Mapping Controversy: A Cartography of Taxonomy and Biodiversity for the Philosophy of Biology

Charles H. Pence[1*] *and Stijn Conix*[1,2]

[1] *Université catholique de Louvain, Institut supérieur de philosophie, Louvain-la-Neuve, Belgium*
[2] *KU Leuven, Hoger Instituut voor Wijsbegeerte, Leuven, Belgium*
[*] *corresponding author, charles@charlespence.net*

Abstract

One potentially extremely fruitful use of the tools of corpus analysis in the philosophy of science is to help us understand disputed terrains within the sciences that we study. For philosophers of biology, for instance, few controversies are as heated as those over the concepts we use in taxonomy to classify the living world, with the definition of 'species' perhaps most fundamental among them. As many understandings of biodiversity, in turn, involve counting the number of species present in a given area, these taxonomic concepts thus become crucially implicated in our efforts in conservation biology and our response to climate change. In this chapter, we present a corpus of taxonomic journal papers, and illustrate how it might be used to make progress in the history and philosophy of taxonomy. What parts of the biological world do taxonomists most often study? Are these areas of focus related to other methodological commitments, or perhaps to their underlying conceptual disagreement? Are species that are more important to conservation, or economic concerns, or more important to local cultures, more or less likely to be the target of biological study? Corpus-based methods, we argue, are uniquely powerful for approaching questions such as these, and we hope that the case we present can serve as a fruitful example for others considering their implementation.

1. Introduction

While biodiversity has become one of the most commonly cited concepts in twenty-first century environmental discourse, its conceptual foundations are recognized—by philosophers and biologists alike—to be murky at best. There are a host of different notions of biodiversity, including those that focus on genetic diversity, ecological diversity, morphological diversity, and, most commonly, various ways in which to count the species present in a given area (Maclaurin and Sterelny 2008). The prevalence of the latter type of definition, then, means that in many cases, our evaluations of biodiversity are dependent upon the concept of 'species' at use in a particular context (Sarkar 2002, 137). The concept of 'species', in turn, is highly controversial, with a small number of distinct viable understandings of what species actually *are* (Zachos 2016; Mayden 1997; Wheeler and Meier 2000), supported by a vast array of ways in which to *empirically delimit* species in practice (Camargo and Sites 2013; Sites and Marshall 2004; 2003). Much of our approach to contemporary taxonomy and conservation, then, is affected by a kind of conceptual uncertainty. The drivers of this uncertainty are highly multifaceted and can be difficult to discern even in close-reading approaches to the taxonomic literature in particular taxa (Cuypers, Reydon, and Artois 2022; Hodgson 1961; McClure et al. 2020). This reflects, in part, the ambiguity in taxonomy's fundamental concepts (and their operationalization), as well as the deep sense in which taxonomy and conservation are enterprises laden with non-epistemic value judgments (Conix 2019; Thiele et al. 2021; Padial and De la Riva 2021).

This much is second nature to philosophers of science who have explored questions of taxonomy, biodiversity, or conservation biology. It has, however, a further implication: it makes it all the more important that we understand the fine-grained detail of taxonomic practice. How do biologists actually *do* taxonomy? Which species do they study? Which methodologies do they employ? Can we find evidence of conceptual disagreement in the products of their work? And how do all of the preceding factors relate to one another? Answering questions such as these can serve a variety of philosophical goals. Our attention might be drawn to particular areas of taxonomic study that raise conceptual or methodological questions, and hence merit more discussion than they have received in the philosophical literature. Or we might come to expect that certain areas of taxonomic practice would have implications for active, public debates over conservation or governance, and thus be sites of productive, real-world intervention (see e.g. Wege et al. 2015).

A range of tools are already available today for philosophers and historians to investigate taxonomy using text and data mining (Bingham et al. 2017). Most importantly, Biodiversity Heritage Library (BHL; Gwinn and Rinaldo 2009; Agosti et al. 2019) has made large parts of the historical biodiversity literature publicly available online for search and download. BHL includes a taxonomic name finding tool to search for names in its enormous corpus, and is invaluable for the study of taxonomy from Linnaeus up to relatively recently. For bibliographic data, philosophers and historians of taxonomy can turn to Clarivate's Zoological Record (https://clarivate.com/webofsciencegroup/solutions/webofscience-zoological-record/) in the Web of Science, probably the oldest database of animal taxonomy and biology more generally. It dates back to 1864, covers over 4000 journals, and through its integration in Web of Science has access to all tools of that database. Coupled to the Zoological Record is the Index of Organism Names (http://www.organismnames.com/), which claims to be the most complete source of organism names.

For those, like us, interested in the taxonomic literature, Plazi is an invaluable resource (Agosti and Egloff 2009; Catapano, Agosti, and Sautter 2007). This platform provides open access to taxonomic and biodiversity data liberated from scholarly publications. It does this through its Biodiversity Literature Repository, and, most importantly, through its TreatmentBank (Agosti et al. 2022). TreatmentBank provides open access to over 650,000 taxonomic treatments, figures, treatment citations and tables in a highly 'FAIR' way (Wilkinson et al. 2016). Plazi is closely integrated with yet another database, OpenBiodiv, which also contains biodiversity data extracted from the scientific literature (Penev et al. 2019).

In addition to these tools, scholars interested in the discipline and its functioning can also use more general databases. Catalogue of Life provides the most complete and authoritative list of the world's species and higher classification (Bisby and Roskov 2010). Encyclopedia of Life provides access to data about all life through its TraitBank (Parr et al. 2014). And GBIF (https://www.gbif.org/), finally, aims at providing open access to all biodiversity data.

While the tools listed above provide scholars with unprecedented opportunities to study the field of taxonomy and its problems, there remains one significant lacuna: a full text, analyzable corpus of recent taxonomic literature. Even if much of the data from taxonomic literature was already available, a lot of additional information gets lost without the full texts. If we want to study the use of concepts, methods, taxonomic reasoning, disagreement and conflict resolution, we need access to the full texts in which all of this is described. Our aim was to fill this gap by creating a representative full text corpus of recent taxonomic literature available for corpus methods.

In this chapter, we will present the corpus of taxonomic documents, as well as a hopefully illuminating example of the kinds of uses to which this corpus may be put to work in contemporary philosophy of science. While the analyses we present here are preliminary (and will be supplemented by further work in the future), we hope they illustrate how such work can be productive in guiding and refining research, in shaping philosophical analyses and providing material to employ both in non-digital philosophical work (e.g., Cuypers, Reydon, and Artois 2022), and in the construction or interpretation of other, non-corpus empirical work (like surveys or vignette studies; **TO CITE**).

2. Building a Corpus

Standard questions of corpus construction are beyond our scope for this chapter (and largely consist in dialogues with publishers and the writing of boilerplate website scraping code). But it is nonetheless important to briefly present the motivations behind the construction of our corpus, and the balance between pragmatism and completeness that characterizes, to some degree, the construction of every large corpus in digital humanities. To begin, we are somewhat lucky in that many formal papers in taxonomy—presenting new species, modifying our classification of extant organisms, and so forth—tend to be published in a relatively small number of highly specialized journals. Especially in the twenty-first century, after almost all publication moved online and into these journals, one can therefore not unreasonably hope to have collected, if not a complete collection, at least a very large and broadly representative set of work in contemporary taxonomy.[1] Furthermore, a number of the relevant journals—especially a collection of journals published by

[1] With the possible exception of publications in taxonomic monographs, some of which are included as “journal articles” in our corpus but for many of which we lack access.

Pensoft, including *ZooKeys*, *PhytoKeys*, and *MycoKeys*—are published open-access. Another publisher, Magnolia Press, while not publishing open-access, has shown extensive interest in digital analyses of their published corpus, and was thus receptive to discussions with us concerning data access, although we did need to negotiate paid subscriptions to the articles in order to have access to the journal.

In the end, combining the Magnolia journal *Zootaxa* with the journals from Pensoft and two other open-access journals (one generalist and one dedicated to insects), we constructed a corpus of a bit more than forty thousand articles, in full text (i.e., including metadata as well as the full content of each article; Table 1). Even this relatively large corpus, however, was the result of a number of trade-offs. While we have limited coverage of botanical work with the eight hundred articles of *PhytoKeys*, this corpus is likely to under-represent botanical taxonomy.[2] Similarly, other journals that we explored ran into other kinds of copyright and access concerns, and had to be excluded.[3] Corpus construction will always necessitate judgment calls of this sort; it is important that they be discussed explicitly, and justified to the fullest extent possible given these compromises.

Journal	**Publisher**	**Size**
Zootaxa	Magnolia Press	31,348
ZooKeys	Pensoft	4,940
PhytoKeys	Pensoft	820
Journal of Hymenoptera Research	Pensoft	382
MycoKeys	Pensoft	315
Zoosystematics and Evolution	Pensoft	153
Insecta Mundi	Center for Systematic Entomology	1,367
European Journal of Taxonomy	Museum National d'Histoire Naturelle	1,105

Table 1. *The journals included in our corpus, with the number of articles specified. Total size of the resulting corpus is N=40,403. After removing documents that have no full text available, the corpus*

[2] One obvious way in which we might have corrected this—including the journal *Phytotaxa* published by Magnolia Press—was ruled out on cost grounds.

[3] Disturbingly common, for instance, is the demand for analysis work to be pre-cleared by writing to an email address that appears to be entirely unmonitored.

used for analysis was of size N=40,316, containing 172,948,456 individual words (tokens) and 983,323 unique types.

A further important aspect in building such corpora is to find ways in which we can validate them, reassuring ourselves that we've arrived at a corpus that is representative enough for present purposes. In our case, it is important to verify that the corpus that we've collected is reasonably representative of the species that taxonomists are interested in. To visualize this, we extracted all of the species names from every article in the corpus using the gnfinder tool (Mozzherin, Myltsev, and Zalavadiya 2022, about which more later). These species names could then be correlated with their corresponding records in the Open Tree of Life, which offers a single phylogenetic tree that is built by a consortium that aims to integrate all extant taxonomic work into a synthetic, consensus output (OpenTreeOfLife et al. 2021). This tree, then, can stand in as a kind of proxy for the global effort of researchers in taxonomy. If we visualize the phylogenetic tree containing only species mentioned in our corpus alongside the entirety of the Open Tree of Life, we will be able to see at a glance the extent to which the collection of species discussed in our corpus is representative of the discipline as a whole.[4]

[4] The Open Tree of Life tree had to be simplified in order to be processed for visualization; for details, see the online data package that accompanies this chapter at **FIXME URL TO OSF PACKAGE**.

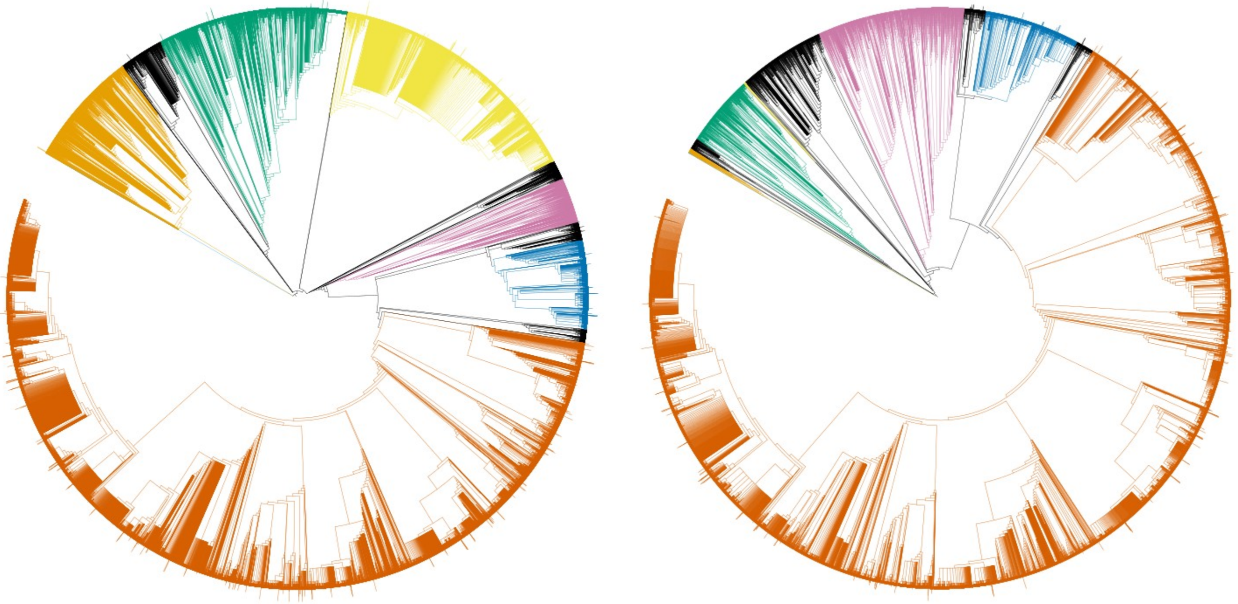

Figure 1. *The complete Open Tree of Life dataset (version 13.4; left) and the set of species in our corpus (right). Visualized using iTOL (Letunic and Bork 2021) after pruning of a number of leaf nodes for computational tractability. Colors correspond to large, colloquial groups: bacteria in gold, archaea (nearly invisible) in pale blue, plants in green, fungi in yellow, molluscs and annelids in dark blue, arthropods in orange, and chordates in pink.*

Comparing the full OTOL (Figure 1, left) with our corpus (Figure 1, right), one sees a few important differences. Most notably, bacterial and fungal taxonomy are almost absent in our database. *A priori*, this is not surprising; bacterial taxonomy, as is well known, functions quite differently, employs different and sometimes quite particular species concepts, and is discussed in different journals (see Franklin 2007). While we hoped that fungal taxonomy would be covered by the journal *MycoKeys*, at only 315 articles it does not appear to have made an appreciable difference. Plants are somewhat underrepresented, and chordates somewhat overrepresented. With these caveats in mind, however, the taxonomic scope of the corpus collected is impressive, and it seems to offer us a reasonably complete approach to the discipline of taxonomy.

3. Methodological Tools

Before detailing the ways in which we plan to analyze this corpus, we begin by briefly presenting the tools that we deployed in order to analyze it. We do not intend to offer a complete introduction to either these methodologies or to the precise ways in which we implemented them here. For the former, we recommend any of a number of contemporary generalist works in digital humanities (e.g., Jockers 2014; Ahnert et al. 2020; Berry and Fagerjord 2017), while for the latter, our data and source code can be consulted via links found in footnotes below.

3.1. Topic Modeling

Central to our approach is the use of *topic modeling*, a procedure which helps us see what the different articles in a particular corpus might be "about." There are a number of ways to describe how topic modeling works, but perhaps the most perspicuous is to present it as offering an idealized model of document creation, then solving for the open parameters in that model using an unsupervised machine-learning approach (i.e., an approach on which researchers do not have to input information in advance about the meaning of terms). Topic models assume that a document is created in the following way (Blei 2012). First, create *N* probability distributions over every word type in the document (call these the *topics*). Then, create another distribution for each document over each of those *N* topics, describing how probable each topic is in that document. To "write" the document, first draw a topic from the latter distribution, then draw a word type from the topic itself. Continue until you have enough words to build your document.

This is, evidently, a massively idealized model. It includes no information about the *order* or *syntactic structure* of the document—the "document" produced is simply a set of word types. It also in no way constrains the distributions in the topics—they can be whatever kind of distribution would produce, in the end, documents that look like the documents found in the corpus. What is surprising—and the reason for which topic models are so popular—is that when we solve for these various distributions, the topics that result act like "topics" in the colloquial sense. That is, when we inspect them—for instance, by looking at the top twenty words that the topic is most likely to select—we tend to find that they capture identifiable, interpretable themes in those documents. A topic model of a journal like *Science*, for instance, might have one topic with the words “women,” “students,” “universities,” and “education” selected as probable, and another that picks out “stars,” “astronomers,” “universe,” and “galaxies” (Blei and Lafferty 2007). We can infer that documents for which the former topic is probable are likely to be reviews of science education and diversity efforts, and those for which the latter topic is probable are likely to be articles in astronomy. This success has been duplicated across a variety of fields, including historical documents, machine translation, social science, (Boyd-Graber, Hu, and Mimno 2017), literature (Erlin 2017), and even the philosophy of science itself (Malaterre, Chartier, and Pulizzotto 2019; Malaterre, Pulizzotto, and Lareau 2020).

The primary choice of the researcher in the application of a topic model is the number of topics *N*, though there are a variety of evaluations that can be performed to help quantify the quality of a given topic model (Röder, Both, and Hinneburg 2015). We used the *gensim* package to perform topic modeling (Řehůřek and Sojka 2010), and evaluated the coherence of each model (using c_v coherence), selecting the model with the highest coherence value. The details of our process of topic model training, model selection, and the code that we used are freely available online, as well as all of the trained models that we did not decide to select.[5]

3.2. Species Name Detection

If we want to understand which species are discussed in the taxonomic literature, we need a way to pick out species names from the full-text content of our corpus. The *gnfinder* package uses a combination of heuristic features (e.g., the patterns of capitalization that are found in species and genus names), a Bayesian classifier which has been trained on a large set of papers for which

[5] See the online data package that accompanies this chapter at **FIXME URL TO OSF PACKAGE**.

taxonomic names had been manually tagged, and a verification system which checks those detected names against a number of online databases containing lists of species, genera, etc. (Mozzherin, Myltsev, and Zalavadiya 2022). It offers extremely fast processing as well as an estimate of the quality of a given match (i.e., the software's own confidence in its assessment of a given string as a taxonomic name). Our corpus was processed for taxonomic name extraction on consumer-grade computing equipment in a matter of several hours.

Species Concept
Phylo-Phenetic Species Concept
Phylogenetic Species Concept
Genic Species Concept
Cohesion Species Concept
Genealogical Concordance Species Concept
Genotypic Cluster Species Concept
Genetic Species Concept
Ecological Species Concept
Recognition Species Concept
Genealogical Species Concept
Biological Species Concept
Differential Fitness Species Concept
Compilospecies Concept
Cladistic Species Concept
Hennigian Species Concept
Internodal Species Concept
Mitonuclear Compatibility Species Concept
Pragmatic Species Concept
Inclusive Species Concept
Biosimilarity Species Concept

Table 2. *A list of all species concepts for which we searched in our dataset. List adapted from Zachos (2016).*

3.3. Species Concept Detection

Finally, if we want to understand when taxonomists explicitly invoke various species concepts, we need a way to search for those concepts in those documents. Here, we are aided by the fact that species concepts have particular proper names, like the Biological Species Concept. If an author is going to explicitly invoke such a concept, they are almost certain to use one of the phrases found in Table 2.[6] Detecting species concepts in the literature therefore requires only a text search algorithm; we used the Python implementation of the FlashText algorithm (Singh 2017).

4. Results

With the corpus and tools in place, we can start answering questions about the role and nature of disagreement in taxonomy, the use of species concepts, or taxonomic method. In this section, we present preliminary results on the basis of a first exploration of the corpus. While more thorough research on all these questions is needed (and will happen, in due time) we hope these can serve as an illustration of how the corpus can be used to answer the kinds of questions listed above. More precisely, we present the basic results, first from the detection of species names and concepts, and then from the topic model. Finally, we explore various correlations between these data sources.

[6] Each of these concepts also tends to come with a classic citation, that is, an article that is usually referenced as having developed or introduced the concept. Searching for these did not prove necessary in our case.

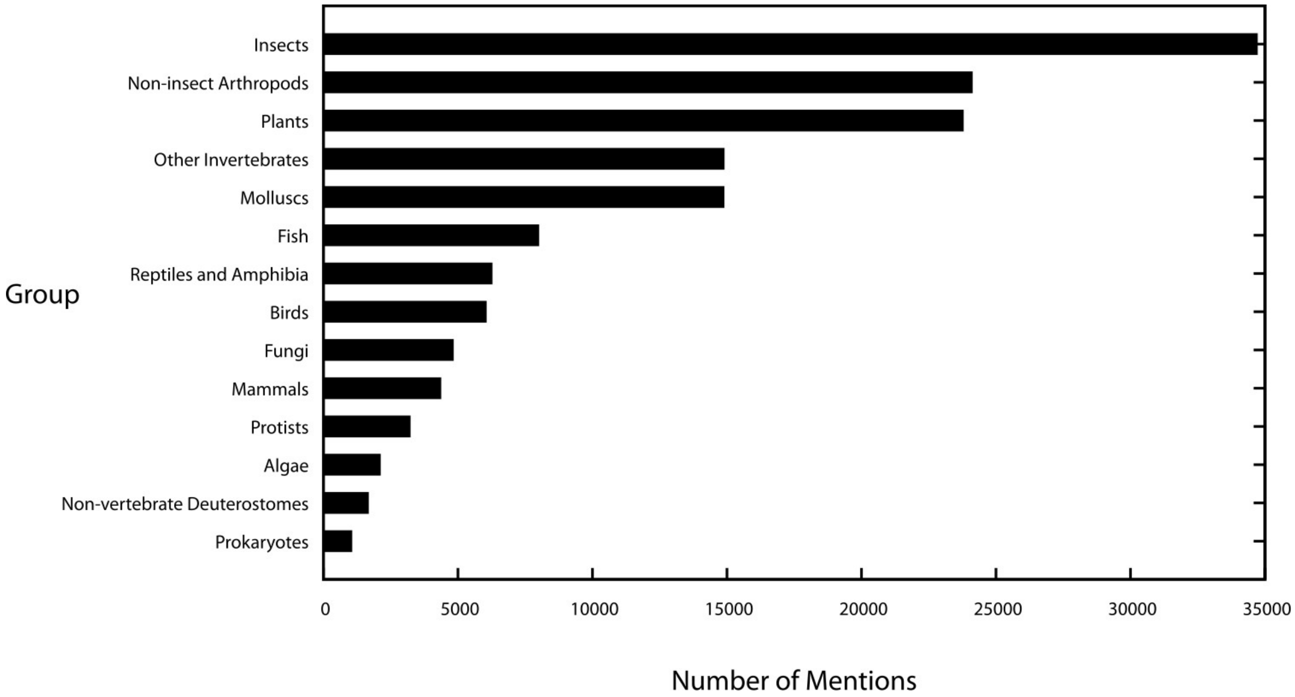

Figure 2. *The number of mentions of each major colloquial group of organisms in our corpus. Extracted using gnfinder as described in Section 3.2.*

4.1. Species Names and Concepts

We categorized the species names appearing in the corpus into a collection of large, colloquial group names, by filtering them according to the complete "taxonomic path" (from kingdom down to species) provided by the validation system in *gnfinder* (see section 3.2). Species names were thus grouped into Mammals, Birds, Reptiles and Amphibia, Fish, Non-Vertebrate Deuterostomes, Insects, Non-Insect Arthropods, Molluscs, Other Invertebrates, Plants, Fungi, Algae, Protists, and Prokaryotes.[7] (In what follows, when referring to these colloquial groups as detected algorithmically in our corpus, we will capitalize them, to avoid confusion with these terms in everyday usage.) The number of occurrences of mentions to each of these groups in our corpus is represented in Figure 2.

A few things are immediately notable from this graph. First, a few groups are extremely rare in our corpus (confirming the intuitive impression that we saw in Figure 1). In particular, the

[7] Algae, Protists, and Reptiles and Amphibia were selected by comparison with an explicit list of taxa. Plants refer to all non-algae plants. Protists also excludes the algae. Prokaryotes refers to all bacteria and archaea. For more precise details, the algorithm for assigning a taxonomic path to these categories can be found in our source code.

number of mentions of Protists, Algae, Non-Vertebrate Deuterostomes, and Prokaryotes are extremely small. Signal from these small groups tended to cause unusual results in a variety of significance- and probability-testing analyses, a well-understood problem in digital humanities (for the case of pointwise mutual information, e.g., see Role and Nadif 2011). In a number of the analyses that follow, these four groups were therefore ignored. Given that we already had reason to believe that our coverage of them would be non-representative, this appears to be a defensible choice. Second, the dominance of taxonomic work on arthropods is as impressively present here as it was in the graphical form of Figure 1. We will see below that interpretation of the topic model often requires knowledge of some relatively fine points of insect and other arthropod anatomy, emphasizing the importance of domain knowledge in the analysis of these corpora.

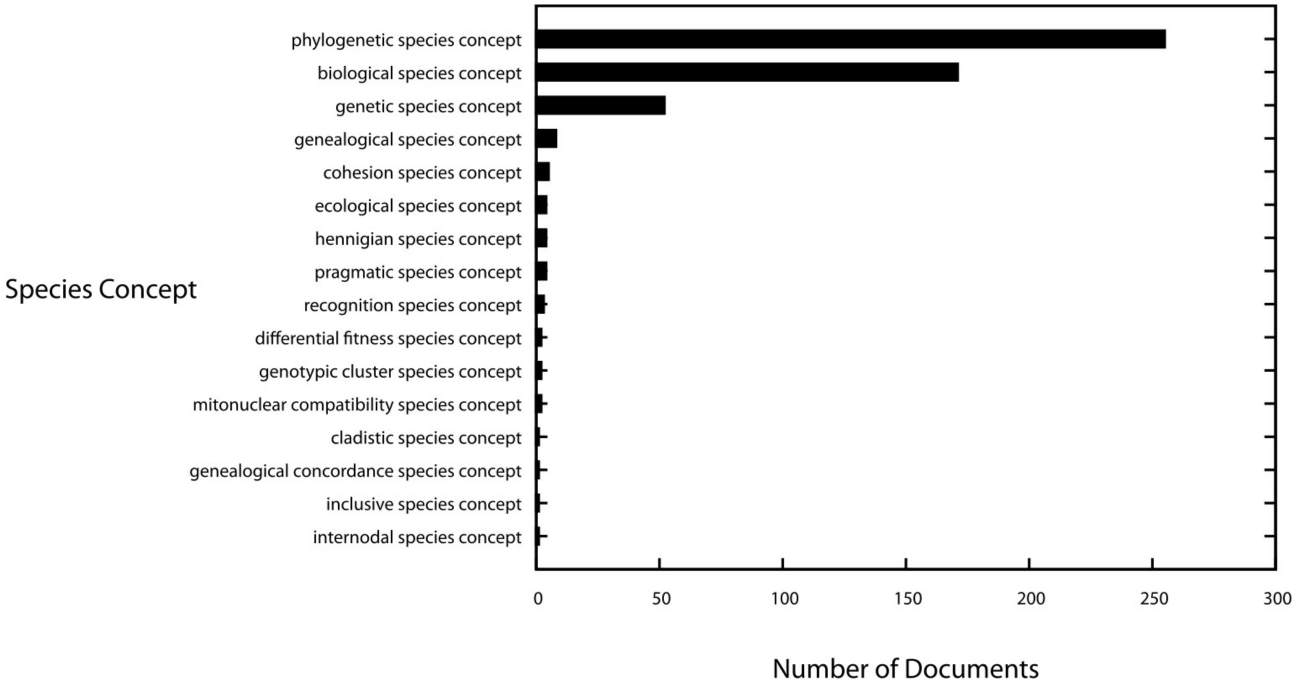

Figure 3. *The number of documents that mention each species concept within the corpus, for all concepts with a non-zero value.*

We can then turn to the detection of species concept terms in the corpus, shown graphically in Figure 3. Strikingly, there are really only three species concepts that are mentioned often, and these combine for a total of only 478 mentions, that is, a species concept is mentioned in at most around 1% of the papers in our corpora. Of these, the phylogenetic species concept dominates, followed by the biological species concept; the genetic species concept ranks a distant third. This is a small enough number of documents that one could envision a close-reading project to explore literally every invocation of a species concept name in our corpus.

Before turning to the topic modeling results, we can look directly at the relationship between invocations of species concepts and invocations of different taxa. Because the frequencies of use of species concept terms are so small that they are likely to pose problems for traditional significance testing, the most reliable way to explore this relationship is to look at the values for the proportion of documents that invoke both a particular species concept and mention a particular taxon. When we take the top 5% of those proportions, we find the most "significant" pairings to be the use of the phylogenetic species concept in papers that mention Mammals, Birds, Reptiles and Amphibia, Fish, and Protists, and the use of the biological species concept in papers that mention Mammals and Birds.

4.2. Topic Model

After evaluating a collection of topic models over our corpus, of sizes from $N = 2$ to 250, we found that the highest quality model was a 125-topic model; we have used this model for all of the analyses that follow. We followed standard methodology for interpreting the topics in this model—most prominently, examining the top twenty words marked out as most probable for this topic, inspecting the documents for which this topic was most probable, and examining (as we will discuss below) the correlations between the probability of a topic and taxon and species concept mentions. In rare cases where this did not suffice for establishing the identity of a topic, we informally consulted taxonomists.

The majority of topics found in the topic model refer to the anatomy of particular groups. When we examine the most probable words chosen by Topic 44, for instance, we find words like ‘ray’, ‘dorsal’, ‘river’, ‘fin’, ‘fish’, and ‘scale’: clearly a topic picking out discussions of fish anatomy. Similar topics are present for, among others, reptiles, rodents, parasites, worms, jewel beetles, flowering plants, water mites, and crustacea. It is unsurprising that the technical terms used to describe these various groups would produce a particularly clear signal within a topic model.

Interesting problems occur, however, when this tight identification between anatomical terms and groups begins to break down. Consider, as a representative example, Topic 31. Its most probable terms include words like ‘lepidoptera’, ‘forewing’, and ‘scale’, indicating a clear focus on the butterflies, as well as (even more probable than these) the words ‘male’, ‘genitalia’, and ‘female’—we have here what seems to be a topic referring to the reproductive anatomy of butterflies. But when we proceed to investigate the species mentioned in papers where this topic is particularly probable, we find that the topic is particularly prominent *in Molluscs*, not in Insects. But both butterflies and molluscs have reproductive anatomy called ‘bursa’ (Latin simply for ‘sac’), and an entire genus of snails is known as *Bursa*. The topic model, then, is incapable of distinguishing these uses of the same term.

A second type of topic found in the model concerns *locations* in which taxonomic work is performed. Topic 16, for instance, includes words like ‘colombia’, ‘peru’, ‘rica’, ‘del’, and ‘san’—a topic describing work which took place in Central and South America. Similar topics exist collecting words that describe Southeast Asia, Brazil, South America as a whole, Mexico, and the Western US and Canada.[8] Types of habitat also make an appearance, including freshwater rivers, the forest floor, Pacific marine habitats, and more.

Finally, some topics related to the practice of taxonomic methodology are present. Two of the topics that are most common throughout the corpus—that is, which are highly probable for a

[8] In future work, we hope to extract mentions of references to particular places from the corpus and use these to explore the geospatial dimensions of taxonomic work in greater depth.

significant portion of papers regardless of their other subjects—concern two core methods. Topic 9, with terms such as ‘find’, ‘collect’, ‘site’, ‘study’, ‘record’, ‘population’, ‘range’, and ‘sample’, seems to describe terrestrial specimen collection, and is thus highly probable in every taxon except Non-Insect Arthropods (i.e., crustacea), Fish, and Fungi. Even more striking is Topic 64 which, picked out by terms like ‘sequence’, ‘analysis’, ‘molecular’, ‘dna’, ‘phylogenetic’, and ‘clade’, describes molecular phylogenetics. This topic is not only highly probable in every single taxon, it is among the top twenty largest absolute probability values for any topic in five different groups. Put briefly: *all* of contemporary taxonomy makes appeal, at least to some degree, to molecular methodologies.

Put briefly, the topic model seems to describe a number of facets of taxonomic practice interesting for philosophers of science: differential attention paid to various taxonomic groups, indications of popular geographic sites for study, as well as the distribution of common varieties of methodology.

4.3. Evaluating Disagreement

As we noted in the introduction, taxonomy is also interesting in that it is a decidedly controversial science. The proliferation of methods of species delimitation leads to a lack of consensus or shared standards for the introduction of new species, as well as a pervasive feeling of “taxonomic disorder” or “taxonomic anarchy” that hampers efforts in other fields like conservation (Garnett and Christidis 2017; Isaac, Mallet, and Mace 2004; Agapow et al. 2004; McClure et al. 2020).

Can signal of this disagreement be detected in the corpus using the methods we’ve pursued here? To some extent, yes. Given the rarity of mentions of explicit species concepts, it will be very unlikely that disagreement at this conceptual level is clearly expressed in the corpus. But this is not the only way to hunt for evidence of disagreement. We can also explore the signal of particular words across various topics—in our case, helping us answer a question like “what is the context in which words indicating disagreement or dispute are often used?”

Detecting disagreement textually, however, is not necessarily a simple matter. Of course, we can look at words like ‘disagree’, ‘disagreement’, or ‘dispute’, which would offer a direct signal. We could also look for words that might be used in taxonomy to indicate the response to disagreement, such as when taxonomists propose a ‘revision’ of a taxonomy (or to ‘revise’ it). Some more complex methods for detecting the “sentiment” of texts exist, but these are often very tightly related to their commercial applications, such as the analysis of product reviews or social media posts mentioning a particular brand; efforts to extend these methods beyond this context meet with mixed success (Boyd-Graber, Hu, and Mimno 2017, 83–85). After making a thorough search of the corpus, we will use the presence of the words ‘disagree’, ‘disagreement’, ‘dispute’, ‘disputable’, and ‘argument’ to mark “disagreement,” and combine this with ‘revision’ and ‘revise’ as markers of “revision.”

Remarkably, the top two topics for *all* of the “disagreement” terms are the same: Topics 120 and 43. The “revision” terms are *much* more common overall, and thus their signal is more difficult to interpret. Notably, both Topics 120 and 43 are important for the term ‘revise’, and 43 is also important for ‘revision’; we thus chose not to add any further topics to the analysis in order to prevent over-interpreting the presence of revision-terms in these other cases. The two topics selected thus are attested as important across a large number of terms relating to disagreement and taxonomic revision (and are the only two topics picked out by such a standard).

Topic	Most Probable Words
Topic 120 (“Ranking”)	character, genera, taxon, group, specie, genus, phylogenetic, include, analysis, family, relationship, phylogeny, clade, morphological, classification, support, press, new, consider, present
Topic 43 (“New Species”)	specie, name, description, new, publish, author, nomenclature, code, publication, type, article, zoological, original, synonym, work, list, valid, international, available, note

Table 3. The top twenty most probable words for Topics 120 (“Ranking”) and 43 (“New Species”). Most probable words are listed in descending order of probability.

The most probable words for Topics 120 and 43 are presented in Table 3. Interestingly, the two topics have a significantly different character. Topic 120, which we will call “Ranking,” seems to pick out words that one might expect in disputing the ranking of a particular clade. For example, the words ‘species’,[9] ‘genus’, ‘genera’, and ‘family’ are present. Further, appeals to at least some of the kinds of evidence that one might use to support or refute the ranking of a particular group at a particular taxonomic level are included, including ‘analysis’, ‘phylogeny’, ‘clade’, ‘morphological’, and ‘support’. Put briefly, this topic—the most common one for a majority of the disagreement terms—seems to nicely encapsulate disagreements over the ranking of particular groups of organisms.

Topic 43, on the other hand, includes terms like ‘name’, ‘description’, ‘new’, ‘publish’, ‘code’, ‘list’, and ‘valid’—in short, the words that one would expect when discussing the introduction of a new species. Notably, the word ‘synonym’ also appears. Declaring that one species name is, formally, a synonym of another is the inverse process to the introduction of a new species name, and it is unclear from the words chosen in this topic whether the topic is indiscriminately picking out both the introduction and synonymization of species names, or whether authors arguing for the introduction of a new name also argue that this name should not be read as a synonym of an existing one.

[9] The preprocessing system that we have used, erroneously believing ‘species’ to be the plural version of ‘specie’, removes the trailing ‘s’. This is a harmless transformation, given that the term ‘specie’ does not appear in taxonomic literature.

Once again, we can attempt to investigate the correlation between these two topics and other document features. If these two topics broadly represent the contexts in which disagreement is discussed, it turns out that disagreement is largely constant across taxa. Both the Ranking and the New Species topics differ by only around 25–30% between the most and least probable taxa, and the values in between lie on an essentially smooth gradient—put differently, there are no grounds for saying that some taxa are "more disagreement-prone" than others in our corpus, and any differences that we do find are difficult to distinguish from noise.

When we look at the relationship between these two topics and species concepts, we can analyze the question in a different way. Since mentions of *individual* species concepts are so rare, we can ask instead the following question: if we sum up the probability, for each topic, of every document that mentions *any species concept at all*, which topics are marked out as particularly interesting? Informally, what topics are involved in documents that mention which species concept(s) they use?

The top such topic is the one which refers to molecular phylogenetic methods, as a result of its being so prevalent across the entire corpus. The second-most important such topic, however, is our Ranking topic. When disputes break out over the ranking of a group, it seem that authors begin to refer much more often than they normally would to explicit species concepts. On the strength of such evidence, one might even begin to consider the Ranking topic as one that directly picks out taxonomic disagreement. New Species, on the other hand, is relatively average on this score. Even though disagreement might be present in the addition or removal of new species names, it does not seem that in this context authors are pressed to make their commitment to particular species concepts more explicit.

5. Discussion

To begin, we should consider the interest of these results for the study of taxonomy itself; we will then conclude the discussion with some ideas about the usefulness of the methodology described here to other areas of philosophy. It is important to emphasize once again that the results presented here are preliminary, and will serve as the basis for a more thorough mapping of taxonomy. For example, close inspection of some of the documents found with the searches about disagreement and species concepts can serve to make a more exhaustive list of terms connected to these subjects. These lists can then be used for a second, more exhaustive search. Similarly, other groupings of taxa than the colloquial groupings used here are possible. For example, grouping taxa roughly into kingdoms could yield other interesting results.

Still, even before approaching more complex analyses, these preliminary results show how corpus methods can yield interesting claims. It is already noteworthy that, as we saw in Figure 3, only around one percent of papers in taxonomy clearly note which concept of species they use. Frank Zachos has argued, on the basis of two other studies in botany dating from the 1990s, that "often the species concept or delimitation method is not explicitly stated" in taxonomic articles, and that "the theoretical underpinning on which researchers base their decisions should at least be explicitly stated" (Zachos 2016, 160). We can empirically confirm here that such a practice has not, in general, caught on in the taxonomic literature that we have studied. While it is certainly possible that papers are working with an operative definition of species that could be discerned upon close reading by a competent professional, such discussions are at best relatively camouflaged: no topic in our topic model finds among its top twenty most probable words terms such as 'concept', 'method', 'methodology', or 'delimitation'. Even implicit references to species concepts, then, seem to be relatively rare.

The detection and evaluation of disagreement in the corpus seems to be more successful than one might have initially expected. As nearly every term indicating disagreement seems to be clustered around our two topics Ranking and New Species, within the context of our corpus it seems entirely justified to conclude that these are the two cases in which taxonomic disagreement is particularly prominent. Future work could consider analyzing the fine-grained details of the other topics that mention ‘revise’ and ‘revision’, which could also indicate hot-spots for taxonomic disagreement, this time picked out by anatomical or geographic terms.

To close, we should consider the potential applicability of this method as it might generalize to other areas in philosophy. We cannot, to be sure, claim that we have performed here something like “empirical conceptual analysis.” We are not equipped to answer, for instance, questions about the precise drivers of taxonomic disagreement in a particular case, or the implicit definitions of species that might be in place when taxonomists study various taxa. While other textual analysis tools might be able to illuminate these details, topic modeling appears to be too general for this purpose.

What we have, instead, is something like a first stab at *conceptual cartography*—an approach that can help us understand the broad-brush ways in which taxonomists operate, and one which can give us a departure point for further analyses. Species concepts appear not individually—we do not have precise ways to approach the biological or the phylogenetic species concept—but species concepts as a category can indeed be seen, and their relationships with broad areas of taxonomic effort like the large colloquial groups which appear in Figure 2 can be described. In short, what we have is a way that philosophers can take the measure of a field of source material, at a glance.

What, in turn, might this kind of conceptual cartography be useful for? As we have already seen, it can occasionally support interesting hypotheses on its own merits. The relatively rare invocation of species concepts in general, for instance, is a claim commonly made in studies on taxonomic practice but rarely explicitly empirically supported. At a lower level, however, we argue that this is useful as a way to direct other sorts of philosophical work. Throughout this chapter, we have noted tantalizing leads which could be followed up with the classic work of close reading—for example, exploring in detail a collection of articles which score highly for the Ranking or New Species topics to see the ways in which they might instantiate taxonomic disagreement. We could also dig into the topics themselves in greater depth. The topic describing molecular phylogenetic methods, for instance, also includes among its top-twenty most probable words the term ‘morphological’, which one might think *a priori* was a signal of *non*-moecular methods. Is this term mentioned in a negative or pejorative sense in articles scoring highly for that topic, or is there perhaps a tighter integration between molecular and traditional taxnomic methods than originally acknowledged? This question arises not from armchair philosophical theorizing, but from direct examination of the literature in taxonomy. While there are many ways in which to engage in philosophy that hews closely to scientific practice, we believe that this is one of particular interest.

Acknowledgments

SC was supported by the Research Foundation – Flanders (FWO), under grant number 3H200026. CHP was supported by the Fonds de la Recherche Scientifique - FNRS under grant number T.0177.21. We would like to thank an audience at the Workshop on Taxonomic Disorder, KU Leuven, 2022, for helpful discussions and commentary, especially Thierry Backeljau, Brent Mishler, Leen Vandepitte, and Frank Zachos. Other members of the FNRS project DisSurvey (Maximilia Bautista Perpinyà) and the FWO project Tackling Taxonomic Disorder (Andreas De Block, Thomas Reydon, Tom Artois, and Vincent Cuypers) also provided much helpful discussion and feedback.

References

Agapow, Paul-Michael, Olaf R. P. Bininda-Emonds, Keith A. Crandall, John L. Gittleman, Georgina M. Mace, Jonathon C. Marshall, and Andy Purvis. 2004. “The Impact of Species Concept on Biodiversity Studies.” *The Quarterly Review of Biology* 79 (2): 161–79. https://doi.org/10.1086/383542.

Agosti, Donat, Terry Catapano, Guido Sautter, Puneet Kishor, Lars Nielsen, Alexandros Ioannidis-Pantopikos, Chiara Bigarella, Teodor Georgiev, Lyubomir Penev, and Willi Egloff. 2019. "Biodiversity Literature Repository (BLR), a Repository for FAIR Data and Publications." *Biodiversity Information Science and Standards* 3 (June): e37197. https://doi.org/10.3897/biss.3.37197.

Agosti, Donat, and Willi Egloff. 2009. "Taxonomic Information Exchange and Copyright: The Plazi Approach." *BMC Research Notes* 2 (1): 53. https://doi.org/10.1186/1756-0500-2-53.

Agosti, Donat, Patrick Ruch, Jose Benito Gonzalez Lopez, and Lyubomir Penev. 2022. "Enabling Published Taxonomic Data to Be Used to Address the Biodiversity Crisis: Biodiversity Literature Repository and TreatmentBank." *Biodiversity Information Science and Standards* 6 (February): e91167. https://doi.org/10.3897/biss.6.91167.

Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart. 2020. *The Network Turn: Changing Perspectives in the Humanities*. Cambridge: Cambridge University Press.

Berry, David M., and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity.

Bingham, Heather, Michel Doudin, Lauren Weatherdon, Katherine Despot-Belmonte, Florian Wetzel, Quentin Groom, Edward Lewis, et al. 2017. "The Biodiversity Informatics Landscape: Elements, Connections and Opportunities." *Research Ideas and Outcomes* 3 (September): e14059. https://doi.org/10.3897/rio.3.e14059.

Bisby, Frank Ainley, and Yury Roskov. 2010. "The Catalogue of Life: Towards an Integrative Taxonomic Backbone for Biodiversity." In *Tools for Identifying Biodiversity: Progress and Problems.*, edited by Pier Luigi Nimis and Regine Vignes Lebbe, 37–42. Paris: EUT - Edizioni Universita di Trieste. http://centaur.reading.ac.uk/15706/.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77. https://doi.org/10.1145/2133806.2133826.

Blei, David M., and John D. Lafferty. 2007. "A Correlated Topic Model of *Science*." *The Annals of Applied Statistics* 1 (1): 17–35. https://doi.org/10.1214/07-AOAS114.

Boyd-Graber, Jordan, Yuening Hu, and David Mimno. 2017. "Applications of Topic Models." *Foundations and Trends in Information Retrieval* 11 (2–3): 143–296. https://doi.org/10.1561/1500000030.

Camargo, Arley, and Jack Jr Sites. 2013. "Species Delimitation: A Decade After the Renaissance." *The Species Problem - Ongoing Issues*, February. https://doi.org/10.5772/52664.

Catapano, Terry, Donat Agosti, and Guido Sautter. 2007. "Plazi.Org: Using DSpace as a Repository of Species Descriptions. Dspace User Group Meeting in Rome, Rome, Italy." Zenodo. https://doi.org/10.5281/zenodo.3552076.

Clarivate. n.d. "Zoological Record." Web of Science Group. Accessed December 15, 2022. https://clarivate.com/webofsciencegroup/solutions/webofscience-zoological-record/.

Conix, Stijn. 2019. "Taxonomy and Conservation Science: Interdependent and Value-Laden." *History and Philosophy of the Life Sciences* 41 (2): 15. https://doi.org/10.1007/s40656-019-0252-3.

Cuypers, Vincent, Thomas A. C. Reydon, and Tom Artois. 2022. "Deceiving Insects, Deceiving Taxonomists? Making Theoretical Sense of Taxonomic Disagreement in the European Orchid Genus *Ophrys*." *Perspectives in Plant Ecology, Evolution and Systematics* 56 (September): 125686. https://doi.org/10.1016/j.ppees.2022.125686.

Erlin, Matt. 2017. "Topic Modeling, Epistemology, and the English and German Novel." *Journal of Cultural Analytics*, May. https://doi.org/10.22148/16.014.

Franklin, L. R. 2007. "Bacteria, Sex, and Systematics." *Philosophy of Science* 74 (1): 69–95. https://doi.org/10.1086/519476.

Garnett, Stephen T., and Les Christidis. 2017. “Taxonomy Anarchy Hampers Conservation.” *Nature* 546 (7656): 25–27. https://doi.org/10.1038/546025a.

Gwinn, Nancy E., and Constance Rinaldo. 2009. “The Biodiversity Heritage Library: Sharing Biodiversity Literature with the World.” *IFLA Journal* 35 (1): 25–34. https://doi.org/10.1177/0340035208102032.

Hodgson, Robert W. 1961. “Taxonomy and Nomenclature in Citrus.” *International Organization of Citrus Virologists Conference Proceedings (1957-2010)* 2 (2). https://doi.org/10.5070/C58mc9c8bp.

Isaac, Nick J.B., James Mallet, and Georgina M. Mace. 2004. “Taxonomic Inflation: Its Influence on Macroecology and Conservation.” *Trends in Ecology & Evolution* 19 (9): 464–69. https://doi.org/10.1016/j.tree.2004.06.004.

Jockers, Matthew L. 2014. *Text Analysis with R for Students of Literature*. Cham: Springer.

Letunic, Ivica, and Peer Bork. 2021. “Interactive Tree Of Life (ITOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation.” *Nucleic Acids Research* 49 (W1): W293–96. https://doi.org/10.1093/nar/gkab301.

Maclaurin, James, and Kim Sterelny. 2008. *What Is Biodiversity?* Chicago: University of Chicago Press.

Malaterre, Christophe, Jean-François Chartier, and Davide Pulizzotto. 2019. “What Is This Thing Called Philosophy of Science? A Computational Topic-Modeling Perspective, 1934–2015.” *HOPOS* 9 (2): 215–49. https://doi.org/10.1086/704372.

Malaterre, Christophe, Davide Pulizzotto, and Francis Lareau. 2020. “Revisiting Three Decades of Biology and Philosophy: A Computational Topic-Modeling Perspective.” *Biology and Philosophy* 35 (5). https://doi.org/10.1007/s10539-019-9729-4.

Mayden, R. 1997. “A Hierarchy of Species Concepts: The Denouement in the Saga of the Species Problem.” In *Species, the Units of Biodiversity, Systematics Association Special Volume Series*, edited by M Claridge, H Dawah, and Robert A. Wilson, 381–424. 54. London: Chapman & Hall.

McClure, Christopher J. W., Denis Lepage, Leah Dunn, David L. Anderson, Sarah E. Schulwitz, Leticia Camacho, Bryce W. Robinson, et al. 2020. “Towards Reconciliation of the Four World Bird Lists: Hotspots of Disagreement in Taxonomy of Raptors.” *Proceedings of the Royal Society B: Biological Sciences* 287 (1929): 20200683. https://doi.org/10.1098/rspb.2020.0683.

Mozzherin, Dmitry, Alexander Myltsev, and Harsh Zalavadiya. 2022. “Gnames/Gnfinder: V1.0.1.” Zenodo. https://doi.org/10.5281/zenodo.7131329.

OpenTreeOfLife, Benjamin Redelings, Luna Luisa Sanchez Reyes, Karen A. Cranston, Jim Allman, Mark T. Holder, and Emily Jane McTavish. 2021. “Open Tree of Life Synthetic Tree: V13.4,” June. https://doi.org/10.5281/zenodo.3937742.

Padial, José M., and Ignacio De la Riva. 2021. “A Paradigm Shift in Our View of Species Drives Current Trends in Biological Classification.” *Biological Reviews* 96 (2): 731–51. https://doi.org/10.1111/brv.12676.

Parr, Cynthia S., Nathan Wilson, Patrick Leary, Katja Schulz, Kristen Lans, Lisa Walley, Jennifer Hammock, et al. 2014. “The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth.” *Biodiversity Data Journal* 2 (April): e1079. https://doi.org/10.3897/BDJ.2.e1079.

Penev, Lyubomir, Mariya Dimitrova, Viktor Senderov, Georgi Zhelezov, Teodor Georgiev, Pavel Stoev, and Kiril Simov. 2019. “OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science.” *Publications* 7 (2): 38. https://doi.org/10.3390/publications7020038.

Řehůřek, Radim, and Petr Sojka. 2010. “Software Framework for Topic Modelling with Large Corpora.” In *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks*, 45–50. Valetta, MT: University of Malta. http://is.muni.cz/publication/884893/en.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures.” In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. Shanghai China: ACM. https://doi.org/10.1145/2684822.2685324.

Role, François, and Mohamed Nadif. 2011. “Handling the Impact of Low Frequency Events on Co-Occurrence Based Measures of Word Similarity: A Case Study of Pointwise Mutual Information.” In *International Conference on Knowledge Discovery and Information Retrieval*. Paris: SciTePress.

Sarkar, Sahotra. 2002. “Defining ‘Biodiversity’; Assessing Biodiversity.” *The Monist* 85 (1): 131–55.

Singh, Vikash. 2017. “Replace or Retrieve Keywords in Documents at Scale.” arXiv. https://doi.org/10.48550/arXiv.1711.00046.

Sites, Jack W., and Jonathon C. Marshall. 2003. “Delimiting Species: A Renaissance Issue in Systematic Biology.” *Trends in Ecology & Evolution* 18 (9): 462–70. https://doi.org/10.1016/S0169-5347(03)00184-8.

———. 2004. “Operational Criteria for Delimiting Species.” *Annual Review of Ecology, Evolution, and Systematics* 35: 199–227.

Thiele, Kevin R., Stijn Conix, Richard L. Pyle, Saroj K. Barik, Les Christidis, Mark John Costello, Peter Paul van Dijk, et al. 2021. “Towards a Global List of Accepted Species I. Why Taxonomists Sometimes Disagree, and Why This Matters.” *Organisms Diversity & Evolution*, July. https://doi.org/10.1007/s13127-021-00495-y.

Wege, Juliet A., Kevin R. Thiele, Kelly A. Shepherd, Ryonen Butcher, Terry D. Macfarlane, and David J. Coates. 2015. “Strategic Taxonomy in a Biodiverse Landscape: A Novel Approach to Maximizing Conservation Outcomes for Rare and Poorly Known Flora.” *Biodiversity and Conservation* 24 (1): 17–32. https://doi.org/10.1007/s10531-014-0785-4.

Wheeler, Quentin D., and Rudolf Meier. 2000. *Species Concepts and Phylogenetic Theory: A Debate*. New York, NY: Columbia University Press.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Zachos, Frank E. 2016. *Species Concepts in Biology: Historical Development, Theoretical Foundations and Practical Relevance*. Basel: Springer.