Measuring What We Don't Know: Biodiversity Catalogs Reveal Bias in Taxonomic Effort

JACOB A. GORNEAU, SIDDHARTH KULKARNI, FRANKLYN CALA-RIQUELME, AND LAUREN A. ESPOSITO

Biodiversity catalogs are an invaluable resource for biological research. Efforts to scientifically document biodiversity have not been evenly applied, either because of charisma or because of ease of study. Spiders are among the most precisely cataloged and diverse invertebrates, having surpassed 50,000 described species globally. The World Spider Catalog presents a unique opportunity to assess the disproportionate documentation of spider diversity. In the present article, we develop a taxonomic ratio relating new species descriptions to other taxonomic activity as a proxy for taxonomic effort, using spiders as a case study. We use this taxonomic effort metric to examine biases along multiple axes: phylogeny, zoogeography, and socioeconomics. We also use this metric to estimate the number of species that remain to be described. This work informs arachnologists in identifying high-priority taxa and regions for species discovery and highlights the benefits of maintaining open-access taxonomic databases—a necessary step in overcoming bias and documenting the world's biodiversity.

Keywords: taxonomic databases, World Spider Catalog, species descriptions, spider biology, accessible science

Taxonomists are challenged with the seemingly Sisyphean task of collecting, understanding, and describing species before they go extinct. Although many species remain to be documented by scientists in the face of the present biodiversity crisis, historic efforts have not been in vain, and recent methods continue to advance the pace at which taxonomy can be done (Padial et al. 2010, Bond et al. 2021). For example, focused molecular phylogenetic investigations have helped to delimit species and uncover the backbone of the tree of life (Bickford et al. 2007, de León and Nadler 2010, Fujita et al. 2012, Choi 2016, Hamilton et al. 2016, Fišer et al. 2018, Giribet and Edgecombe 2019, James et al. 2020). This, often paired with detailed morphological analyses and taxonomic revisions, has shone a light on poorly understood lineages in recent years (Grattepanche et al. 2018, Dimitrov and Hormiga 2021). However, inherent biases in how we set about the work of documenting life on Earth can have dramatic effects on our understanding of our biodiversity. For example, taxonomic bias may lead to the neglect of groups that could benefit from further study and an overabundance of taxonomic effort on others. Similarly, geographic bias in historical taxonomic work has affected our understanding of the diversity within and definition of global zoogeographic regions.

Biodiversity catalogs provide a record of taxonomic hierarchy, history, and associated metadata (literature, distribution, museum deposition of types, etc.). The digitization of these catalogs in recent decades has allowed for greater accessibility of taxonomic information that either was previously hidden by paywalls or was not centralized in an easily updated and distributed repository. The associated metadata within these catalogs aid in addressing questions about global environmental change. They hold the potential for highlighting groups of organisms that require more species discovery work and will inform systematists about groups or regions of high priority in the face of the biodiversity crisis. Despite this, these resources can also be valuable in assessing the historical inequity of natural history research. The implications of such insights gleaned from biodiversity catalogs and paired with phylogenomic studies provide a best-case scenario for providing biodiversity data to aid in addressing the present biodiversity crisis and meeting globally established goals for halting their loss (i.e., UN Sustainable Development Goals). Our understanding of biodiversity is likely to improve as new technologies (e.g., parallel sequencing, developments in imaging, molecular analysis) continue to be applied in phylogenetic and population genetic research and may offer expedition to the rate of documentation of species, but these efforts must be focused to yield maximal results.

Spiders are a canvas for the most compelling topics in modern biology, evidenced by their body sizes (from 0.37 to 100 millimeters; Dunlop 2019); their multiple respiratory

BioScience 73: 112–123.
https://doi.org/10.1093/biosci/biac116

systems; their abilities to secrete silk using spigots and spinnerets, to construct complex foraging webs, and to secrete venoms composed of hundreds of protein compounds; and a range of behaviors from sociality to cannibalism (Foelix 2011). This broad phenotypic diversity may underlie why spiders are so diverse (over 50,000 species, World Spider Catalog 2022) when compared with closely related arachnid groups—for example, their sister group, whip spiders (about 260 species; Harvey 2003, Miranda et al. 2022). Despite invoking such consternation and curiosity among humans, our knowledge of spiders is not commensurate with their described diversity, which is greater than seven times that of all described mammals (Burgin et al. 2018). With a group of species so diverse, taxonomy has broad implications for studies in biodiversity, ecology, behavior, venom therapeutics, and biomaterials.

The World Spider Catalog (WSC) is a global taxonomic database with roots in the efforts of Bonnet (1945, 1955, 1956, 1957, 1958, 1959, 1961), Roewer (1942, 1955), Brignoli (1983), and Platnick and Brignoli (1989) to track descriptions of spider species throughout various periods in the history of spider taxonomy. The current WSC first came online in 2000 because of the work of Norman I. Platnick. Now managed by the Natural History Museum of Bern, this resource has grown to be the largest consistently updated taxonomic database for a single organismal group, although more broadly encompassing databases such as the World Register of Marine Species are larger, with just under 250,000 accepted species in the database (Miller et al. 2015, Vandepitte et al. 2018, World Spider Catalog 2022). In addition to being a robust interactive taxonomic resource, users with a WSC account may also access the extensive bibliographic database containing over 16,000 taxonomic publications, permitting greater accessibility to araneological research. Taxonomically, spiders are a unique group among Metazoa because their binomial naming by Clerck (1757) in *Svenska Spindlar* predates Linnaeus' (1758) *Systema Naturae*. Cataloging both the uniquely long history of spider species discovery and taxonomy and the breadth of spider diversity is of perennial importance to araneology because it expedites, makes more accessible, and standardizes our understanding of the group. Without a firm taxonomic foundation, it becomes impossible to evaluate biological discoveries in their most important context—that of evolution. Therefore, the WSC is a critical resource, because taxonomy serves as the basis of all other biological work on the group.

Various attempts have been made to estimate the extant spider diversity, both at the ordinal and familial level. Costello and colleagues (2012) used a nonhomogeneous renewal process model developed by Wilson and Costello (2005) to identify the number of species likely to be described. Their 95% upper confidence interval of species to be described by 2050 (46,700) has already been surpassed by a few thousand species, indicating that Wilson and Costello's (2005) estimate (52,920 species) is likely to be a very conservative estimate of species diversity (Costello et al. 2012). Agnarsson and colleagues (2013) took a combined approach involving averaging (pseudoscientific) estimates provided by taxonomic experts for the 15 most diverse spider families and included a blanket estimator for the rate of synonymies and invalid names across all spiders. As acknowledged by these authors, this metric is not a reliable estimate, because it is unlikely that species descriptions and other taxonomic changes are so evenly distributed across spiders such that the same magnitude of proportional diversity has been described in each family (Agnarsson et al. 2013). It is, rather, likely that families of larger or more charismatic species (e.g., tarantulas, Theraphosidae; jumping spiders, Salticidae) are more thoroughly documented than smaller, more cryptic families. This trend in favor of larger, more charismatic species is further evidenced by the fact that in a generally more popular taxon such as birds (Aves), the percentage of yearly increase in species is only about 0.9% per year from 2007 to 2016 (Sangster 2018). This has also been observed within invertebrates, such as parasitoid wasps, where there are biases in favor of charismatic species in the temperate Global North (Jones et al. 2009). There are also demonstrated biases both in biodiversity occurrence data, where occurrences are at the greatest deficit relative to species diversity in arachnids and insects (Troudet et al. 2017) and in conservation biology studies, where most work is done on vertebrates and charismatic invertebrates (e.g., Lepidoptera, Clark and May 2005).

Where many authors have investigated the rate of species documentation or quantified current described diversity to estimate total species diversity, new species "discovery" (or the scientific documentation of new species) has yet to be examined in the context of taxonomic effort. Understanding our implicit biases in the documentation of biodiversity and where the most critical gaps may lie is critical. Although past estimates of spider species diversity have primarily been concerned with broad estimates of total species diversity, there has been comparatively little attention paid to the influence of our own efforts as a scientific community on these estimates (Costello et al. 2012). In other words, what is the relative influence of taxonomic effort made to date on our estimates of species diversity? We view the concept of taxonomic effort as a lens for interpreting where this effort is lacking, as well as identifying biases in taxonomic effort along the axes of space, time, phylogenetic position, and even the lifetime of an individual specialist. The goal of understanding past taxonomic effort is to highlight taxonomic groups or geographic regions for which spiders (or other groups) require the most taxonomic focus and effort to close our biodiversity knowledge gap. For example, there is a general trend in scientific work, particularly that of systematic biology, to label an organism or system "understudied," but rarely is this done objectively. This research seeks to provide a metric to guide not only the research community in its future directions but also to substantiate claims about understudied groups.

We propose a novel way to assess the bias of taxonomic effort, using a ratio of new species descriptions to other

taxonomic activity as a proxy for taxonomic effort. In this study, we hypothesize that a higher number of species descriptions for some spider families relative to other taxonomic activity (such as synonymy, transfers, redescriptions, new records) is an indication of lower taxonomic effort to date and hypothesize that increased effort will yield a higher relative abundance of undocumented species. Conversely, we hypothesize that if the proportion of primary species descriptions relative to other taxonomic activities in a group is low, the asymptote of effort relative to species abundance has leveled, and relatively few undocumented extant species remain, meaning a high taxonomic effort has been made. This metric is dynamic, meaning that as more species are described or taxonomic changes are made, the metric can be recalculated. We seek to highlight historic trends among spider families and zoogeographic regions and to assess both the effect of newer technologies on species documentation and the effect of resource access on taxonomic research along socioeconomic divides. In proposing this metric, we seek a simple approach that can be easily applied to other groups and further propose that other taxonomic changes in a group only occur when the work of new species description (e.g., in a geographic area or a phylogenetic group) has reached a level wherein factors such as synonymies can begin to be recognized; that is, the fauna is no longer severely undocumented.

Methods: Leveraging data from the World Spider Catalog

We acquired data from the WSC team with an export date of 15 January 2022 in the comma-separated values (.csv) format. The following data were used: a spreadsheet containing a list of all taxonomic references (supplemental table S1), a spreadsheet containing a list of all valid species (supplemental table S2), and a spreadsheet with all references in the WSC bibliography (supplemental table S3). The first two spreadsheets were used to count the number of species in each family and the number of total taxonomic activities in the family. Total taxonomic activities were defined as all taxonomic references listed under the taxonomic references field of a species page on the WSC. Other taxonomic activities included all taxonomic references except the original species description and were transformed from total taxonomic activities by subtracting the number of valid species in each family. The ratio of new species descriptions to other taxonomic activity were log-transformed to reduce the weight of spider families that are dramatically more diverse than others on such a metric and so that positive values indicated species that have a higher amount of species descriptions relative to other taxonomic activities and the converse for negative values. The list with all references in the WSC bibliography was used to detail the number of references held and made accessible by the database.

Taxonomic effort metric. In the present article, we define taxonomic effort from a mechanistic perspective: the total number of taxonomic treatments in the scientific literature at a particular point in time for any given taxon, including both protologue (new species descriptions) and nonprotologue taxonomic work (such as designating synonyms, describing allotypes, new combinations). *Other taxonomic activities* is a broad term, which can be described best as any taxonomic action that does not result in a currently valid species description. This may include any synonymies, new combinations, redescriptions, or other taxonomic works that are not a protologue. Broadly speaking, this includes all texts except the protologue in the taxonomic references section of a species page of the WSC.

If, for a given spider group (e.g., clade, family, or genus), there is a greater amount of species descriptions relative to other taxonomic activity, it may indicate that additional taxonomic effort may yield more species. Conversely, groups with more other taxonomic activity relative to new species descriptions may be reaching a point of saturation for taxonomic effort, such that additional taxonomic work on the group isn't going to result in the description of new species but, rather, subsequent changes that are often the result of revisionary taxonomy. This ratio has the benefit that it can be assessed for each defined group rather than an evenly applied ad hoc estimate across all spider families.

Where many authors have investigated the rate of species documentation or quantified current described diversity to estimate total species diversity, new species "discovery" (or the scientific documentation of new species) has yet to be examined in the context of taxonomic effort—a gap our novel approach seeks to close. This metric was tabulated and then excluded families with less than ten species (Archoleptonetidae, Austrochilidae, Cithaeronidae, Ctenizidae, Hexurellidae, Homalonychidae, Huttoniidae, Mecicobothriidae, Megadictynidae, Megahexuridae, Microhexuridae, Myrmecicultoridae, Penestomidae, Periegopidae, Porrhothelidae, and Trogloraptoridae) to reduce the effects of singleton or low-diversity families and are tabulated in the supplemental material (supplemental tables S4 and S5).

Geographic assignment. To evaluate a potential effect of geographic bias on taxonomic effort, we assigned all 50,000 species to a zoogeographic region. We processed the distribution string in the WSC spreadsheet of all valid species so it could be standardized across the following defined zoogeographic regions (Holt et al. 2013, Morrone 2014, 2015): Andean, Australian, Ethiopian, Malagasy, Melanesian, Micronesian, Nearctic, Neotropical, Novozelandic, Oriental, Palearctic, Polynesian, Subantarctic, and unknown. The authors would like to acknowledge that the colonial legacy and implicit racism of some of these regions, both in name and delineation, is an urgent issue that warrants discussion and consensus within the research community, but that discussion is beyond the scope of this article.

The initial classification applied to this data set involved employing the use of the VLOOKUP function in Excel

and replacing country names with those of zoogeographic regions. Because of the coarse regional scale we were working in, countries that encompassed two zoogeographic regions were treated as being part of both regions (e.g., for China, we used both Palearctic and Oriental). It is beyond the scope of this study to assign zoogeographic regions to 50,000 individual spider species beyond what is present in the WSC distribution string, and such specificity may not be known, especially for species described over a century ago. Following this, a filter was applied to provide additional data cleanup steps as there were multiple strings that included different combinations (e.g., Palearctic, Palearctic, Oriental, where it should say Palearctic, Oriental), as well as for species that VLOOKUP could not automatically assign (approximately 14,000 species or 30% of the data).

For the remaining species, a pivot table was used to group the unassigned species by the distribution string so the proper zoogeographic region could be assigned more expeditiously. Once this was complete, a spreadsheet was exported for each zoogeographic region for further analysis (supplemental tables S6–S20).

An anticipated bias due to these methods includes the fact that some countries contain multiple zoogeographic regions. To rectify this, we classified all families recorded from countries with multiple zoogeographic regions as being present in all those zoogeographic regions. Although this is unlikely true for all species and may contribute to double counting in the respective zoogeographic regions, for many species, it is likely not explicitly known whether they are restricted only to one zoogeographic region or multiple within the country. This is especially true for species described early on in spider taxonomy, where localities were not as detailed as they presently are. A coarse zoogeographic scale was selected because we are looking at high-level taxonomic trends by region rather than asking questions at any finer ecological or taxonomic scale. Although these regions are flawed in that some regions are double counted because of countries that occur in multiple zoogeographic regions being counted twice and are further biased by the colonial history of these terms, these are more relevant than strict geopolitical boundaries of states or provinces, countries, and continents. The inherently political nature of these boundaries also allowed us to explore trends in geographic bias because of colonialism, especially with respect to disparities among regions primarily in the Global North (e.g., Nearctic, Palearctic, Australian, Novozelandic) versus those primarily in the Global South (e.g., Neotropical, Ethiopian, Oriental). Introduced species were considered as belonging only to their original range or what is thought to be their original range.

Estimates of species diversity by family. Using the taxonomic effort metric, we applied the ratio of new species descriptions relative to other taxonomic changes to provide estimates of species diversity by family. This was completed using the following formula: $ES = n + (n \times r)$, where ES is the species diversity estimate, n is the number of species currently in the family, and r is the taxonomic effort ratio described in this article. The version of the ratio not log-transformed was used to avoid negative values that would deflate species diversity and the assumption in the present article is that for each family the species diversity is additive.

Comparisons to prior work. To investigate and make comparisons between Agnarsson and colleagues (2013), we had to address three types of changes that have occurred since the publication of Agnarsson and colleagues (2013) at the family level (Wunderlich 2008, Bond et al. 2012, Griswold et al. 2012, Fernández et al. 2014, 2018, Ramírez 2014, Polotow et al. 2015, Dimitrov et al. 2017, Wheeler et al. 2017, Godwin et al. 2018, Hedin et al. 2018, 2019, Ono and Ogata 2018, Ramírez et al. 2019, Kulkarni et al. 2020, 2021, Opatova et al. 2020, Ledford et al. 2021, Azevedo et al. 2022, Montes de Oca et al. 2022). Either families were split, families were merged, or new families were described. For families that were split after Agnarsson and colleagues (2013), we maintained the family level status at that time and combined our metrics for the families that were split so they were in line with what it currently is (supplemental table S21). For families that were combined after Agnarsson and colleagues (2013), we combined the metrics from Agnarsson and colleagues (2013) so that they could translate to the current composition of the family. For families newly described, they were excluded from this part of the analysis.

Counts were made to examine the amount of species representation in a database by group (Kallal et al. 2021). For ease of examination, these clades and the UDOH (Uloboridae, Deinopidae, Oecobiidae, Hersiliidae) grade were used, but a table with all families is available in the supplemental material (supplemental tables S22–S24). This includes the count of species represented by nucleotide sequences on GenBank (representing primarily single-locus Sanger reads), the number of species represented within each family by the Sequence Read Archive, which includes genomic and transcriptomic data, and the number of species represented within each family by the World Spider Trait Database (Pekár et al. 2021). These represent a variety of information and can help contribute to an understanding of what data is present for what share of species. The reconciliation of spider family names with that of Agnarsson and colleagues (2013) as described above was important in the database examinations as well.

Representation in nontaxonomic databases. We counted the number of species represented by three databases for each spider family: the GenBank Nucleotide (primarily single-locus data generated via Sanger sequencing; acquired 10 April 2022), the GenBank Sequence Read Archive (primarily genomic and transcriptomic data via parallel sequencing; acquired 10 April 2022), and the World Spider Trait Database (containing data about anatomy, biomechanics,

communication, cytology, defense, ecology, life history, morphology, morphometrics, predation, and reproduction; acquired 13 April 2022). These data were tabulated by percentages (supplemental table S22) and then adjusted because some families had species counts that exceeded the described diversity (because of the inclusion of data from undescribed or unidentified specimens).

Landmarks in spider taxonomy. We selected early publications highlighting landmark developments in taxonomic technology, such as one of the earliest papers to include a genitalia illustration, one of the earliest papers to include a photograph in a species description, one of the earliest papers to include a scanning electron microscope photograph alongside a species description, one of the earliest papers to include molecular data alongside a species description. These were identified through a combination of searching references chronologically on the WSC and finding candidate papers via Google Scholar.

Data analysis and visualization. The data were analyzed and visualized using RStudio and the packages dplyr, forcats, ggplot2, scales, phylotools, and phytools (Revell 2012, Zhang et al. 2017, Wickham 2016, 2021, 2022, RStudio Team 2022, Wickham et al. 2022). Figure 1 was created using the contMap function in phytools, which plots the reconstructed ancestral state of a continuous character along a tree on the basis of data at the tips, the data in this article being the ratio of taxonomic effort (Revell 2012).

Results: Employing the taxonomic effort metric

Taxonomic bias is evident in a range of new species descriptions relative to other taxonomic activity ratios (figure 1). However, there is some evidence of a phylogenetic bias in taxonomic effort (figure 2). The Synspermiata and Mygalomorphae clades have slightly higher ratios of new species descriptions relative to other taxonomic changes. The Lycosoidea has slightly less species descriptions relative to other taxonomic activity. We do not recover evidence of geographic bias, although it is important to note that the majority of spider families (approximately 75%) are distributed in the Global South (figure 1). It appears that many families that are primarily found in the Global North are at the top of this metric and warrant further species discovery research. However, see below for caveats of using this metric to define regions in which research should be prioritized.

Estimates of species diversity by family. When summing the estimates by Agnarsson and colleagues (2013), the total estimated spider diversity is 120,333, whereas the total spider diversity estimated from the metric of taxonomic effort is 83,089 (supplemental table S25). The top five most speciose families using the species estimation metric based on taxonomic effort described here are Salticidae (8,872), Linyphiidae (6,195), Oonopidae (5,934), Araneidae (3,952), and the Pholcidae (3,756). In terms of the difference between the species estimated on the basis of taxonomic effort and the number of currently described species, the top five families with the highest number of estimated new species are Oonopidae (4,046 more species), Salticidae (2,479), Pholcidae (1,907), Zodariidae (1,483), and the Linyphiidae (1,474). As with the original taxonomic effort metric, the families with less than 10 currently described species were also excluded. For estimates of every spider family, please see table S25.

World Spider Catalog data and geographic assignment. The zoogeographic regions with the greatest accumulation of new species described to date are the Palearctic, the Neotropical, and the Oriental (figure 3). The Australian and Nearctic regions, although they have similar levels of new species descriptions through time, have differently sloped curves, suggesting an asymptote in Nearctic species but an exponential increase in Australian (figure 3). The Ethiopian region is the fourth most speciose zoogeographic region and shows a constant increase in new species descriptions through time. Other regions, particularly many regions in the Global South, exhibit a plateau (figure 3) occurring between 1900 and 1950.

Representation in nontaxonomic databases. We did not find evidence of a phylogenic correlation in terms of representation in nontaxonomic databases (figure 4). The most well-documented database is the GenBank Nucleotide database. The GenBank Sequence Read Archive is still definitely catching up with the adoption of parallel sequencing technology. The World Spider Trait database closely matches the GenBank Nucleotide database proportions. Aside from groups with two species or less, the largest proportional representation for the GenBank Nucleotide Database is the Avicularioidea (100%, but see below). The smallest proportional representation for the GenBank Nucleotide database is the UDOH grade, with only 14.4% of species represented. Excluding groups with two species or fewer, the largest proportional representation for the GenBank Sequence Read Archive is the Atypoidea. Excluding groups with two species or fewer, the smallest proportional representation for the GenBank Sequence Read Archive is Dionycha clade A, with 0.2% of species represented. Excluding groups with two species or fewer, the largest proportional representation for the World Spider Trait Database is Eresidae (100%, but see below). Excluding groups with two species or fewer, the smallest proportional representation for the World Spider Trait Database is Sparassidae (10.6%).

Modernizing approaches to taxonomy in the age of big data

This study presents a novel way of assessing taxonomic effort and allows for a quantification of the degree to which taxonomic groups have been understudied. Within spider families, we have done this by capitalizing on the publicly available WSC and examining spider diversity and taxonomic effort of spider families. If more species have been described relative to the other taxonomic changes,

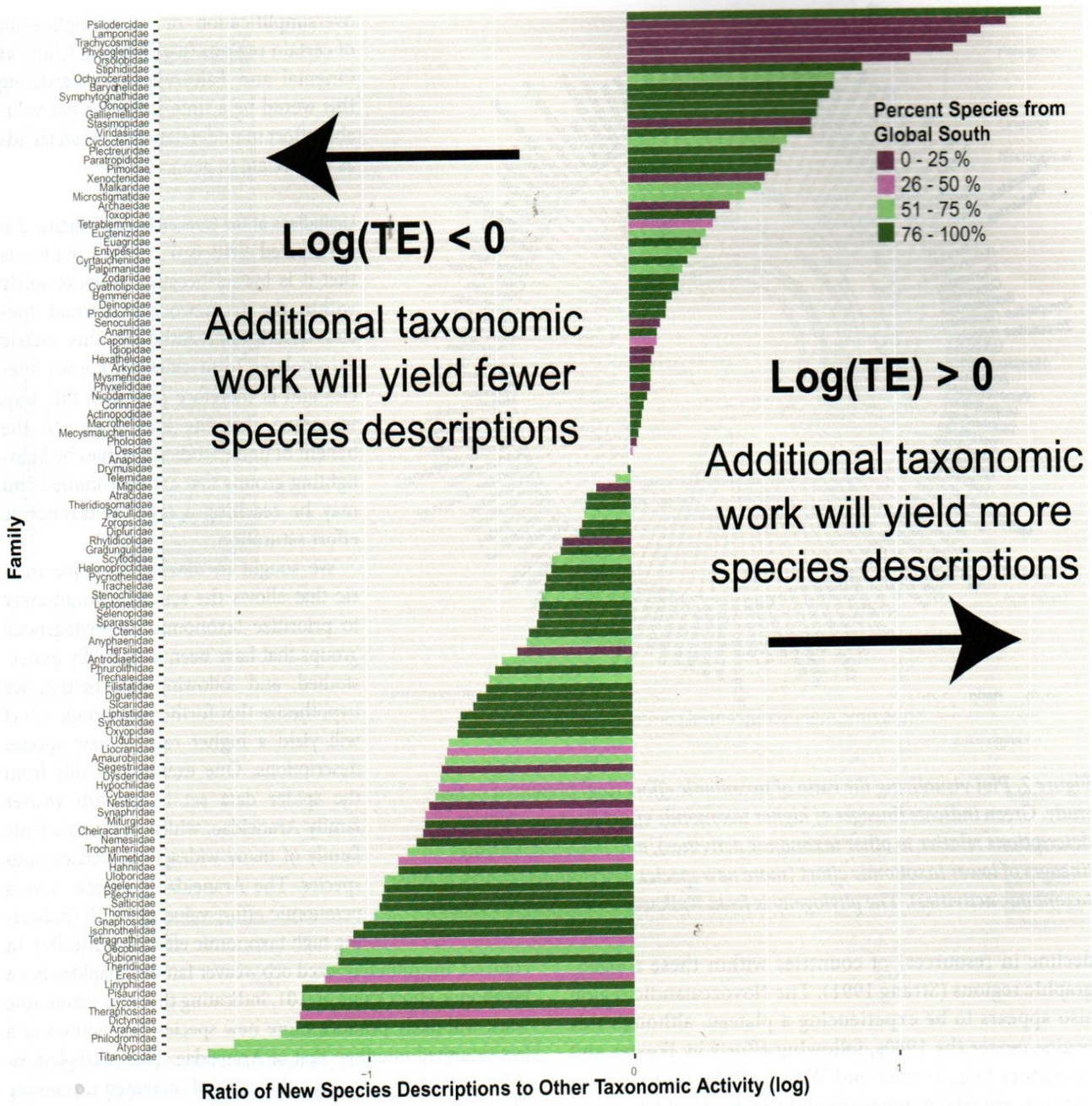

Figure 1. All spider families included in taxonomic effort metric, with ratio of new species descriptions relative to other taxonomic activity (log) by family on x-axis and all families included in y-axis. Families colored by socioeconomic region assigned from zoogeographic region, with families distributed primarily in the Global South (as defined by socioeconomic and political characteristics) shown in green, and families distributed primarily in the Global North shown in magenta.

we view that as indicative that further taxonomic effort will continue to yield more new species before or in synchrony with other taxonomic changes (e.g., a monographic or phylogenetic work on the group may yield a combination of new species and other taxonomic changes). The opposite, fewer new species relative to other taxonomic changes, may indicate that the family is reaching saturation in species discovery and future work may result in more nonprotologue work. It is important to note that we view nonprotologue taxonomic work and other taxonomic changes or revisionary taxonomy as equally important in our understanding and framing of the classification and evolution of life on Earth.

On the basis of the species accumulation curve partitioned by zoogeographic region (figure 3), several regions in the Global South appear to be in a plateau of new species descriptions (Polynesian, Melanesian, Subantarctic, Micronesian). The Andean, Australian, and Oriental zoogeographic regions, despite now having an increase in species descriptions, also experienced a plateau. This plateau ranges from 75 to 120 years ago and corresponds with the decolonization (and subsequent

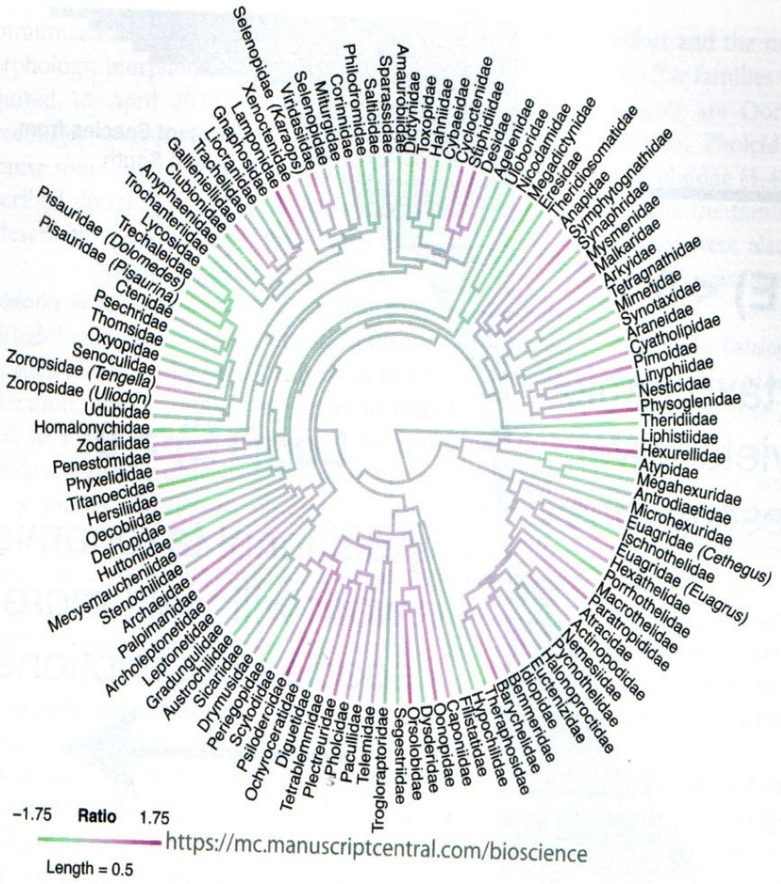

Figure 2. Plot visualizing the ratio of taxonomic effort (log) presented in this study. Green indicate lineages of higher taxonomic effort (less new species descriptions relative to other taxonomic activities), and magenta indicates lineages of lower taxonomic effort (more new species descriptions relative to other taxonomic activities). The phylogeny is from Kulkarni and colleagues (2023).

decline in resources) of countries within these zoogeographic regions (Strang 1991). The Novozealandic region also appears to be experiencing a plateau, although this begins nearer the 1980s, following efforts by Forster and colleagues (e.g., Forster and Wilton 1973). Of particular concern are island zoogeographic regions, which, given their rate of extinction, are of paramount importance for documenting endemic diversity (Fernández-Palacios et al. 2021). Islands are also tractable for complete documentation, relative to other species-rich regions exhibiting exponential (or other growth) in species descriptions (Coddington et al. 2009).

We found that the coarse nature of zoogeography boundaries obscures some trends of colonialism, racism, and exclusion on taxonomy. For example, although Iran and Iraq are considered part of the Global South from a socioeconomic and political perspective, they are zoogeographically Palearctic. An improvement to the WSC (and other biotic catalogs) would be the inclusion of distribution data beyond a simple, nonstandardized, distribution string. The distribution string, although it is flexible for the variety of information it may contain, obfuscates broad zoogeographic or biogeographic trends and results in both the oversimplification or overcomplication of certain regions (e.g., treating China as Oriental and Palearctic). Standardizing this would be a time-intensive but valuable effort in understanding broad trends as they relate to geography.

Limitations of the taxonomic effort metric. An anticipated criticism of this metric is that it is biased because it works solely within the framework of formal species description. Although this metric extrapolates from existing, known species and is therefore biased in this way, we argue that the bias works to the benefit of understudied lineages by highlighting groups that are well studied and may be reaching a level of taxonomic effort saturation.

We sought to develop a simple metric that allows the scientific community to prioritize taxonomic or phylogenetic groups that have been historically understudied, and following this metric, we hypothesize that further taxonomic effort will yield a higher rate of new species descriptions. One example of this from the spider data set is the orb weaver family Araneidae, which is a charismatic family of many widespread, conspicuous species. The Araneidae currently have a taxonomic effort value of −1.27 (indicating high taxonomic effort historically). In contrast, the minutely sized orb weaver family Anapidae has a taxonomic effort value of 0.01, indicating that new taxonomic work will likely produce more new species descriptions at a less saturated rate than that of Araneidae. This example provides tangible evidence for the value of increased taxonomic effort in the latter group.

Although this metric may be powerful for examining bias by clade or taxonomic group and for predicting species diversity, it has limitations and may not be useful for examining trends by zoogeographic or socioeconomic regions. We found that using this metric to examine trends socioeconomically (Global North versus Global South), many families that warrant further species description by this metric are primarily Global North in distribution (figure 1). Conversely, the species descriptions through time plots show that many of the zoogeographic regions in the Global South had a sharp decline in rate beginning with the period when colonized countries gained independence, and for many regions, this plateaued rate has persisted to the present day (figure 3). Given the fact that many of these regions hold a disproportionately high amount of global biodiversity, our interpretation is that the taxonomic effort ratio cannot be used to assess geographic regions because of the deeply

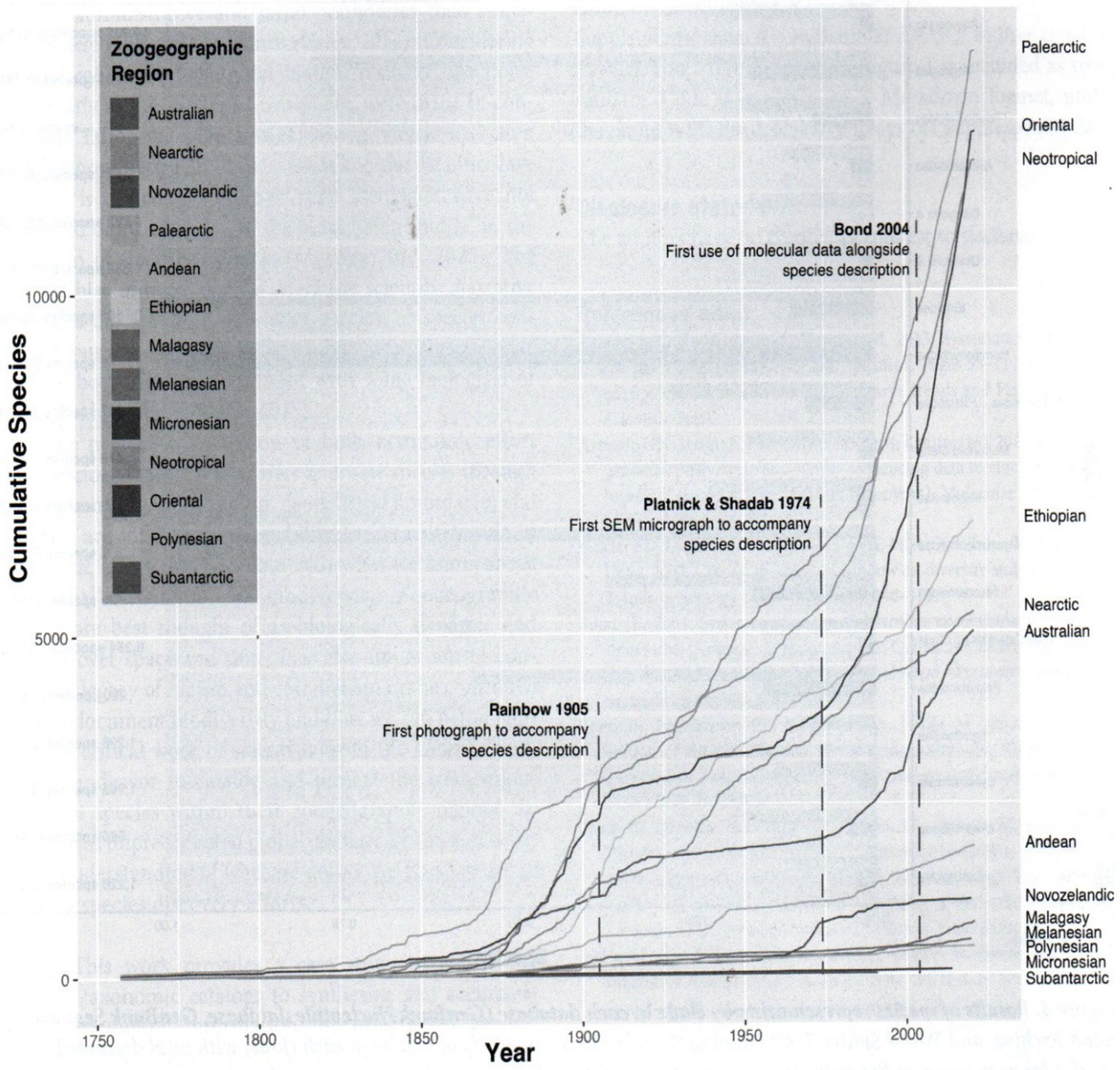

Figure 3. Spider species described through time, illustrated by zoogeographic region, from 1757 through to 2022. Notable events are depicted with dotted lines.

interwoven colonial history of taxonomy and the impact of disproportionate resources available in these regions before and after the end of colonialism.

Implications for araneology. Our estimate of species diversity by family using the taxonomic effort metric are more conservative than those proposed by Agnarsson and colleagues (2013), with about 40,000 fewer species proposed when the species diversity estimates per family are summed. Furthermore, estimating the diversity of each family using the metric based on the history of taxonomic work in that family. Including the other taxonomic changes in this metric provides a more accurate estimate. In some cases, these estimates are consistent and very close to the estimates proposed by Agnarsson and colleagues (2013), such as the Pholcidae (Agnarsson et al. 2013 proposed 3,631 species, whereas this metric proposes 3,756). In other cases, such as the Hexathelidae, Agnarsson and colleagues (2013) proposed 300 species, whereas this metric proposes 94. Our results indicate that there is some phylogenetic trend regarding taxonomic effort, in that there are clades or families that are proportionally overstudied (Lycosoidea, Eresoidea, and some parts of the Araneoidea, and the Theraphosidae) and some that warrant increased taxonomic effort (e.g., Psilodercidae, Lamponidae, Trachycosmidae, Physoglenidae, Orsolobidae, Synspermiata, Mygalomorphae generally). It is likely that future taxonomic effort for any spider family will yield new species, but this metric is intended to underscore those families that will result in maximal new species description from added taxonomic effort.

Some other notable discoveries from our research relate to the impact of taxonomic tools in arachnological

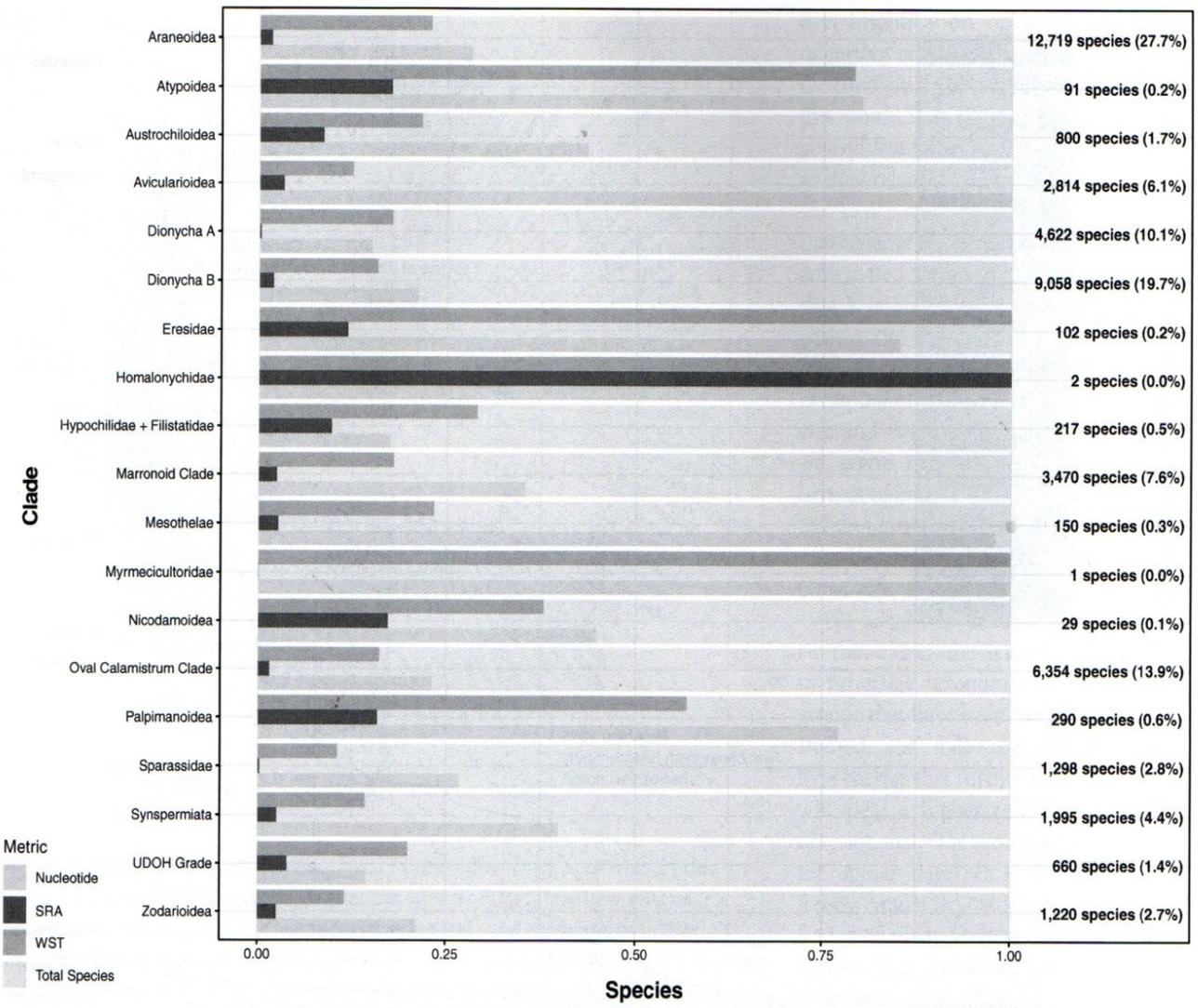

Figure 4. Results of species representation by clade in each database (GenBank Nucleotide database, GenBank Sequence Read Archive, and World Spider Trait Database), scaled to the diversity of species in each clade, with total described species for each group to the right, and the proportion of described spider diversity in parentheses.

research. Clerck's (1757) *Svenska Spindlar*, the first taxonomic work in araneology, was also the first to include an illustration of genitalia, demonstrating that even at the very beginning of spider taxonomy, it was understood that the genitalia of spiders were valuable for delimiting species. The first photograph in a taxonomic work was published by Rainbow (1905) and is, in fact, not a photograph of the spider itself (*Badumna socialis*; Rainbow 1905) but of the web structures, highlighting that web structure has long been viewed as an important characteristic in spider taxonomy. The first paper to include a scanning electron microscope micrograph in a species description was in 1974, and this technique has been widely used in the subsequent years (Platnick and Shadab 1974). The first paper to include molecular data alongside a species description was in 2004 (Bond 2004).

In exploring other databases, some taxa (e.g., Avicularioidea, Eresidae) had 100% representation relative to the current described species. This is in part because these databases allow nonspecific assignments (representing undescribed species or undetermined specimens, e.g., *Araneus* sp. A) that inflates the total number of species for these taxa. For example, the Homanolychidae had the highest proportional representation in each database (100% representation), but this family includes only two species. The same is true for the Myrmecicultoridae in the GenBank and the World Spider Trait Database. As such, these database metrics are probably overestimates of the species representation rather than underestimates, meaning for most groups there is still significant gaps in our knowledge beyond taxonomy.

Where zoogeographic regions inform the natural delineations of biodiversity, there are significant geopolitical consequences of the history of colonialism in the Global South and its detrimental effects (Holt et al. 2013, Morrone 2014, 2015, Gorneau et al. 2022). For example, delineations between the Nearctic and Neotropical regions during the nineteenth century relate more directly to that of British

colonial administrative and naval boundaries than something solely biological in nature (Greer 2015). Considering that a large amount of initial natural history collections, particularly in the Western hemisphere, were made in concert with European colonialist expansion, violent exclusion of Indigenous people from their lands and the Atlantic slave trade, it is difficult to imagine such ideologies were not impressed on or extended to the biota being studied in the formation of these boundaries (Murphy 2013, 2020). The political bias of these regions alters the scientific narrative of species distributions, migratory species, or species considered vagrant, because it may be that organisms crossing political boundaries are conflated with some biological or geological hypothesis (Greer 2015).

In the present examination of both taxonomic effort along socioeconomic boundaries and new species through time by zoogeographic region, geopolitical boundaries and histories are influencing our ability to observe trends in taxonomic effort that correlate with what we know about the global distribution of biodiversity. Zoogeographic regions are best thought of as biologically dynamic and variable over space and time. It is also important to consider the legacy of racism and colonialism in the continued work to document biodiversity and how we can better support the critical work of scientists from the Global South as they endeavor to catalog and protect the tremendous wealth of species within their zoogeographic regions. In this time of unprecedented global change, we must develop a firm understanding of what we do not yet know to aid in focusing species discovery efforts.

Closing. This work provides a case study regarding the value of taxonomic catalogs to synthesize and accurately record taxonomic information for organismal groups. Biodiversity catalogs aid in species discovery, provide open access taxonomic information, and allow for inferences to be made about global and phylogenetic trends in taxonomic effort. The advancement of such resources is only going to improve the state of biological research and equip the future generation of taxonomists in their effort to document global biodiversity. This proposed taxonomic effort metric is widely applicable to organisms for which a biodiversity catalog is available. In the face of global biodiversity decline, it is imperative that we identify ways to prioritize species discovery research in taxonomy to document biodiversity before it goes extinct. We encourage biodiversity scientists to support taxonomic catalogs for their own organismal groups and to mine taxonomic data to better inform future investments in research.

Acknowledgments

This research was supported by a National Science Foundation grant no. DEB-2026623 to Sarah C. Crews and LAE. The authors would also like to thank the World Spider Catalog team for providing the data sets from the database and Lyra Wallace for assistance with spreadsheet manipulation.

Supplemental material

Supplemental data are available at *BIOSCI* online as tables S1–S24, and a tree file used for figure 1 is included as tree_ratio.tre. Code is available, in R Markdown format, under the name WSC.Rmd, and it employs the supplemental data.

Disclosure statement

The authors have no conflict of interest to declare.

References cited

Agnarsson I, Coddington JA, Kuntner M. 2013. Systematics: Progress in the study of spider diversity and evolution. Pages 58–111 in Penney D, ed. Spider Research in the 21st Century: Trends and Perspectives. Siri Scientific Press.

Azevedo GH, Bougie T, Carboni M, Hedin M, Ramírez MJ. 2022. Combining genomic, phenotypic and Sanger sequencing data to elucidate the phylogeny of the two-clawed spiders (Dionycha). Molecular Phylogenetics and Evolution 166: 107327.

Bickford D, Lohman DJ, Sodhi NS, Ng PK, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution 22: 148–155.

Bond JE. 2004. Systematics of the Californian euctenizine spider genus *Apomastus* (Araneae: Mygalomorphae: Cyrtaucheniidae): The relationship between molecular and morphological taxonomy. Invertebrate Systematics 18: 361–376.

Bond JE, Hendrixson BE, Hamilton CA, Hedin M. 2012. A reconsideration of the classification of the spider infraorder Mygalomorphae (Arachnida: Araneae) based on three nuclear genes and morphology. PLOS ONE 7: e38753.

Bond JE, Godwin RL, Colby JD, Newton LG, Zahnle XJ, Agnarsson I, Hamilton CA, Kuntner M. 2021. Improving taxonomic practices and enhancing its extensibility: An example from araneology. Diversity 14: 5.

Bonnet P. 1945. Bibliographia Araneorum: Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. I. Douladoure.

Bonnet P. 1955. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. II A–B: Systématique des Araignées (Étude par Ordre Alphabétique). Douladoure.

Bonnet P. 1956. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. II C–F. Systématique des Araignées (Étude par Ordre Alphabétique). Douladoure.

Bonnet P. 1957. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. II G–M. Systématique des Araignées (Étude par Ordre Alphabétique). Douladoure.

Bonnet P. 1958. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. II N–S. Systématique des Araignées (Étude par Ordre Alphabétique). Douladoure.

Bonnet P. 1959. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. II T–Z. Systématique des Araignées (Étude par Ordre Alphabétique). Douladoure.

Bonnet P. 1961. Bibliographia Araneorum. Analyse Méthodique de Toute la Littérature Aranéologique jusqu'en 1939, vol. III. Index Alphabétiques, Résultats, Conclusions, Considérations Diverses. Douladoure.

Brignoli PM. 1983. A Catalogue of the Araneae Described between 1940 and 1981. Manchester University Press.

Burgin CJ, Colella JP, Kahn PL, Upham NS. 2018. How many species of mammals are there? Journal of Mammalogy 99: 1–14.

Choi SC. 2016. Methods for delimiting species via population genetics and phylogenetics using genotype data. Genes and Genomics 38: 905–915.

Clark JA, May RM. 2002. Taxonomic bias in conservation research. Science 297: 191–192.

Clerck C. 1757. Svenska Spindlar: Uti sina hufvud-slågter indelte samt under några och sextio särskildte arter beskrefne och med illuminerade figurer uplyste. Laurentius Salvius.