Integrative taxonomy of araneomorph spiders: Breathing new life into an old science

Rivera Quiroz, F.A.

Chapter 2

Mining data from legacy taxonomic literature and application for sampling spiders of the *Teutamus* group (Araneae; Liocranidae) in Southeast Asia

F. ANDRES RIVERA-QUIROZ*[1, 2], **BOOPPA PETCHARAD**[3] and **JEREMY A. MILLER**[1,4]

[1] Understanding Evolution group, Naturalis Biodiversity Center, Darwinweg 2, 2333CR Leiden, The Netherlands

[2] Institute for Biology Leiden (IBL), Leiden University, Sylviusweg 72, 2333BE Leiden, The Netherlands.

[3] Faculty of Science and Technology, Thammasat University, Rangsit, Pathum Thani, 12121 Thailand.

[4] Plazi, Zinggstrasse 16, CH 3007, Bern, Switzerland.

***Scientific Reports* (10:15787).**

doi:10.1038/s41598-020-72549-8

Abstract

Taxonomic literature contains information about virtually every known species on Earth. In many cases, all that is known about a taxon is contained in this kind of literature, particularly for the most diverse and understudied groups. Taxonomic publications in the aggregate have documented a vast amount of specimen data. Among other things, these data constitute evidence of the existence of a particular taxon within a spatial and temporal context. When knowledge about a particular taxonomic group is rudimentary, investigators motivated to contribute new knowledge can use legacy records to guide them in their search for new specimens in the field. However, these legacy data are in the form of unstructured text, making it difficult to extract and analyze without a human interpreter. Here, we used a combination of semi-automatic tools to extract and categorize specimen data from taxonomic literature of one family of ground spiders (Liocranidae). We tested the application of these data on fieldwork optimization, using the relative abundance of adult specimens reported in literature as a proxy to find the best times and places for collecting the species (*Teutamus politus*) and its relatives (*Teutamus* group, TG) within Southeast Asia. Based on these analyses we decided to collect in three provinces in Thailand during the months of June and August. With our approach, we were able to collect more specimens of *T. politus* (188 specimens, 95 adults) than all the previous records in literature combined (102 specimens). Our approach was also effective for sampling other representatives of the TG, yielding at least one representative of every TG genus previously reported for Thailand. In total, our samples contributed 231 specimens (134 adults) to the 351 specimens previously reported in the literature for this country. Our results exemplify one application of mined literature data that allows investigators to more efficiently allocate effort and resources for the study of neglected, endangered, or interesting taxa and geographic areas. Furthermore, the integrative workflow demonstrated here shares specimen data with global online resources like Plazi and GBIF, meaning that others can freely reuse these data and contribute to them in the future. The contributions of the present study represent an increase of more than 35% on the taxonomic coverage of the TG in GBIF based on the number of species. Also, our extracted data represents 72% of the occurrences now available through GBIF for the TG and more than 85% of occurrences of *T. politus*. Taxonomic literature is a key source of undigitized biodiversity data for taxonomic groups that are underrepresented in the current biodiversity data sphere. Mobilizing these data is key to understanding and protecting some of the less well-known domains of biodiversity.

Introduction

In the aggregate, traditional taxonomic publications can be thought of as a repository that has accumulated vast amounts of biological data linked to specific taxonomic names. These units of taxonomic knowledge, information linked to a name within a publication, are known as taxonomic treatments [1–3]. This makes taxonomic literature not only crucial for the exchange and growth of biodiversity knowledge, but also capable of being used to detect and understand larger biodiversity patterns with historical perspective.

In recent years, great efforts have gone into the digitization of legacy taxonomic literature [4–6]. This combined with digital publications have greatly improved access to taxonomic literature. Nevertheless, although easy to share, PDF publications still have most biodiversity data embedded in strings of text making them less dynamic and difficult or impossible to read and analyze without a human interpreter [7]. This difficulty to access and use core specimen data is what we define as PDF prison [8]. Recently developed tools allow text in PDF documents to be interpreted and categorized in XML format (mark-up) allowing information to be mobilized, aggregated and reanalyzed [9–12]. Plazi Treatment Bank [8,13,14], is a project dedicated to creating a comprehensive compendium of taxonomic and biological data extracted from primary literature [15]. This platform permits mined treatment data to be accessed, queried, compared, and reused in a customized way. The strategy for data extraction can be prospective: where journals generate new data in XML format that can be uploaded directly to repositories (as has been implemented by Zookeys [2] and EJT [8,13]). or retrospective: where data is mined from legacy taxonomic literature [3,11–13] through a process called semantic enhancement [9,13]. This retrospective approach is more complicated and time consuming since the semi-automatic process of text recognition and tagging needs to be checked by a human operator [3,15]. However, it can provide useful information by extracting, integrating and using biodiversity data contained in the hundreds of years of accumulated taxonomic literature. Data integration is achieved by linking records from Plazi treatment bank to the Global Biodiversity Information Facility (GBIF) [8,16] where they are aggregated with other type of records, mainly natural history institution specimen collections and observation data based on GBIF's taxonomic backbone [17].

Here we combined several of these cybertaxonomic tools to test the data extraction process and its potential application on the design and planning of an expedition to collect fresh material in the field. We targeted the ground spider *Teutamus politus* Thorell 1890 and its relatives from the so called *Teutamus* group (TG) (Araneae, Liocranidae) [18]. This group of spiders is mostly distributed in Southeast Asia [19–23] and is composed of seven genera: *Jacaena*, *Koppe*, *Oedignatha*, *Sesieutes*, *Sphingius*, *Sudharmia* and *Teutamus* [18]. These spiders have been cataloged in the family Liocranidae; however, their phylogenetic relationships, biology and evolution are still poorly understood[18,24]. Therefore, collection of fresh specimens of the target taxa was necessary

for building a molecular phylogeny of the TG. The species *T. politus*, besides being the type species of the genus *Teutamus*, is an example of the extremely rare phenomenon of directional genital asymmetry [25]. For this reason, the collection of live adult specimens was crucial to study, document, and test the behavioral implications of their abnormal genital morphology.

Our study aimed to highlight the importance of making biodiversity data contained within taxonomic treatments accessible and reusable in accordance with the FAIR data principles [26]. This approach can help bridge gaps and focus efforts in the study of particularly interesting taxa or geographic regions. The usability of taxonomic literature data, potential applications, and its limitations and biases are discussed.

Material and Methods

Literature data extraction– We accessed all taxonomic literature of the family Liocranidae available in the World Spider Catalog [27]. We selected 55 publications that contained taxonomic treatments of the family Liocranidae [19–23,28–80] (for full list, see Supplementary Table 1). We selected and processed all publications that provided taxonomic treatments with specimen data and usable geographical references. Publications written in a language other than English were not processed since OCR parsing, as implemented by the programs used here, has mostly been developed in this language. From the marked-up documents, 21 contained information on members of the TG and two on the species *T. politus*. We used the program GoldenGATE Imagine V.3 (GGI; http://plazi.org/resources/treatmentbank/goldengate-editor/) to semantically enhance PDF documents, allowing atomization and categorization of data. In some cases, ABBYY FineReader V. 11 was used first to extract and correct text from the PDF document using optical character recognition (ORC) and text editing functions. Once the PDF documents were marked and revised, we used GoldenGATE to upload the files to Plazi's TreatmentBank [14].

Data analysis– We used Plazi Treatment Collection Statistics tool (http://tb.plazi.org/GgServer/srsStats) to download all the information relevant to our study in an excel spreadsheet to facilitate fine-grained management and analysis, largely following the approach described by Miller et al. [12]. We used these specimen based data to create profiles of the TG species allowing us to visualize where and when these taxa had been collected. Also, we used the GBIF occurrence search tool (https://www.gbif.org/occurrence/search) to look for records on our relevant TG taxa. The specific datasets we used can be found in the Data Accessibility section.

Site selection– Literature data were used to design our field collection in a way that allowed us to optimize the collection of adult specimens of our target taxa in Southeast Asia (SEA). We explored the number of specimens of the TG reported per country, province and location whenever possible. We favored those locations with a higher

representation of genera from the TG but also those where *T. politus* had been reported. Finally, we analyzed the total number of adult specimens collected per month for both the TG species and *T. politus* in order to increase the chances of finding adult spiders. Based on this, we decided to sample in three provinces in Thailand between July 16 and August 12, 2018.

Sampling– Following the results of our literature analysis, we prioritized collections in national parks and protected areas. Precise geographical coordinates and specific habitat information was scarce or missing altogether in most taxonomic treatments. Therefore, we further divided each site in four different vegetation types (collecting sites details in the Supplementary Table 2) allowing us to cover a wide range of available habitats. We combined pitfall traps, Winkler extractors (for soil arthropods; www.entowinkler.at), and direct collecting targeting ground spiders. A mixture of propylene-glycol and ethanol was used in the pitfalls to avoid excessive evaporation and help with DNA preservation [81]; all specimens were collected and stored in 96% ethanol. All liocranid spiders were identified to species level. Juvenile spiders were assigned to a species only when they were at a pre-adult or late juvenile instar

Results

Literature data analysis– Data extracted from 55 analyzed publications represent in total 23 genera and ca. 160 species of the family Liocranidae with ca. 3000 specimens collected worldwide (Fig. 2.1a). A visual summary of the data extraction process and data display in Plazi's Treatment Bank and GBIF can be found in Supplementary Figure 2.1. These include treatments of all currently valid genera and 90 species of the TG based on 1,309 specimens; out of 137 currently valid species [27]. The TG was mostly distributed in East and Southeast Asia (Fig. 2.1b) with the exception of two species of the genus *Oedignatha* found in the Seychelles. Within SEA, six genera of the TG have a broad distribution being reported from India and the southern region of mainland Asia to the Malay Archipelago (Fig. 2.1c-e, g-h). Two exceptions are *Jacaena* that has not been reported south of Thailand (Fig. 2.1f) and Sudaharmia that has only been reported within Indonesia (Fig. 2.1i). Indonesia (Six genera, 386 specimens), Thailand (Five, 351) and Malaysia (Four, 212) were the countries with a highest richness and abundance of TG spiders accounting for 72.5% of all the TG records (Fig. 2.2a). Thailand was the country that combined most occurrences of the TG genera and *T. politus* having 66% of all the known specimens of this species reported in literature. Within Thailand, the best sampled province is Chiang Mai accounting for 35% of all the TG specimen records for the country. Other relatively well known provinces were Krabi, Nakhon Ratchasima and Phuket, adding up to 30% of the country records (Fig. 2.2a). Chiang Mai had reports of four TG genera and 11 species, Krabi and Phuket had relatively less representation of the TG; however, these two provinces had 66 of the 68 specimens of *T. politus* recorded for the country.

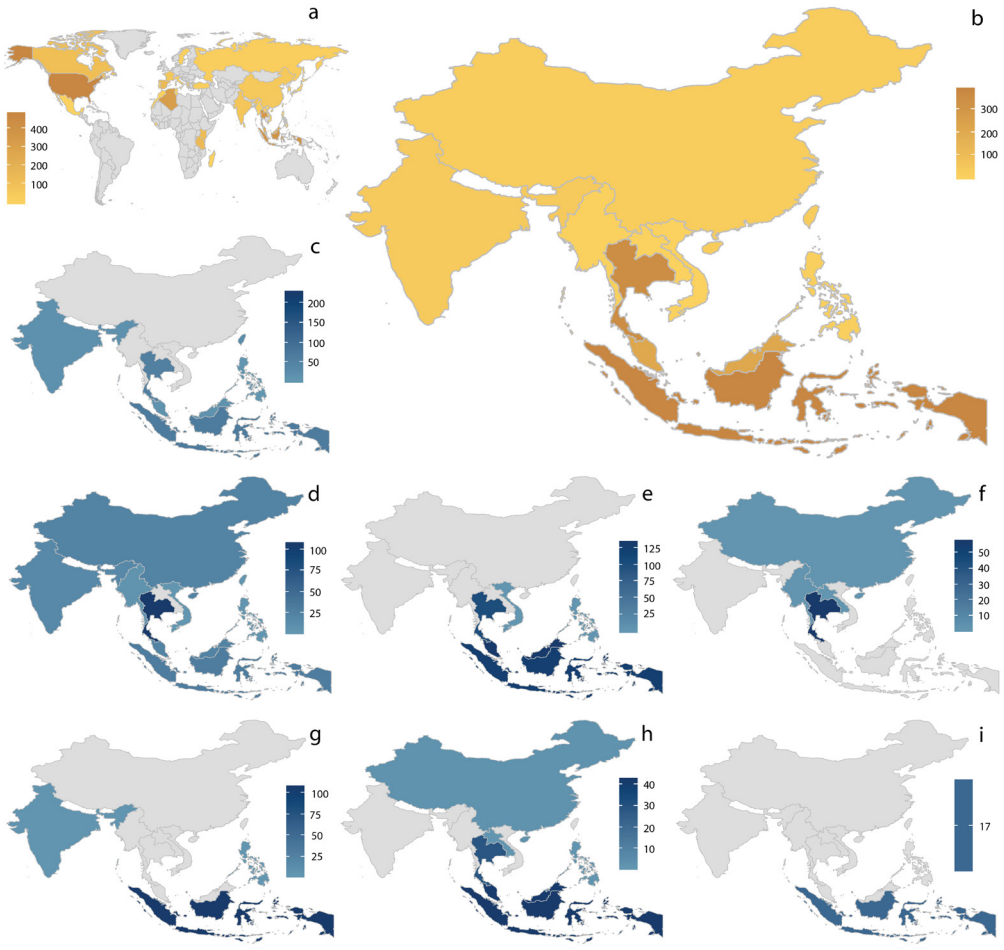


Figure 2.1.–Maps of liocranid spiders distribution. Based on geographic data extracted from taxonomic literature using Plazi's retrospective workflow (see Supplementary Table 1 for the whole set of documents used). Maps generated in RStudio [82–84]. a) Family: Liocranidae worldwide. b) Family Liocranidae in Southeast Asia (SEA). c) Genus: *Oedignatha*. d) *Sphingius*. e) *Teutamus*. f) *Jacaena*. g) *Koppe*. h) *Sesieutes*. i) Sudaharmia. Brown shades represent family distribution and blue shades represent genus distributions. Color intensity corresponds to numbers of specimens per country.

The majority of species treatments that we semantically enhanced contained collecting dates that allowed us to plot temporal distribution of the group within Thailand. Most specimens were collected between 1980 and 2009. These dates together with collecting locations allowed us to plot the known temporal and geographic distribution of our target taxon (Fig. 2.2b). For instance, most collecting is concentrated between May and December, with February and March being the least represented months. Similarly, Indonesia, Malaysia and Thailand are the best sampled countries in Southeast Asia.

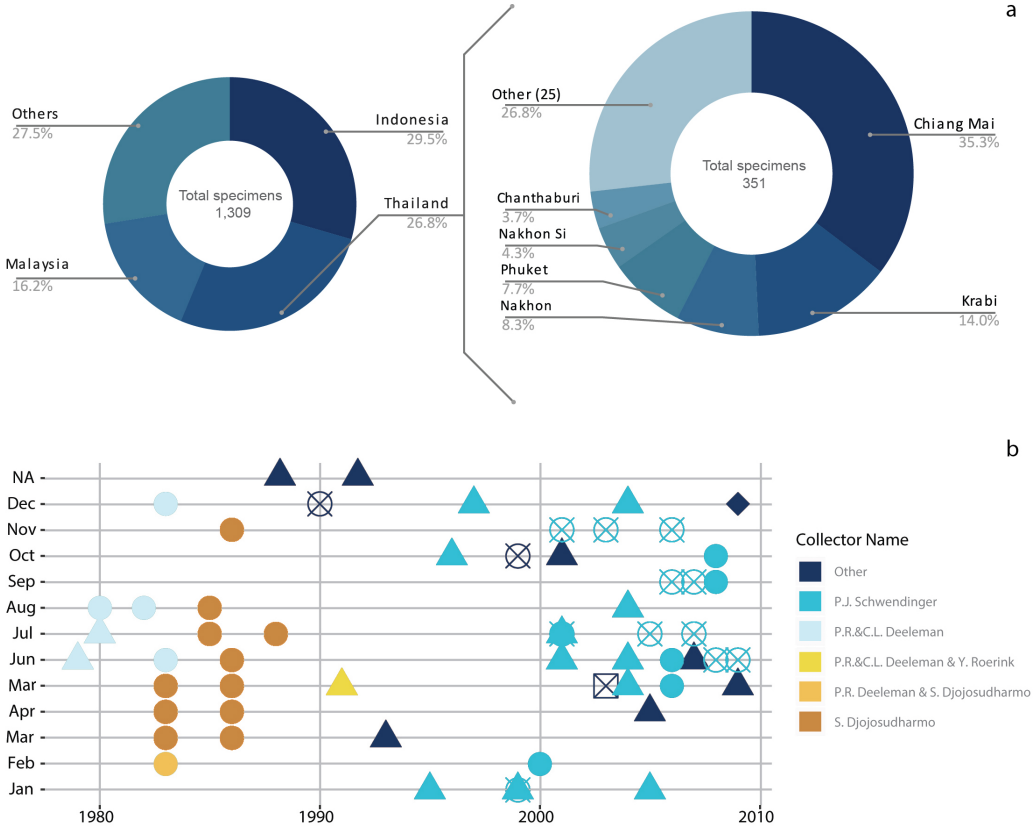


Figure 2.2.–Distribution of the *Teutamus* group in Southeast Asia according to taxonomic literature. Based on data extracted from 23 studies [19–23,28–30,39,42,50,56,59,61,65,68,73–75,77–80] using Plazi's retrospective workflow. a) Proportion of specimens reported per country, with detail of provinces in Thailand. b) Temporal and spatial distribution of collections for the past 40 years. ● = Indonesia, ▲=Malaysia, ⊗=Thailand, ♦ =Philippines, =Vietnam.

From an historical perspective, Indonesia was clearly the most sampled area during the 80s and Malaysia during the 90s, with more heterogeneous and international records appearing during the 2000s.Total monthly abundances suggest that adults of the TG are mostly found in between June and July, and October to January (Fig. 2.3a). A more detailed visualization at genus level shows that most TG genera have similar seasonal variations, with the exception of *Teutamus* that is most common between June and July (Fig. 2.3a). The species *T. politus* has adults reported mostly between June and July, and some specimens from September to December but none have been recorded between January and May (Fig. 2.3b).

Fieldwork– Our sampling produced 134 adult liocranid specimens from the following genera: *Jacaena* (3), *Oedignatha* (32), *Sesieutes* (3), *Sphingius* (1), *Teutamus* (95) (Table 1). Some juvenile specimens of *Oedignatha* and *Teutamus* could be matched to adults in the same sample and assigned to the same species adding up to a total of 229

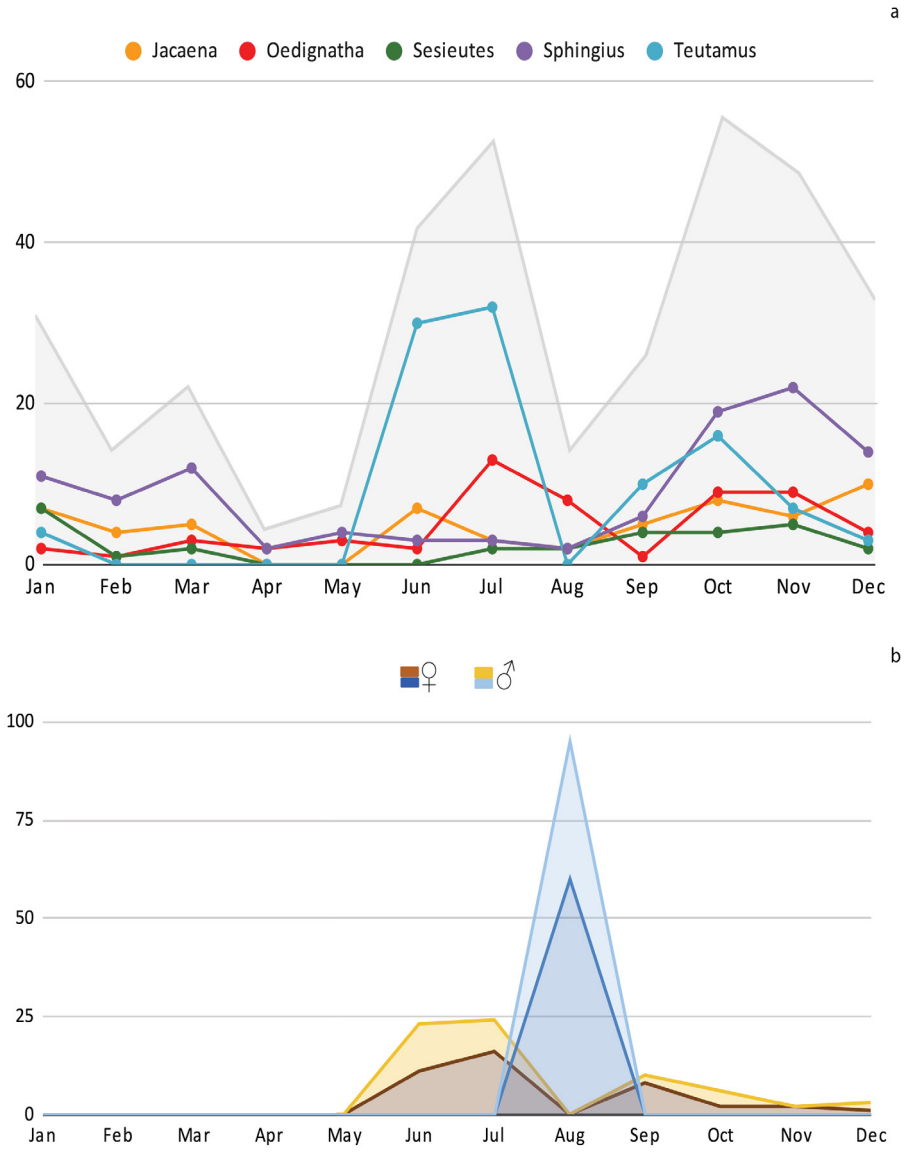


Figure 2.3.–Seasonal distribution of adult specimens of the *Teutamus* group in Thailand. Based on data extracted from 2 studies [19,21] using Plazi's retrospective workflow. a) Grey area indicates total number of specimens; lines detail richness per genus in literature. b) Relative abundances of males and females of *Teutamus politus*. Brown shades indicate specimens in literature; blue shades indicate specimens in our study.

identified specimens of the Liocranidae. We found four species of the TG in Chiang Mai: *Jacaena* lunulata, *Oedignatha* barbata, O. jocquei, and *Sphingius* cf. vivax; three species in Phuket: O. spadix, *Sesieutes* cf. minuatus, and *Teutamus politus*; and two species in Krabi: O. sp. and *T. politus*. Most of them were represented by males and females with the exception of J. lunulata and S. cf. vivax, where only males were found. These

two, along with O. barbata and O. sp., were the rarest species having three or fewer individuals in our sample. The most abundant species were O. spadix and *T. politus* with 21 and 95 adults respectively.

Discussion

Literature data analysis– Detecting and understanding biodiversity patterns require large amounts of high quality data. In recent years global databased like GBIF and Plazi have set standards for collection, curation and dissemination of these biological data. GBIF, the largest biodiversity data repository, has aggregated digitized specimen records from many of the world's most important biodiversity collections institutions. In addition, records from observation networks such as iNaturalist are aggregated on GBIF. However, legacy taxonomic literature as a source of biodiversity data has remained relatively unexplored until recent years. Taxonomic literature holds a vast amount of high-quality biodiversity data [12,85,86]. Like data from institutional collections and unlike data from observations networks, these data typically point to specimen objects archived in a natural history institution. Such records have the potential to be re-evaluated in a way that records from observation networks cannot be. It is worth noting that many specimens cited in the taxonomic literature, although archived in a natural history collection, are not necessarily among the institutional collections data shared with GBIF.

Table 1.–Records of Teutaumus group (TG) species from three Thai provinces. Total records from taxonomic literature (Spp. in literature) vs. Literature records from June-August (Spp. July-August) vs. our field samples (Spp. in our study). * indicates new geographic distribution for the species.

Province	Species	Spp. in literature			Spp. in lit. (July- August)			Spp. in our Study		
		♂	♀	Total	♂	♀	**Total**	♂	♀	**Total**
Chiang Mai	*Jacaena angoonae*	-	4	**4**	-	-	**-**	-	-	**-**
	Jacaena lunulata	8	5	**13**	-	-	**-**	3	-	**3**
	Jacaena mihun	3	3	**6**	-	-	**-**	-	-	**-**
	Jacaena schwendingeri	3	9	**12**	-	3	**3**	-	-	**-**
	Oedignatha barbata	6	5	**11**	2	2	**4**	1	1	**2**
	Oedignatha jocquei	8	15	**23**	6	9	**15**	1	6	**7**
	Sesieutes zhui	5	4	**9**	-	-	**-**	-	-	**-**
	Sphingius gothicus	16	6	**22**	-	-	**-**	-	-	**-**
	Sphingius penicillus	17	3	**20**	-	-	**-**	-	-	**-**

	*Sphingius vivax**	-	-	-	-	-	-	1	-	**1**
Krabi	*Oedignatha* sp.*	-	-	-	-	-	-	1	1	**2**
	Sesieutes aberrans	2	-	**2**	2	-	**2**	-	-	-
	Sphingius punctatus	-	1	**1**	-	-	-	-	-	-
	Teutamus politus	20	19	**39**	1	-	**1**	5	14	**19**
	Teutamus rama	4	3	**7**	-	-	-	-	-	-
Phuket	*Oedignatha spadix**	-	-	-	-	-	-	6	15	**21**
	Sesieutes cf. *minuatus**	-	-	-	-	-	-	2	1	**3**
	Teutamus politus	8	19	**27**	7	16	**23**	30	46	**76**
Total specimens		100	96	**196**	18	30	**48**	50	84	**134**

Data extraction from taxonomic literature can proceed along two major pathways: 1) prospective, where data is mobilized and shared with GBIF as part of the routine publication process, as has been implemented some journals like EJT [13] and ZooKeys [2,8] and some revisionary studies [87]; and 2) retrospective, where data is mined from legacy taxonomic data [11,12]. This retrospective approach was tested in our study by semantically enhancing records from more than 50 legacy taxonomic documents. From these sources, ca. 3000 specimens of the family Liocranidae were structured and mobilized, including more than 1300 records from about 100 treatments of TG taxa (Supplementary Table 1). These data included relevant biodiversity information, such as geographical distribution, date of collection, sex, and number of specimens.

Although the data contained in taxonomical treatments has been curated by specialists and is highly dependable, it is not free from error and methodological bias. Meyer, Weigelt, and Kreft [88], in their study of land plant data available on GBIF, documented data biases in two major groups: coverage (geographical and temporal documentation gaps) and uncertainty (accuracy or credibility). Another bias observed in GBIF, as well as biodiversity studies and funding in general, is related to the taxonomic coverage and over representation of some groups like birds and plants and under representation of megadiverse groups like insects and arachnids [89–92] (Supplementary Figure 2.2; see also Data Aggregation, below).

In our analysis we did not find clear cases of uncertainty bias with the exception of the absence of geographical coordinates that made some of the occurrences spatially ambiguous. However, geographical and temporal coverage bias was observed. Scientists do not sample randomly or evenly from the whole world; therefore, it should be expected that some areas and times are studied more than others. This makes it difficult to distinguish seasonal changes in abundance from uneven sampling effort at different times of the year. Nevertheless, existing records at least indicate the time of year when

specimens have been found in the past, and might therefore be found again. Overall, records of TG taxa were not evenly spread throughout the year. For example, zero specimens of *T. politus* are recorded for the month of August, suggesting that this might not be best time of year to search for this species in Thailand (Figs. 2.2-2.3). Although we had planned our sampling during the highest abundance peak (June-July; Fig. 2.3b), logistic constrains forced us to carry our sampling one month later. Nevertheless, we found a total of 188 specimens of this species during our collection, of which 95 were adults. Our results give evidence of the presence of these taxa during this time of the year, suggesting that the variation observed in legacy records is most probably due to temporal coverage bias and must be interpreted with care.

Another temporal coverage bias was observed when assessing specimen contributions per collector (Fig. 2.2b). We found P.J. Schwendinger to be the collector with most specimens contributed to the TG [19–23]; between 1983 and 2009 he collected 231 TG specimens in Thailand. However, most of his specimens, presumably, due to logistics, were reported around June and July, and December. Therefore, temporal distribution patterns, as observed in literature-extracted data (Figs. 2.2 and 2.3), could be an artifact of sampling bias and not necessarily reflect real seasonal variation of the taxa. Even taking into account these methodological biases, we consider specimen records in taxonomic literature to be among the best curated evidence of presence and, to some extent, relative abundances; and for many understudied and megadiverse taxa, this is the only source of specimen records available. Identifying and understanding data biases can help to identify temporal and spatial gaps were further sampling effort is needed.

Fieldwork– Data extracted from taxonomic literature on the family Liocranidae were used to create detailed profiles for the TG. These helped us to plan a collection that specifically targeted the re-collection of these taxa. Our analysis showed that within Southeast Asia, three provinces in Thailand, Chiang Mai, Phuket and Krabi were the best choice for targeting *T. politus* and its relatives.

This selection of times and places, in combination with specific methods for collecting ground spiders showed a high efficiency for sampling the TG. Our one-month expedition captured 134 adult spiders of the TG (Table 1) representing all TG genera previously reported for Thailand and six out of seven liocranid genera reported for this country (only missing Paratus Simon, 1898). In total, 351 adults of the TG had been reported from Thailand [19–23,73]; from these, ca. 200 had been reported in the same provinces we sampled (Chiang Mai, Krabi and Phuket) (Table 1). When comparing only the collections reported for the same months where we sample, we can observe that our approach was much more efficient, collecting 134 adults vs. 48 in literature. We collected a total of nine TG species vs. 14 reported from the same provinces and six reported from the same provinces and times. From these, *Teutamus politus* was the most abundant species in both literature and our study with 66 and 95 adults respectively (Fig. 2.3b).We collect more specimens of this species (188) than all the previous

records in literature combined (102 specimens) [19,21]. *Oedignatha* spadix was the second most abundant in our study with 21 adult specimens; *Oedignatha* spadix is previously known only from Indonesia [19].

Data aggregation– The interoperable network of Plazi allows the extracted data to be automatically shared with other biodiversity databases like GBIF. This allows taxonomic literature data to be analyzed together with data from Natural History collections and observation networks. Many studies have explored the limits and capabilities of GBIF data for setting conservation priorities [93–96], modeling [93,97,98], aggregation of different kinds of data and its biases [88,92,95,96,99,100], among others. The major GBIF data domains (institutional collections databases, observation networks, taxonomic literature, and, in some cases, DNA sequence databases), each have their particular biases, but taken together are complementary enough to serve as a basis for building more complete biodiversity knowledge. In the case of the *Teutamus* group, virtually all records in GBIF were originated from digitized collection data with only five records contributed through human observation and one through iBOL [101]. Even in groups where other sources of data are not available, digitized collection data can give important insights on aspects like the group taxonomy and distributions. Two studies in the Amazonia highlight the importance of collection-based data, by aggregating museum specimen data of several unrelated taxa collected in Amazonia comparing their richness, distribution and endemism [102,103]. This approach allowed them to identify undersampling bias taxonomically and spatially, and map priority areas for conservation based on biodiversity data. They also observed that even when individual datasets might be imperfect, the aggregation of different approaches and sources can help to better assess and allocate conservation efforts.

In our study, the addition of records from the taxonomic literature, aggregated with complementary data from other sources available on GBIF, improved the taxonomic, geographic, and seasonal coverage of TG taxa (Table 2), giving us an improved picture of their overall biodiversity pattern. Semantic enhancement of taxonomic literature cannot compete in volume against the millions of records sourced from natural history collections databases and especially observation networks. But records from taxonomic literature may be the only source of data available for the vast portion of biodiversity about which we know very little. In other words, observation network records tend to be copious but dominated by few species, while specimen records from natural history collections and especially taxonomic literature tend to be fewer in number, but are often the only source of data on rare species. The Plazi approach gives free and persistent access to high quality data curated by taxonomic experts that might potentially help to identify and close knowledge gaps for some underrepresented groups.

Observation networks are some of the largest contributors to GBIF in terms of total records, but these tend to be quite limited in taxonomic focus and rarely include any but the most conspicuous and recognizable representatives of small bodied, high diver-

sity groups like spiders. Here we emphasize the usefulness of the Plazi retrospective approach to close those gaps. Comparing a list of the currently valid species of the TG from the world spider catalog [27], the Plazi approach contributed with records on 89 out of 137 species. By contrast, only 41 species of the TG were present in GBIF before our study. Our contributions to the knowledge of these spiders can be also observed in the number of occurrences in GBIF. Literature extracted data on the TG currently represents 470 occurrences in GBIF versus the 180 occurrences that were available from collection-based data, observation and iBOL combined. Our marked-up documents account for 72% of the occurrences of the TG and the genus *Teutamus*, and 85% of records of our target species, *Teutamus politus* (Fig. 2.4). This gives evidence of the complementarity of these data sources and the importance of mobilizing and making publicly available all the specimen data contained in taxonomic literature.

It is worth noting that this complementarity can also mean that some records from literature and digitized collection data could be overlapping. However, ruling out these cases demands unambiguous collection numbers or specimen identifiers; or, in case this number is absent, comparing probable matches by collection date, locality, specimen count, and other data. For the *Teutamus* group, some records available in GBIF do have a unique collection number (e.g. *Teutamus politus* RMNH.ARA.15194). However, these identifiers are not always available (either in GBIF, on literature or on both) making difficult to reconcile data from different sources. Therefore setting unique identifiers and strengthening publication standards must be a top priority for the future [12,105–108]. This will help to generate usable and reliable datasets that can help to observe, study, and ultimately preserve biodiversity.

Structured, digitized specimen data extracted from taxonomic literature remains a small portion of the overall biodiversity data sphere, but it complements more mainstream data sources in important ways and has the potential to grow into a major source of data in its own right. Our study shows the importance of taxonomic literature records that, in combination with data from other sources, contributes to the most complete available assessment of spatial and temporal biodiversity pattern. Using this data for field work planning is but one possible application, but conservation risk assessment and species distribution modeling could be important in this context as well. The Plazi approach makes these data permanently available for others to re-use and add to in ways that we may or may not be able to currently imagine. Despite decades of ambitious and largely successful digitization efforts, much of the knowledge that biologists have accumulated about global biodiversity remains undigitized and unstructured, unqueryable, and difficult to access. The challenges presented by the global biodiversity crisis are daunting, and our best hope for addressing it begins with building a data infrastructure that faithfully represents the knowledge that generations of scientists have accumulated; specimen records from taxonomic literature are a key element in such an infrastructure.

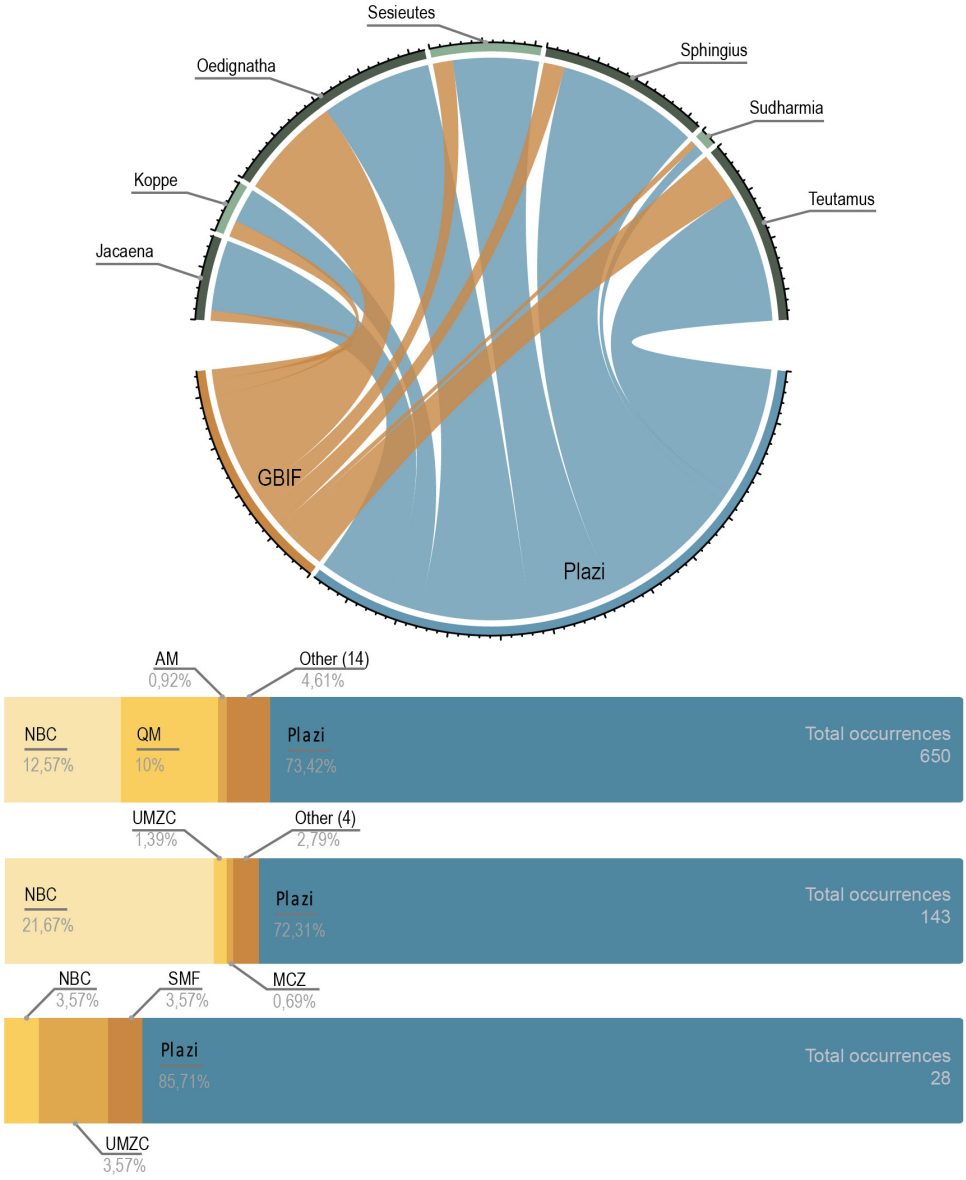


Figure 2.4.–Proportion of occurrences of the *Teutamus* group in GBIF [101]. Color indicates data source: digitized collection data (brown shaded) and taxonomic literature mined data (blue). Circle: Proportion per data source for the whole *Teutamus* group and each TG genera. Generated in RStudio[82,104]. Bars: detail of proportions and total occurrences TG (top), genus *Teutamus* (middle), and *Teutamus politus* (middle). Note the high proportion of data contributed through our mark-up and integration using Plazi's retrospective workflow). Collection abbreviations explained in Table 2.

Table 2.–*Teutamus* group in GBIF per collection/database comparing number of occurrences, total of specimens, geographical distribution and taxonomic coverage. Blue shaded squares indicate presence of each genus. J= *Jacaena*; K= *Koppe*; O= *Oedignatha*; Se= *Sesieutes*; Sp= *Sphingius*; Su= *Sudharmia*; T= *Teutamus*. Collection names: AM– Australian Museum, Australia; CAS– California academy of Sciences, USA; MACN– Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Argentina; MCZ– Museum of Comparative Zoology, Harvard, USA; MNHN–P–Muséum national d'Histoire naturelle-Paris, France; NBC– Naturalis Biodiversity Center (formerly RMNH), The Netherlands; NMNS– National Museum of Nature and Science, Japan; QM– Queensland Museum, Australia; SMF– Senckenberg Museum Frankfurt, Germany; SMNK– Staatliches Museum für Naturkunde Karlsruhe, Germany; UMZC– The University Museum of Zoology, Cambridge, UK; WAM– West Australia Museum, Australia; ZMUC– Zoological Museum, Natural History Museum, Denmark.

Source	Occurrences	Total Specimens	Geographical distribution	Genera representation						
				J	K	O	Se	Sp	Su	T
Plazi (23 documents in GBIF)	467	1035	China, India, Indonesia, Laos, Malaysia, Myanmar, Philippines, Seychelles, Singapore, Taiwan, Thailand, Vietnam	■	■	■	■	■	■	■
NBC	79	166	Indonesia, Malaysia, Sri Lanka, Netherlands, Thailand	■	■	■	■	■	■	■
QM	65	135	Australia, Malaysia, New Caledonia			■				
AM	6	ND	Australia, Papua New Guinea			■				
MCZ	5	5	Indonesia, Malaysia, Thailand	■		■				■
SMF	5	ND	Cambodia, India, Indonesia, Laos	■		■				■
MACN	2	ND	Thailand	■		■				
MNHN-P	2	4	Singapore, Sri Lanka		■		■			
UMZC	2	ND	Malaysia							■
WAM	2	3	Christmas Island			■				
NMNS	1	ND	Vietnam							■
CAS	1	1	Thailand					■		
SMNK	1	ND	Malaysia				■			
ZMUC	1	ND	Thailand							■
TOTAL	639	1349								

Acknowledgements

Thanks to Joe Dulyapat and Choojai Petcharad for their great assistance and participation during our fieldwork in Thailand. Thanks to editor Uwe Fritz, reviewer Torsten Dikow, and three anonymous reviewers for their valuable comments and suggestions. Funding for the first author was provided by CONACyT Becas al extranjero 294543/440613, Mexico. All specimens collected by us in Thailand were authorized under permit 5830802 emitted by the Department of National Parks, Wildlife and Plant Conservation, Thailand.

Data accessibility

Extracted data is available from Plazi [14] tb.plazi.org/GgServer/srsStats (refining search as needed) and GBIF [101,109,110]. A list of all the Plazi document UUID used in this study can be found in the Supplementary Table 1.

References

1. Catapano T. (National Center for Biotechnology Information, 2010). NoTaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. in *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010* DOI: 10.5281/zenodo.3484285.

2. Penev L, Agosti D, Georgiev T, Catapano T, Miller JA, Blagoderov V, Roberts D, Smith VS, Brake I, Ryrcroft S, Scott B, Johnson NF, Morris RA, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson FC, Erwin TL. (2010). Semantic tagging of and semantic enhancements to systematics papers: Zookeys working examples. *Zookeys* **50**, 1–16 DOI: 10.3897/zookeys.50.538.

3. Penev L, Lyal CHC, Weitzman A, Morse DR, King D, Sautter G, Georgiev T, Morris RA, Catapano T, Agosti D. (2011). XML schemas and mark-up practices of taxonomic literature. *Zookeys* **150**, 89–116 DOI: 10.3897/zookeys.150.2213.

4. Dikow T, Agosti D. (2006). Utilizing online resources for taxonomy: a cybercatalog of Afrotropical apiocerid flies (Insecta: Diptera: Apioceridae). *Biodivers. Data J.* **3**, e5707.

5. Creech J. (2012). Biodiversity Heritage Library. *Coll. Res. Libr. News* **73**, 626–627.

6. Gwinn NE, Rinaldo C. (2009). The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA J.* **35**, 25–34 DOI: 10.1177/0340035208102032.

7. Page RDM. (2010). Enhanced display of scientific articles using extended metadata. *J. Web Semant.* **8**, 190–195 DOI: 10.1016/j.websem.2010.03.004.

8. Agosti D, Catapano T, Sautter G, Egloff W. (2019). The Plazi Workflow: The PDF prison break for biodiversity data. *Biodivers. Inf. Sci. Stand.* **3**, e37046 DOI: 10.3897/biss.3.37046.

9. Cui H. (2008). Converting Taxonomic Descriptions to New Digital Formats. *Biodivers. Informatics* **5**, 20–40 DOI: 10.1111/j.1468-2478.2011.00691.x.

10. Thessen AE, Patterson D. (2011). Data issues in the life sciences. *Zookeys* **150**, 15–51 DOI: 10.3897/zookeys.150.1766.

11. Miller JA, Dikow T, Agosti D, Sautter G, Catapano T, Penev L, Zhang Z-Q, Pentcheff D, Pyle R, Blum S, Parr C, Freeland C, Garnett T, Ford LS, Muller B, Smith L, Strader G, Georgiev T, Bénichou L. (2012). From taxonomic literature to cybertaxonomic content. *BMC Biol.* **10**, 87 DOI: 10.1186/1741-7007-10-87.

12. Miller JA, Agosti D, Penev L, Sautter G, Georgiev T, Catapano T, Patterson D, King D, Pereira S, Vos RA, Sierra S. (2015). Integrating and visualizing primary data from prospective and

legacy taxonomic literature. *Biodivers. data J.* **3**, e5063 DOI: 10.3897/BDJ.3.e5063.

13. Chester C, Agosti D, Sautter G, Catapano T, Martens K, Gérard I, Bénichou L. (2019). EJT editorial standard for the semantic enhancement of specimen data in taxonomy literature. *Eur. J. Taxon.* **586**, DOI: 10.5852/ejt.2019.586.

14. Plazi. (2020). PLAZI Home Page. Available from http://plazi.org/.org [20th June 2020].

15. Agosti D, Egloff W. (2009). Taxonomic information exchange and copyright: the Plazi approach. *BMC Res. Notes* **2**, 53 DOI: 10.1186/1756-0500-2-53.

16. GBIF. (2019) Global Biodiversity Informaton Facility Home Page. Available from: http://www.gbif.org [4th April 2019].

17. GBIF Secretariat. (2019) GBIF Backbone Taxonomy. *Checklist dataset* https://doi.org/10.15468/39omei accessed via GBIF.

18. Ramírez MJ. (2014). The morphology and phylogeny of dionychan spiders (Araneae, Araneomorphae). *Bull. Am. Museum Nat. Hist.* **390**, 1–374 DOI: doi.org/10.1206/821.1.

19. Deeleman-Reinhold C. (Leiden : Brill, 2001). *Forest spiders of South East Asia: with a revision of the sac and ground spiders (Araneae: Clubionidae, Corinnidae, Liocranidae, Gnaphosidae, Prodidomidae and Trochanterriidae)*.

20. Dankittipakul P, Tavano M, Singtripop T. (2011). Neotype designation for *Sphingius thecatus* Thorell 1890 synonymies new records and descriptions of six new species from Southeast Asia(Araneae Liocranidae). *Zootaxa* 1–20.

21. Dankittipakul P, Tavano M, Singtripop T. (2012). Seventeen new species of the spider genus *Teutamus* Thorell, 1890 from Southeast Asia (Araneae: Liocranidae). *J. Nat. Hist.* **46**, 1689–1730 DOI: 10.1080/00222933.2012.681314.

22. Dankittipakul P, Tavano M, Singtripop T. (2013). Revision of the spider genus *Jacaena* Thorell, 1897, with descriptions of four new species from Thailand (Araneae: Corinnidae). *J. Nat. Hist.* **47**, 1539–1567 DOI: 10.1080/00222933.2012.763059.

23. Dankittipakul P, Deeleman-Reinhold C. (2013). Delimitation of the spider genus *Sesieutes* Simon, 1897, with descriptions of five new species from South East Asia (Araneae: Corinnidae). *J. Nat. Hist.* **47**, 167–195 DOI: 10.1080/00222933.2012.742165.

24. Wheeler WC, Coddington JA, Crowley LM, Dimitrov D, Goloboff PA, Griswold CE, Hormiga G, Prendini L, Ramírez MJ, Sierwald P, Almeida-Silva L, Alvarez-Padilla F, Arnedo MA, Benavides Silva LR, Benjamin SP, Bond JE, Grismado CJ, Hasan E, Hedin M, Izquierdo MA, Labarque FM, Ledford J, Lopardo L, Maddison WP, Miller JA, Piacentini LN, Platnick NI, Polotow D, Silva-Dávila D, Scharff N, Szüts T, Ubick D, Vink CJ, Wood HM, Zhang J. (2017). The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics* **33**, 574–616 DOI: 10.1111/cla.12182.

25. Rivera-Quiroz FA, Schilthuizen M, Petcharad B, Miller JA. (2020). Imperfect and askew: A review of asymmetric genitalia in araneomorph spiders (Araneae: Araneomorphae). *PLoS One* **15**, e0220354 DOI: 10.1371/journal.pone.0220354.

26. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Mons B, *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, DOI: 10.1038/sdata.2016.18.

27. WSC. http://www.wsc.nmbe.ch/ (2020) World Spider Catalog Version 21.0. *Natural History Museum Bern, online at http://wsc.nmbe.ch* DOI: 10.24436/2.

28. Rstudio T. (2020). RStudio: Integrated Development for R. *Rstudio Team, PBC, Boston, MA URL http://www.rstudio.com/* DOI: 10.1145/3132847.3132886.

29. Becker RA, Wilks AR, Brownrigg R, Minka TP, Deckmyn A. (2017) CRAN - Package maps. *CRAN R-Project*.

30. Wickham H. (2016). *ggplot2 Elegant Graphics for Data Analysis (Use R!). Springer* DOI: 10.1007/978-0-387-98141-3.

31. Biswas V, Roy R. (2008). Description of six new species of spiders of the genera Lathys (Family: Dictynidae), Marpissa (Family: Salticidae), Misumenoides (Family: Thomisidae), Agroeca (Family: Clubionidae), Gnaphosa (Family: Gnaphosidae) and *Flanona* (Family: Lycosidae) - F. *Rec. Zool. Surv. India* **108**, 43–57.

32. Biswas B, Biswas K. (1992). Araneae: Spiders. in *State Fauna series 3: Fauna of West Bengal 3* 357–500.

33. Biswas B, Majumder SC. (1995). Araneae:

Spider. in *Fauna of Meghalaya, State Fauna Series. Zoological Survey of India Kolkata* 93–128.

34. Chen SH, Huang WJ. (2009). A newly recorded spider *Oedignatha platnicki* Song et Zhu 1998 from Taiwan, with description of the female (Araneae, Corinnidae). *BioFormosa* **44**, 31–36.

35. Dankittipakul P, Deeleman-Reinhold C. (2012). A new spider species of the genus *Sudharmia* from Sumatra, Indonesia (Araneae, Liocranidae). *Dongwuxue Yanjiu* **33**, 187–190 DOI: 10.3724/SP.J.1141.2012.02187.

36. Jäger P. (2007). Spiders from Laos with descriptions of new species (Arachnida: Araneae). *Acta Arachnol.* **56**, 29–58.

37. Ono H. (2009). Three new spiders of the family Clubionidae, Liocranidae and Gnaphosidae (Arachnida, Araneae) from Vietnam. *Bull. Natl. Museum Nat. Sci. Tokyo* **35**, 1–8.

38. Reddy TS, Patel BH. (1993). Two New Species Of The Genus *Oedignatha* Thorell (Araneae: Clubionidae) From Coastal Andhra Pradesh, India. *Entomon* **18**, 47–51.

39. Saaristo MI. (2002). New species and interesting new records of spiders from Seychelles (Arachnida, Araneaea). *Phelsuma* **10**, 1–32.

40. Tso I, Zhu MS, Zhang J, Zhang F. (2005). Two new and one newly recorded species of Corinnidae and Liocranidae from Taiwan (Arachnida: Araneae). *Acta Arachnol.* **54**, 45–49.

41. Zhang F, Fu JY. (2010). First Report of the Genus *Sesieutes* Simon (Araneae: Liocranidae) from China, with Description of One New Species. *Entomol. News* **121**, 69–74 DOI: 10.3157/021.121.0114.

42. Zhang F, Fu JY, Zhu MS. (2009). Spiders of the genus *Sphingius* (Araneae: Liocranidae) from China, with description of two new species. *Zootaxa* 31–44.

43. Zhao Y, Peng XJ. (2013). Three new species of spiders of the family Liocranidae (Arachnida: Araneae) from China. *Orient. Insects* **47**, 176–183 DOI: 10.1080/00305316.2013.811021.

44. Barrion AT, Litsinger JA. (1995). Family Clubionidae Wagner- Genera *Alaeho*, *Castianeira*, *Agroeca*, *Phrurolithus* & *Scotinella*. in *Riceland spiders of South and Southeast Asia* 170–180 DOI: DOI: 10.5281/zenodo.897849.

45. Bastawade DB. (2006). Replacement name for *Amaurobius indicus* Bastawade and its transfer to family Corinnidae (Arachnida: Araneae). *Zoo's Print J.* **21**, 2307.

46. Bastawade DB. (2002). Three new species from the spider families Amaurobiidae, Thomisidae and Salticidae (Araneae: Arachnida) from India. *J. Bombay Nat. Hist. Soc.* **99**, 274–281.

47. Bennett R, Copley C, Copley D. (2013). *Apostenus ducati* (Araneae: Liocranidae) sp. nov.: A second Nearctic species in the genus. *Zootaxa* **3647**, 63–74 DOI: 10.11646/zootaxa.3647.1.3.

48. Biswas V, Raychaudhuri D. (2000). Sac spiders of Bangladesh-II: Genera *Castianeira* Keyserling, *Sphingius* Thorell and *Trachelas* Koch (Araneae: Clubionidae). *Rec. Zool. Surv. India* **98**, 131–139.

49. Meier R, Dikow T. (2004). Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conserv. Biol.* **18**, 478–488 DOI: 10.1111/j.1523-1739.2004.00233.x.

50. Dikow T, Meier R, Vaidya GG, Londt JGH. (2009). Biodiversity Research Based on Taxonomic Revisions - A Tale of Unrealized Opportunities. in *Diptera Diversity: Status, Challenges, and Tools* 323–345.

51. Markee A, Dikow T. (2018). Taxonomic revision of the assassin-fly genus Microphontes Londt, 1994 (Insecta, diptera, asilidae). *African Invertebr.* DOI: 10.3897/afrinvertebr.59.30684.

52. Meyer C, Weigelt P, Kreft H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 DOI: 10.1111/ele.12624.

53. Leather SR. (2009). Taxonomic chauvinism threatens the future of entomology. *Biologist* **56**, 10–13.

54. Cardoso P, Erwin TL, Borges PAV, New TR. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* **144**, 2647–2655 DOI: 10.1016/j.biocon.2011.07.024.

55. Titley MA, Snaddon JL, Turner EC. (2017). Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PLoS One* **12**, e0189577 DOI: 10.1371/journal.pone.0189577.

56. Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.*

7, DOI: 10.1038/s41598-017-09084-6.

57. Bartomeus I, Stavert JR, Ward D, Aguado O. (2019). Historical collections as a tool for assessing the global pollination crisis. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, DOI: 10.1098/rstb.2017.0389.

58. Shirey V, Seppälä S, Branco VV, Cardoso P. (2019). Current GBIF occurrence data demonstrates both promise and limitations for potential red listing of spiders. *Biodivers. Data J.* **7**, e47369 DOI: 10.3897/BDJ.7.E47369.

59. Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, Possingham HP, Lindenmayer DB. (2019). Do big unstructured biodiversity data mean more knowledge? *Front. Ecol. Evol.* **6**, DOI: 10.3389/fevo.2018.00239.

60. Iannella M, D'Alessandro P, Biondi M. (2019). Entomological knowledge in Madagascar by GBIF datasets: Estimates on the coverage and possible biases (Insecta). *Fragm. Entomol.* **51**, 1–10 DOI: 10.4081/fe.2019.329.

61. Beck J, Böller M, Erhardt A, Schwanghart W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecol. Inform.* DOI: 10.1016/j.ecoinf.2013.11.002.

62. Smith JA, Benson AL, Chen Y, Yamada SA, Mims MC. (2020). The power, potential, and pitfalls of open access biodiversity data in range size assessments: Lessons from the fishes. *Ecol. Indic.* **110**, DOI: 10.1016/j.ecolind.2019.105896.

63. Meyer C, Jetz W, Guralnick RP, Fritz SA, Kreft H. (2016). Range geometry and socio-economics dominate species-level biases in occurrence information. *Glob. Ecol. Biogeogr.* **25**, 1181–1193 DOI: 10.1111/geb.12483.

64. Hochmair HH, Scheffrahn RH, Basille M, Boone M. (2020). Evaluating the data quality of iNaturalist termite records. *PLoS One* **15**, e0226534 DOI: 10.1371/journal.pone.0226534.

65. GBIF.org (11 August 2020). GBIF Occurrence Download (Liocranidae) https://doi.org/10.15468/dl.jfy7sp.

66. Kress WJ, Heyer WR, Acevedo P, Coddington J, Cole D, Erwin TL, Meggers BJ, Pogue M, Thorington RW, Vari RP, Weitzman MJ, Weitzman SH. (1998). Amazonian biodiversity: Assessing conservation priorities with taxonomic data. *Biodivers. Conserv.* **7**, 1577–1587 DOI: 10.1023/A:1008889803319.

67. Heyer WR, Coddington JA, Kress WJ, Acevedo P, Cole D, Erwin TL, Meggers BJ, Pogue M, Thorington RW, Vari RP, Weitzman MJ, Weitzman SH. (1999). Amazonian biotic data and conservation decisions. *J. Brazilian Assoc. Adv. Sci.* **51**, 372–385.

68. Gu Z, Gu L, Eils R, Schlesner M, Brors B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics* DOI: 10.1093/bioinformatics/btu393.

69. Page RDM. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* **9**, 345–354 DOI: 10.1093/bib/bbn022.

70. Page RDM. (2009). BioGUID: Resolving, discovering, and minting identifiers for biodiversity informatics. *BMC Bioinformatics* DOI: 10.1186/1471-2105-10-S14-S5.

71. Guralnick RP, Cellinese N, Deck J, Pyle RL, Kunze J, Penev L, Walls R, Hagedorn G, Agosti D, Wieczorek J, Catapano T, Page RDM. (2015). Community next steps for making globally unique identifiers work for biocollections data. *Zookeys* 133–154 DOI: 10.3897/zookeys.494.9352.

72. Nelson G, Sweeney P, Gilbert E. (2018). Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. *Appl. Plant Sci.* **6**, e1027 DOI: 10.1002/aps3.1027.

73. Bosmans R. (1999). The genera *Agroeca*, *Agraecina*, *Apostenus* and *Scotina* in the Maghreb countries (Araneae: Liocranidae). *Bull. l' Inst. R. Sci. Nat. Belgique* **69**, 25–34.

74. Bosmans R. (2011). On some new or rare spider species from Lesbos, Greece (Araneae: Agelenidae, Amaurobiidae, Corinnidae, Gnaphosidae, Liocranidae). *Arachnol. Mitteilungen* 15–22 DOI: 10.5431/aramit4003.

75. Bosmans R, van Keer J. (2012). On the spider species described by L. Koch in 1882 from the Balearic Islands (Araneae). *Arachnol. Mitteilungen* **43**, 5–16 DOI: doi:10.5431/aramit4306.

76. Bosselaers J. (2009). Studies in Liocranidae (Araneae): Redescriptions and transfers in *Apostenus* Westring and *Brachyanillus* Simon, as well as description of a new genus. *Zootaxa* 37–55.

77. Bosselaers J. (2012). Two interesting new ground spiders (Araneae) from the Canary Islands and Greece. *Serket* **13**, 83–90.

78. Bosselaers J, Dierick M, Cnudde V, Masschaele B, Van Hoorebeke L, Jacobs P. (2010). High-resolution X-ray computed tomography of an

extant new *Donuea* (Araneae: Liocranidae) species in Madagascan copal. *Zootaxa* 25–35.

79. Bosselaers J, Jocqué R. (2013). Studies in Liocranidae (Araneae): a new afrotropical genus featuring a synapomorphy for the Cybaeodinae. *Eur. J. Taxon.* **40**, 1–49.

80. Candek K, Gregorič M, Kostanjšek R, Frick H, Kropf C, Kuntner M. (2013). Targeting a portion of central European spider diversity for permanent preservation. *Biodivers. Data J.* **1**, e980 DOI: 10.3897/BDJ.1.e980. eCollection 2013.

81. Crespo LC, Domènech M, Enguídanos A, Malumbres-Olarte J, Cardoso P, Moya-Laraño J, Frías-López C, Macías-Hernández N, de Mas E, Mazzuca P, Mora E, Opatova V, Planas E, Ribera C, Roca-Cusachs M, Ruiz D, Sousa P, Tonzo V, Arnedo MA. (2018). A DNA barcode-assisted annotated checklist of the spider (Arachnida, Araneae) communities associated to white oak woodlands in Spanish National Parks. *Biodivers. Data J.* **6**, e29443 DOI: doi:10.3897/BDJ.6.e29443.

82. Danilov SN. (1998). The spider family Liocranidae in Siberia and Far East (Aranei). *Arthropoda Sel.* **7**, 313–317.

83. Deltshev C, Komnenov M, Blagoev G, Georgiev T, Lazarov S, Stojkoska E, Naumova M. (2013). Faunistic diversity of spiders (Araneae) in Galichitsa mountain (FYR Macedonia). *Biodivers. Data J.* **1**, e977 DOI: 10.3897/BDJ.1.e977.

84. Deltshev C, Wang C. (2016). A new *Agraecina* spider species from the Balkan Peninsula (FYR Macedonia) (Araneae: Liocranidae). *Zootaxa* **4117**, 135–140 DOI: 10.11646/zootaxa.4117.1.8.

85. Elverici M, Özkütük RS, Kunt KB. (2013). Two new liocranid species records from Turkey (Araneae: Liocranidae). *Munis Entomol. Zool.* **1**, 305–308.

86. Esyunin SL, Kazantsev DK. (2007). On the spider (Aranei) fauna of the Pechoro-Ilychskiy Reserve (north Urals), with the description of a new *Agroeca* species (Liocranidae). *Arthropoda Sel.* **16**, 245–250.

87. Felton C, Judd S, Merrett P. (2004). *Agroeca dentigera* Kulczynski, 1913, a liocranid spider new to Britain (Araneae, Liocranidae). *Bull. Br. Arachnol. Soc.* **13**, 90–92.

88. Fu JY, Zhang F, Zhu MS. (2009). Redescription of a little-known spider species, *Mesiotelus lubricus* (Simon, 1880) (Aranei: Liocranidae) from China. *Arthropoda Sel.* **17**, 169–173.

89. Hayashi T. (1992). Three species of the genus *Agroeca* (Araneae: Clubionidae) from Japan, including a new species. *Acta Arachnol.* **41**, 133–137.

90. Jonsson LJ. (2005). *Agroeca dentigera* and *Entelecara omissa* (Araneae: Liocranidae, Linyphiidae), found in Sweden. *Arachnol. Mitteilungen* 49–52.

91. Marusik YM, Koponen S. (2000). New data on spiders (Aranei) from the Maritime Province, Russian Far East. *Arthropoda Sel.* **9**, 55–68.

92. Marusik YM, Omelko MM, Koponen S. (2016). Rare and new for the fauna of the Russian Far East spiders (Aranei). *Far East. Entomol.* **317**, 1–15.

93. Marusik YM, Zheng G, Li S. (2008). A review of the genus *Paratus* Simon (Araneae, Dionycha). *Zootaxa* **1965**, 50–60.

94. Namkung J. (1989). A new species of the genus *Agroeca* (Araneae: Clubionidae) from Korea. *Korean Arachnol.* **5**, 23–27.

95. Platnick NI, Di Franco F. (1992). On the relationship of the spider genus *Cybaeodes* (Araneae, Dionycha). *Am. Museum Novit.* 9.

96. Reboleira AS, Pérez AJ, López H, Macías-Hernández N, de la Cruz S, Oromí P. (2012). Catalogue of the type material in the entomological collection of the University of La Laguna (Canary Islands, Spain). I. Arachnida. *Zootaxa* **3556**, 61–79.

97. Ribera C, de Mas E. (2015). Description of three new troglobiontic species of *Cybaeodes* (Araneae, Liocranidae) endemic to the Iberian Peninsula. *Zootaxa* **3957**, 313–323 DOI: 10.11646/zootaxa.3957.3.4.

98. Sankaran PM, Malamel JJ, Joseph MM, Sebastian PA. (2017). A new species of *Paratus* Simon, 1898 (Araneae: Liocranidae, Paratinae) from India. *Zootaxa* **4286**, 139–144 DOI: doi:10.11646/zootaxa.4286.1.12.

99. Seo BK. (2011). Description of three liocr0anid spider species from Korea (Araneae: Liocranidae). *Entomol. Res.* **41**, 98–102 DOI: 10.1111/j.1748-5967.2011.00326.x.

100. Seyyar O, Oba A, Demir H, Turkes T. (2016). *Arabelia* Bosselaers, 2009 and *Arabelia pheidoleicomes* Bosselaers, 2009 (Araneae: Liocranidae) are new records for the Turkish Spider Fauna. *Serket* **15**, 30–32.

101. Ubick D, Platnick NI. (1991). On *Hesperocranum*, A New Spider Genus from Western North America (Araneae, Liocranidae). *Am. Museum Novit.* 1–12.

102. Ubick D, Vetter RS. (2005). A New Species of *Apostenus* From California, With Notes on the Genus (Araneae, Liocranidae). *J. Arachnol.* **33**, 63–75 DOI: 10.1636/H03-24.

103. Vetter RS. (2001). Revision of the spider genus *Neoanagraphis* (Araneae,Liocranidae). *J. Arachnol.* **29**, 1–10 DOI:10.1636/0161-8202(2001)029[0001:ROTSGN]2.0.CO;2.

104. Warui C, Jocqué R. (2002). The First Gallieniellidae (Araneae) from Eastern Africa. *J. Arachnol.* **30**, 307–315 DOI:10.1636/0161-8202(2002)030[0307:TFGAFE]2.0.CO;2.

105. Wunderlich J. (2011). On European spiders of the nominal families Liocranidae, Miturgidae and Zoridae (Araneae), with descriptions of new taxa. *Beiträge zur Araneologie* **6**, 108–120.

106. Zapata L V., Ramírez MJ. (2010). A new species of the genus *Paratus* Simon (araneae: liocranidae) from Thailand. *Zootaxa* 65–68.

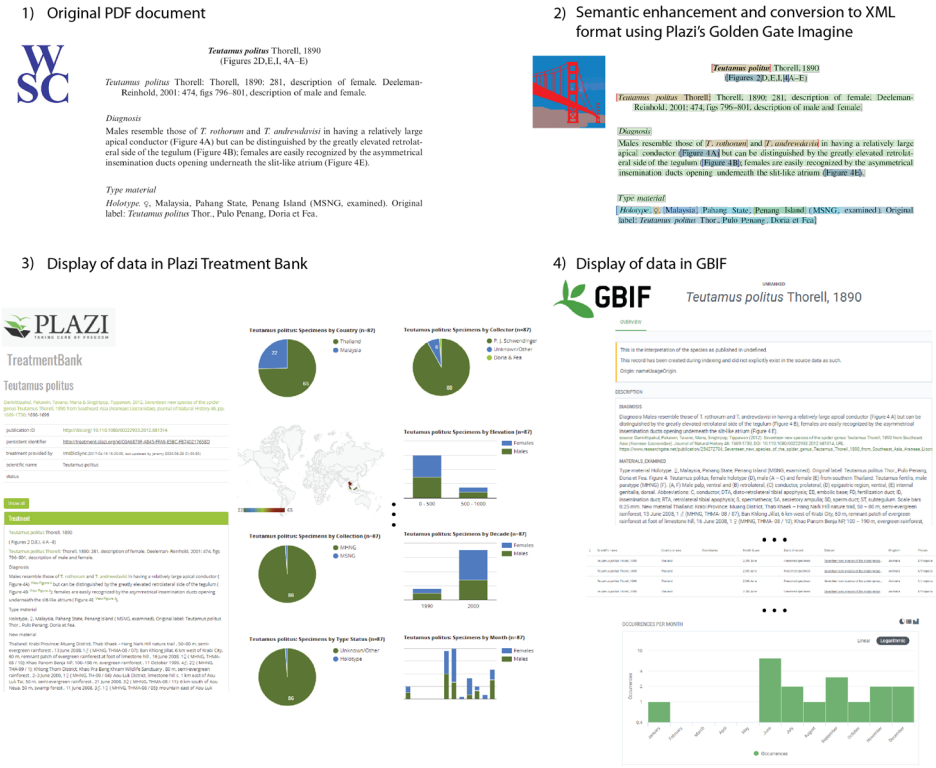
107. Zonstein SL, Marusik YM, Omelko M. (2015). A survey of spider taxa new to Israel (Arachnida: Araneae). *Zool. Middle East* **61**, 372–385 DOI: 10.1080/09397140.2015.1095525.

108. Vink CJ, Thomas SM, Paquin P, Hayashi CY, Hedin M. (2005). The effects of preservatives and temperatures on arachnid DNA. *Invertebr. Syst.* **19**, 99–104 DOI: 10.1071/IS04039.

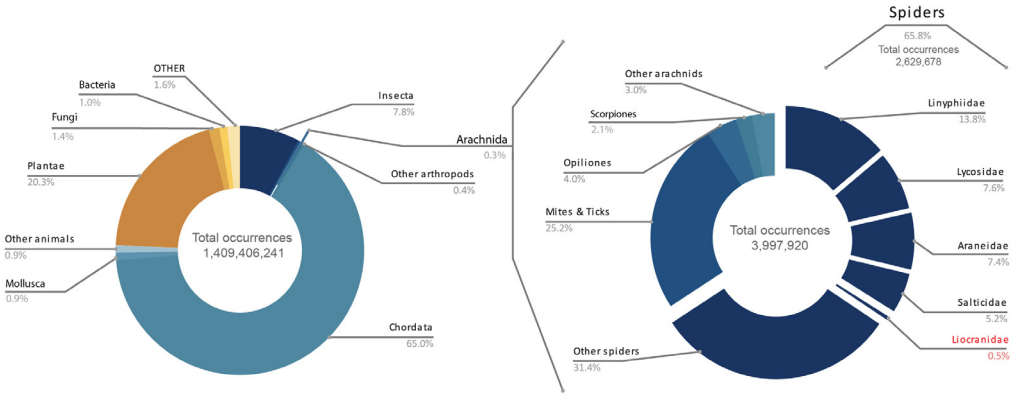
109. GBIF. org (12 July 2019). GBIF Occurrence Download (Liocranidae) https://doi.org/10.15468/dl.fcpcw9.

110. GBIF.org (20 March 2020). GBIF Occurrence Download (Sudharmia, Sesieutes, Jacaena, Koppe, Sphingius, Oedignatha, Teutamus) https://doi.org/10.15468/dl.3eh0rl.

Supplementary Figure 2.1.–Visual summary of the data extraction process for *Teutamus politus* treatment. From Dankittipakul, Tavano, and Singtripop (2012). 1- Taxonomic document in PDF format downloaded from the World Spider Catalog https://wsc.nmbe.ch/species/7486. 2- Conversion to XML format using Golden Gate Imagine. Each color in the figure text represents a semantic tag in the extraction process (Sautter, Böhm, and Agosti 2007). 3- Extracted treatment as displayed in Plazi http://tb.plazi.org/GgServer/html/03A6879FA845FFA9E5BCFB740217658D. To the left, whole treatment text, illustrations and link to the original source, to the right charts and maps based on the specimen data. 4- Specimen data and taxonomic treatment text displayed in GBIF https://www.gbif.org/species/130509488.



Supplementary Figure 2.2.–Proportion of GBIF records per taxonomic group. Left circle represent the whole GBIF database. Right circle represent spiders and other arachnids detailing the proportion of the best represented spider families and the Liocranidae (in red).



Supplementary Table 1.–Complete list of processed publications of the family Liocranidae. Plazi UUID code (a unique persistent identifier given to documents), number of liocranid genera and species treated, and specimens listed per study. The UUID can be added to the prefix this prefix "http://tb.plazi.org/GgServer/summary/" to access the index of linked treatments for that source. A list of the references used in this table can be found at the end of this supplementary document.

Study	Article UUID	Non- TG			TG			Total Specimens
		Genera	Species	Specimens	Genera	Species	Specimens	
Barrion and Litsinger 1995	D6116953413AA950FFBB4926252AFFE1	-	-	-	3	3	8	8
Bastawade 2006	E370626DFFC6FFDCFFB4FFFFFFC3E666	-	-	-	1	1	12	12
Bastawade 2002	FFCC8B0CFFA3FF80A43F537CFFC6F93F	-	-	-	1	1	3	3
Bennett, Copley, and Copley 2013	8F7EFFBB8872FFB96C32B8600B71A62C	1	1	114	-	-	-	114
Biswas and Raychaudhuri 2000	FFEEFFB22C17FFEEFF9E8F29FFCBFFD8	-	-	-	1	1	14	14
Biswas and Roy 2008	FFF2D503FFD8154B7903FF201A02FF92	1	1	6	-	-	-	6
Biswas and Biswas 1992	06598512FFD6FFEFFF80FFD3F611FFF8	-	-	-	2	2	9	9
Biswas and Majumder 1995	8F48FF87FF9A9839B635FFD6FF85DD64	-	-	-	1	1	1	1
Bosmans and van Keer 2012	DE294F64CD29FFFCCA795F1BFF93FFAA	1	1	2	-	-	-	2
Bosmans 1999	FFAAFF87FFCDFFD9FF87461AFFA12804	4	7	339	-	-	-	339
Bosselaers 2009	FF805162FFD1F162FFF3F327FF8FD505	4	5	109	-	-	-	109
Bosselaers 2012	FF9CFFB9FB60691AFFEAFF9C85151220	1	1	1	-	-	-	1
Bosselaers et al. 2010	7F15FF9E4C7DFF817E28ED77920EEC4D	1	1	2	-	-	-	2
Candek et al. 2013	622F99677618FFF81D122429FFE4FF91	2	2	6	-	-	-	6
Chen and Huang 2009	FFEEFFF5C520246D3960E4674C4EE517	-	-	-	1	1	24	24
Crespo et al. 2018	FFA7F823FFB5FFF26305FFECFFC5B633	4	9	70	-	-	-	70
Danilov 1998	733D3830FFB3FFFFFFF4FFC8FFAFFF8F	1	3	21	-	-	-	21

Dankittipakul and Deeleman-Reinhold 2012	FA379502FFBEFFB0FFF8FF-C06931FFFB	-	-	-	1	1	9	9
Dankittipakul and Deeleman-Reinhold 2013	FFF7FFD5FFEF5F40FF83FFDAFF-BE5660	-	-	-	1	9	54	54
Dankittipakul, Tavano, and Singtripop 2011	566E5A667339121BA634FFD-C3833FF8F	-	-	-	1	13	79	79
Dankittipakul, Tavano, and Singtripop 2012	FF9FFFE7A842F-FA3E422FFF90169654C	-	-	-	1	18	205	205
Dankittipakul, Tavano, and Singtripop 2013	10136F22FFB5F-F85AE6DD059FFD39809	-	-	-	1	9	43	43
Deeleman-Reinhold 2001	FF8A860AC93EFFE765528965D D50FFD6; FF8FE1734262FFA4F-F8A4255DE29FF94; FFC9FF-C63A1BFFDDFFD5FFC6DB37324B	1	1	6	7	38	564	570
Deltshev et al. 2013	FFD9FFE35975FFA9FFF1FFC1FF-CDFF82	2	2	2	-	-	-	2
Deltshev and Wang 2016	FFB8FFF5C324E437FFF7FF910D10F-C7A	1	1	4	-	-	-	4
Elverici, Özkütük, and Kunt 2013	FFA7C048B65EFFF5AE774A45FF9F-F24A	2	2	25	-	-	-	25
Esyunin and Kazantsev 2007	5141AC78642A420DDF-7CFFDFFFFB3311	1	1	4	-	-	-	4
Felton, Judd, and Merrett 2004	FFC-CFF95D602FF82FF84FF82FFF1E808	1	1	6	-	-	-	6
Fu, Zhang, and Zhu 2009	4C1488709C14A741FF-9CFFE8FF98FFB9	1	1	17	-	-	-	17
Hayashi 1992	FF9896040E0FFF90512CFFD1FFE9FF-DC	1	2	4	-	-	-	4
Jäger 2007	2E4DFFC7FF823B04FF8434627515FFF5	-	-	-	1	1	1	1
Bosselaers and Jocqué 2013	9641FFF01026FF93FFC1CF4E123C-D00A	1	7	227	-	-	-	227
Jonsson 2005	FFB49220FFB2FF91FFD2FFA9C165D-B4A	1	1	7	-	-	-	7
Marusik and Koponen 2000	BE33BE057603DB13FFDE024BC-B37C220	1	1	1	-	-	-	1
Marusik, Omelko, and Koponen 2016	3F632B49FFD3296A0D-5CA5685D01FFB9	1	2	3	-	-	-	3
Marusik, Zheng, and Li 2008	F43D9C-12004C613DA434AC5324270C62	1	2	16	-	-	-	16
Namkung 1989	FFBDFF-D5063A094CFF937B5D13709146	1	1	3	-	-	-	3
Ono 2009	3513FF9301781725463FFF92FFA2FF89	-	-	-	1	1	1	1
Platnick and Di Franco 1992	4C59FFB0F-FE19E6FFFB7FFD0FFEDFF8B	1	6	19	-	-	-	19

Reboleira et al. 2012	D40BFFF1FF9DFF-C7FFB3FFD2FFE3B822	2	4	40	-	-	-	40
Reddy and Patel 1993	AF35FFE0FF850E68FFA5145FFFE6F-FEB	-	-	-	1	1	3	3
Ribera and de Mas 2015	6864FFE79E0A1704FFFCFF8EFF-BAF854	1	9	19	-	-	-	19
Bosmans 2011	FFE1FF97FFF5FFBFF-FA5FFC935544357	2	2	11	-	-	-	11
Saaristo 2002	FF809205FFF21921FFC75A4699324D69	-	-	-	1	2	236	236
Sankaran et al. 2017	8618FF9AFF914734FFD0FFC-F3A19FF86	1	1	17	-	-	-	17
Seo 2011	FF8BB520FFA9743EFFF97A74FF-BAF37E	2	3	9	-	-	-	9
Seyyar et al. 2016	FFA1FFC04B7C6416FFAEE21F200EFF-DC	1	1	3	-	-	-	3
Tso et al. 2005	BD41FFED-E627A13FB559E159D628093C	-	-	-	2	2	2	2
Ubick and Platnick 1991	FFB7AC78821CFFC2FFC2FFD7B-54DAD40	1	1	41	-	-	-	41
Ubick and Vetter 2005	F117A4286D45FFE7233FFFA2F-F906E1E	1	1	109	-	-	-	109
Vetter 2001	FF9BFFDBFFB54A58AC05FF-CDFFD2FFAB	1	2	320	-	-	-	320
Warui and Jocqué 2002	FFDE4808FFABFFDCFFEDFFAF-F36A181D	1	2	28	-	-	-	28
Wunderlich 2011	FFC1DD35277DCD7FBE6DFE7EFF-CAFFE0	2	3	13	-	-	-	13
Zapata and Ramírez 2010	5924FF98422FFFB9FF823E088233FFC8	1	1	1	-	-	-	1
Zhang and Fu 2010	FF806164FFE5FFDF2F48FF90FFF 6D054	-	-	-	1	1	15	15
Zhang, Fu, and Zhu 2009	FF8A9941EF368C07FFC8FFC4FF-B1A240	-	-	-	1	4	23	23
Zhao and Peng 2013	8163FFFD6B76FFA8FFA5FFEE764AF-FB6	1	1	3	2	2	3	6
Zonstein, Marusik, and Omelko 2015	FFA288738C19FFC1FFA7FF-D054431A6B	1	1	8	-	-	-	8
		55	**94**	**1636**	**32**	**112**	**1309**	**2945**

Supplementary Table 2.–Detail of our sampling sites in Thailand.

Province	Site details	Geographic Coordinates and elevation	Date
Chiang Mai	Pha Daeng NP. Riparian tropical forest.	19º37.768'N 98º57.257'E, 560m.	16-19 July 2018.
	Pha Daeng NP. Bamboo forest.	19º37.668'N 98º57.131'E, 573m.	16-19 July 2018.
	Pha Daeng NP. Mixed Teak forest.	19º34.320'N 98º57.340'E, 474m.	16-19 July 2018.
	Pha Daeng NP. Dipterocarpus forest.	19º36.132'N 98º56.980'E, 571m.	17-19 July 2018.
	Doi Inthanon NP. Cloud forest.	18º35.268'N 98º29.240'E, 2572m.	21-24 July 2018.
	Doi Inthanon NP. Montane evergreen forest.	18º30.454'N 98º30.584'E, 1605m.	21-24 July 2018.
	Doi Inthanon NP. Mixed pine forest.	18º32.606'N 98º34.479'E, 995m.	21-24 July 2018.
	Doi Inthanon NP. Mixed oak-pine tropical forest.	18º32.436'N 98º31.858'E, 1279m.	21-24 July 2018.
	Doi Suthep NP. Montane evergreen forest with pine.	18º48.502'N 98º53.528'E, 1409m.	24-28 July 2018.
	Doi Suthep NP. Mixed oak tropical forest.	18º48.164'N 98º54.081'E, 1300m.	24-28 July 2018.
	Doi Suthep NP. Mixed bamboo tropical forest.	18º49.045'N 98º55.296'E, 802m.	25-28 July 2018.
	Doi Suthep NP. Dipterocarpus forest.	18º48.780'N 98º55.928'E, 643m.	25-28 July 2018.
Phuket	Ton Sai Waterfall. Mixed bamboo tropical forest.	8º1.673'N 98º22.019'E, 144m.	29 July - 2 August 2018.
	Ton Sai Waterfall. Mixed Kerriodoxa elegans tropical forest.	8º1.816'N 98º22.375'E, 215m.	29 July - 2 August 2018.
	Bang Pae Waterfall. Mixed bamboo tropical forest.	8º2.310'N 98º23.407'E, 135m.	30 July - 3 August 2018.
	Bang Pae Waterfall. Mixed tropical forest.	8º2.353'N 98º23.365'E, 173m.	31 July - 4 August 2018.
	Siray Island. Mixed tropical forest.	7º53.355'N 98º26.083'E, 132m.	2-6 August 2018.
	Siray Island. Rubber plantation.	7º53.384'N 98º26.102'E, 104m.	2-6 August 2018.
	Siray Island. Mixed tropical forest near banana plantation.	7º53.169'N 98º26.108'E, 88m.	3-6 August 2018.
	Siray Island. Mixed tropical forest near rubber plantation.	7º53.409'N 98º26.067'E, 117m.	4 August 2018.
Krabi	Community Forest near Than Bok Khorani NP. Mixed tropical forest.	8º29.536'N 98º44.353'E, 93m.	7-12 August 2018.
	Community Forest near Than Bok Khorani NP. Mixed bamboo tropical forest.	8º29.572'N 98º44.367'E, 85m.	8-12 August 2018.
	Community Forest near Than Bok Khorani NP. Mixed young tropical forest.	8º29.655'N 98º44.001'E, 60m.	9-12 August 2018.
	Community Forest near Than Bok Khorani NP. Oil palm plantation.	8º29.592'N 98º43.907'E, 56m.	9 August 2018.