Insect Systematics and Diversity, 7(6), 1; 2023, 1–9
https://doi.org/10.1093/isd/ixad023
Method

ENTOMOLOGICAL SOCIETY OF AMERICA
SHARING INSECT SCIENCE GLOBALLY

OXFORD

Taxonomy

Building a community-based taxonomic resource for digitization of parasites and their hosts

Kathryn A. Sullivan[1,2,*], Erika M. Tucker[2,3], Nicolas J. Dowdy[2], Julie M. Allen[4], Vijay Barve[5,6], James H. Boone[†], Sarah E. Bush[7], Neal L. Evenhuis[8], Michael Hastriter[9], Jessica E. Light[10], Teresa Mayfield-Meyer[11], Barry M. OConnor[12], Jorrit H. Poelen[13,14], Gabor R. Racz[15], Katja C. Seltmann[14], Jennifer M. Zaspel[1,2]

[1]Department of Biological Sciences, Marquette University, Milwaukee, WI, USA, [2]Department of Zoology, Milwaukee Public Museum, Milwaukee, WI, USA, [3]Biodiversity Outreach Network, Flagstaff, AZ, USA, [4]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA, [5]Marine Biodiversity Center, Natural History Museum of Los Angeles County, Los Angeles, CA, USA, [6]Department of Entomology, Purdue University, West Lafayette, IN, USA, [7]School of Biological Sciences, University of Utah, Salt Lake City, UT, USA, [8]Department of Entomology, Bishop Museum, Honolulu, HI, USA, [9]Monte L. Bean Life Science Museum, Brigham Young University, Provo, UT, USA, [10]Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA, [11]Museum of Southwestern Biology, The University of New Mexico, Albuquerque, NM, USA, [12]Museum of Zoology, Insect Collection, University of Michigan, Ann Arbor, MI, USA, [13]Ronin Institute, Montclair, NJ, USA, [14]Cheadle Center for Biodiversity and Ecological Restoration, University of California Santa Barbara, Santa Barbara, CA, USA, [15]University of Nebraska State Museum, Harold W. Manter Laboratory of Parasitology, Lincoln, NE, USA, *Corresponding author, mail: kathrynsully95@gmail.com

[†]Deceased.Subject Editor: Jeffrey Lozier

Received on 17 May 2023; revised on 5 September 2023; accepted on 2 November 2023

Classification of the biological diversity on Earth is foundational to all areas of research within the natural sciences. Reliable biological nomenclatural and taxonomic systems facilitate efficient access to information about organisms and their names over time. However, broadly sharing, accessing, delivering, and updating these resources remains a persistent problem. This barrier has been acknowledged by the biodiversity data sharing community, yet concrete efforts to standardize and continually update taxonomic names in a sustainable way remain limited. High diversity groups such as arthropods are especially challenging as available specimen data per number of species is substantially lower than vertebrate or plant groups. The Terrestrial Parasite Tracker Thematic Collections Network project developed a workflow for gathering expert-verified taxonomic names across all available sources, aligning those sources, and publishing a single resource that provides a model for future endeavors to standardize digital specimen identification data. The process involved gathering expert-verified nomenclature lists representing the full taxonomic scope of terrestrial arthropod parasites, documenting issues experienced, and finding potential solutions for reconciliation of taxonomic resources against large data publishers. Although discordance between our expert resources and the Global Biodiversity Information Facility are relatively low, the impact across all taxa affects thousands of names that correspond to hundreds of thousands of specimen records. Here, we demonstrate a mechanism for the delivery and continued maintenance of these taxonomic resources, while highlighting the current state of taxon name curation for biodiversity data sharing.

Key words: taxonomy, parasite, bioinformatics, arthropod, biological nomenclature

Introduction

Classifying biological diversity provides an essential foundation for research in the life sciences. In the age of digital datasets and rapid biodiversity informatics development, it is especially important that the biological nomenclature linking biodiversity data to taxonomic names be standardized and accessible. Despite the importance of such taxonomic resources, sustainable systems for managing taxonomic lists that are persistent through time are generally narrow in scope (e.g., taxon-specific, geographically limited) or difficult to find (Miralles et al. 2020). Biological research often relies on integrating datasets, including genetic, trait, and occurrence data derived from previously published literature and databases, community science observations, and natural history collections (Peterson et al. 2015). For example, integrating occurrence records with specimen vouchers can provide a historical account of species distributions, while observation records from photo vouchers collected by community scientists supplement data for recent distribution patterns in the wake of climate change (Troudet et al. 2018, Bakker et al. 2020). Many biodiversity databases also serve as taxonomic list repositories to aggregate data and apply standardized taxonomy, such as Global Biodiversity Information Facility (GBIF: The Global Biodiversity Information Facility 2023), National Center for Biotechnology Information (NCBI; Sayers et al. 2022), Open Tree of Life (Hinchliff et al. 2015), Encyclopedia of Life (EOL; Parr et al. 2014), and iNaturalist (inaturalist.org). However, availability and maintenance of expert-verified nomenclatural resources varies greatly and taxonomic lists are often incomplete due to unavailability of expertise, organizational infrastructure and capacity, and the dynamic nature of taxonomy itself (Bánki et al. 2019). An additional complication is that taxonomic concepts, which are the author's opinion on what encompasses a taxon, are not always the same as the taxonomic name, which can only be tied directly to type specimens (Kennedy et al. 2006), and competing taxonomic concepts may lead to nomenclatural instability.

Biodiversity assessment and conservation also rely on accurate taxonomic designations and standardized application of names (Hortal et al. 2015, Vogel Ely et al. 2017). Numerous conservation entities rely on taxonomic names to confer protections; therefore, the biota must be formally described to be protected legally. With so much yet to be described and new studies resulting in constant adjusting of accepted taxonomy, it is important to continue improving methods for biodiversity description management systems as the taxonomic literature continues to grow.

Taxonomic list formatting and quality is a concern for downstream data users, but the lack of tools to standardize resources is also a major impediment to providing both taxonomic and specimen data (Nelson and Ellis 2019). Natural history collections are important sources of data that unlock tremendous amounts of data via digitization initiatives (Soltis 2017, Ball-Damerow et al. 2019). Up-to-date, expert-vetted, and accessible nomenclatural and taxonomic resources are essential in facilitating the supply and use of biological data from collections. The hesitancy to provide incomplete or dated taxonomic assignments can prevent many natural history collections from making specimen data public (Peterson et al. 2015). Additionally, these data providers often have limited resources to ensure that their taxonomic names stay current with modern scientific consensus (Vollmar et al. 2010, Tulig et al. 2012).

Many of the steps involved in creating digitally available specimen data from natural history collections have been streamlined through automation and support via the US National Science Foundation's Advancing the Digitization of Biodiversity Collections (ADBC) program (now under the Infrastructure and Capacity for Biological Research (Capacity) Program) and subsequent workflow sharing (e.g., BugFlow; https://entcollnet.github.io/BugFlow). Yet taxonomic list synthesis continues to be a time consuming step as names in collections may largely be out of date with current accepted taxonomic concepts (Tulig et al. 2012). Previous solutions for gathering taxonomic lists for new digitization projects often required building from the ground up. For projects focusing on a small number of taxa which lack available and reliable resources, new project-specific resources are often created, or a list is produced as an output after the specimens were curated and digitized (Favret 2014, Mason et al. 2020). Building taxonomic lists has proved instrumental in improving the quality of digitized data from natural history collections and other sources (Zermoglio et al. 2016, Nelson and Ellis 2019). Unfortunately, many projects create resources that are difficult for users outside of the field to find and quickly become outdated after the project ends (Ball-Damerow et al. 2019).

Numerous groups have developed databases and tools that can be used to manage taxonomic names; however, to date none have succeeded in emerging as an all-encompassing source for regularly updating names spanning all biological diversity (Zermoglio et al. 2016). There are some resources that are regularly updated, but they tend to be focused on specific taxonomic groups. For example, World Spider Catalog (WSC) houses all relevant literature for this group and maintains a list of taxonomic names as new literature is made available (World Spider Catalog 2023). It is supported by the Naturhistorisches Museum Bern and arachnological societies around the world, with updates almost daily. A great model for sustainable upkeep of taxonomic lists, this database is exceptional in its ability to provide a consistently updated catalog of names and taxonomic concepts in a diverse group. Additional standalone databases curated by taxonomic experts for various groups of organisms include: AmphibiaWeb (AmphibiaWeb 2023), Reptile Database (Uetz et al. 2021), AntBase (Agosti and Johnson 2005), International Plant Names Index (Croft et al. 1999), World Register of Marine Species (WoRMS Editorial Board 2023), and Discover Life bee species guide and world checklist (Ascher and Pickering 2020). These communities are active and successfully synthesize taxonomic lists for public use. However, independent resources can be more difficult to find compared with the general biodiversity databases, and often provide data in different formats, making dataset synthesis problematic for downstream users.

Other efforts to improve visibility and coordination of taxonomic working groups have been successful in centralizing and updating taxonomic lists over time. The Species File Group (University of Illinois; Illinois Natural History Survey) has developed a model for collaborative assembly of taxonomic lists for many insect groups via TaxonWorks (TaxonWorks Community 2022). While access to these lists and integration into other repositories has been an ongoing process, they provide a valuable role in assembling taxonomic lists in a comprehensive database with standard Darwin Core (Wieczorek et al. 2012) fields while providing support for taxonomists. Accessibility and ease of use are still challenges for many providers and users of these taxonomic lists, as the TaxonWorks model requires the community to adopt a specific software when many taxonomists and data users have preexisting management systems.

Some taxonomy services index taxonomic names from multiple sources (aggregators) into one data stream, catalog, or taxonomic list. These aggregators seek to serve as a general repository for all types of taxonomic groups, such as Catalogue of Life (CoL; Hobern et al. 2021, Bánki et al. 2023), the Integrated Taxonomic Information System (ITIS 2023), Global Names (Pyle 2016), and the GBIF Backbone Taxonomy (GBIF Secretariat 2022). However,

most of these aggregators rely on the smaller databases such as WSC to share data to the aggregator site and provide updates over time. Smaller name databases send data to aggregators in various formats and some of these lists are only indexed once or very infrequently. Many become obsolete over time and cease updates or routine maintenance. Both aggregators and smaller databases encounter similar problems; it is difficult to find experts that can curate taxonomy and maintain digital databases across the tree of life, and taxonomic groups with few people actively conducting research have fewer resources to invest in generating and disseminating their data (Ball-Damerow et al. 2019).

Besides taxonomic lists themselves, downstream tools, commonly implemented as R packages, have been developed to access taxonomic databases and standardize names that come from multiple sources. These range from the more general, (e.g., *taxlist;* Alvarez and Luebert 2018) to the more taxon-specific, (e.g., *vegdata*; Jansen and Dengler 2010). There are many data and taxon name management tools where the same basic functionality has been recreated in different packages for different databases (Grenié et al. 2023). Although these tools can improve access to names and align data from multiple sources, the outputs of these endeavors often solely remain in a research publication rather than a standalone resource that can be reused. Therefore, if a user wants to use a synthesized taxonomic list for publicly available data, they may have to repeat the same process of consolidating and standardizing names as was previously completed. These persistent issues of assembling and maintaining taxonomic lists apply across many disciplines of biological research, yet attempts to solve this problem are rarely successful over the long-term (Garnett et al. 2020).

Taxonomically related problems are often most pronounced in data deficient groups, most notably invertebrate groups (Chapman 2009). Parasitism is one of the most common lifestyles exhibited by organisms on the planet (Dobson et al. 2008, Weinstein and Kuris 2016) and has great economic and health implications (Gubler 1998, Weissenböck et al. 2010), but data about parasitic arthropods are especially underrepresented among digitized specimens (SCAN 2018). The Terrestrial Parasite Tracker (TPT) is a multi-institutional collection digitization project funded by the National Science Foundation's ADBC Thematic Collections Network program focused on arthropod parasites and their vertebrate hosts. A crucial goal of the TPT project was to provide expert-vetted and accessible taxonomic lists for each of the parasite groups targeted in the project that can be used as a reference taxonomy. These taxonomic lists encompass multiple diverse arthropod groups totaling thousands of species including Acari (2 groups: ticks and mites), Siphonaptera (fleas), Phthiraptera (lice), and Diptera (select biting fly families). Additionally, taxonomic lists for vertebrate hosts of these parasites were also required for standardization of biotic interaction data. It is essential that expert-verified taxonomic names for these groups are accessible and can be easily applied to the specimen data generated via TPT. To achieve this goal, TPT developed and managed curated lists of taxonomic names that project participants used as a reference for curating the taxonomic names in their databases as well as for identifications. We produced a dataset that includes taxonomic names aligned from multiple sources making it easier to use and integrate into future biodiversity studies. Here, we outline a workflow for managing curated and sustainable taxonomic lists, designed during the TPT digitization project. These lists are easily accessible, immediately usable, citable, and sustainable, and we propose that these workflows can be used as a template for other taxonomic groups.

Materials and Methods

The taxonomy synthesis workflow we created for the TPT network consists of the following stages: ingestion, data quality assurance, primary review, reconciliation analysis, publishing, continuous review, and maintenance (Fig. 1). Each step of the process is detailed as it was applied to our case study.

Ingestion

Relevant names were obtained from openly accessible databases (ITIS), digital aggregators (GBIF, CoL), and lists from taxonomic experts within the TPT network. These resources were compiled from

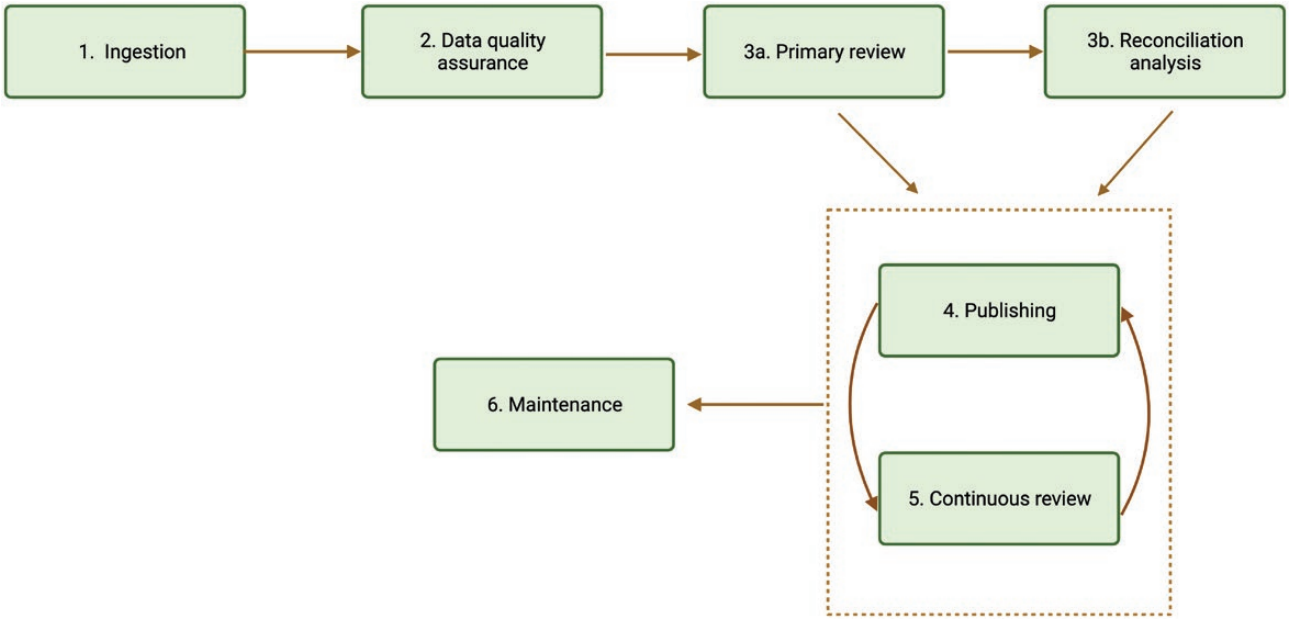

Fig. 1. Taxonomic Workflow produced by the TPT group. Initial steps (1–3) are followed by a revolving Publishing and Continuous Review cycle (4–5). The Publishing and Continuous Review cycle are necessary for long-term Maintenance (6) and stability. Created with BioRender.com.

a variety of flat file formats (e.g., .tab, .csv, .txt, .xlsx, .docx, .pdf, etc.). Some taxonomic datasets could be exported with specified data fields or modified to Darwin Core standards, while other lists were provided in nonstandardized formats as experts had unique systems that required more extensive reformatting.

Taxonomy resources for each of these focal groups had varying levels of active expertise available, and many taxonomic experts were involved in TPT as data providers from collections as well. Several taxonomic groups examined during this project had limited sources of data, much of which were not publicly accessible or accurate. When comprehensive lists for taxonomic groups were not publicly available, independent resources were gathered based on expert knowledge available to the TPT group.

Data Quality Assurance

After taxonomic resources from all sources were ingested, lists for each major taxonomic group were combined and quality checked. The quality assurance process involved checking taxonomic name validity per each source and removing duplicate or erroneously included names. This was primarily accomplished using the R package, *taxotools* (Barve 2021), which was specifically developed for the TPT project. Using *taxotools*, we wrote custom scripts to transform the files received in the ingestion phase into the discipline standard Darwin Core taxon class and HTML format to facilitate list combination and name comparison for data providers curating specimen data. Definitions for each column and taxonomic term were provided for clear guidelines on use of each field. The tools and scripts used in this phase were used once to create a format that allows for ease of future changes by experts without needing additional technical processes and are not required to use this resource or its outputs. Once the lists were completed, they were delivered to taxonomic experts for review in the subsequent steps of the workflow with the ability to maintain the lists in their native systems. Lists are also in HTML format with discrepancies and potential errors flagged so experts could review them in a format more like a traditional revisionary manuscript.

We assessed congruence of various taxonomic lists by first concatenating all lists, so each name was listed once every time it appeared in a resource with standard columns and fields. Next, for groups that had multiple expert sources, we determined which source should take priority, if any, in our final reconciliation of names to use for TPT digitization as the standard. The standard source for a given group was chosen based primarily upon the advice of TPT experts (when available), and then upon the perceived modernity of the resource (whether it was currently or recently maintained). Instead of choosing one list over another, this process allowed all published names to be accounted for, whether or not the sources agreed on validity or application of names to concepts. These quality checked taxonomic lists are directly tied to the experts who assemble and maintain the original ingested sources, as well as the history of TPT personnel involved in producing synthesized versions via the dataset metadata.

R scripts used for each taxonomic group are available in the TPT Resource Hub (https://github.com/njdowdy/tpt-taxonomy) and detailed in Supplementary Material (S1).

Primary Review

Five quality checked taxonomic lists were sent back to designated taxonomic experts to review and clarify any discrepancies identified in the quality assurance phase. We found this part of the process was the most time consuming, especially for groups with numerous competing sources to reconcile. While the TPT project did not have taxonomic experts available for all groups, the taxonomic experts that were involved contributed extensive expertise for Siphonaptera, Acari, Ixodida, Phthiraptera, and Diptera. The Siphonaptera list underwent numerous revisions within this phase, primarily completed by coauthor Hastriter, over the course of 6 mo. For Phthiraptera, the chewing louse checklist by Price et al. (2003) was converted into an electronic format as a first version with some updates by TPT experts. While these lists were in review by experts, the quality checked versions were available to the TPT participants so that the digitization process could begin. Upon completion of primary expert review and list return, new versions of the taxonomic lists were disseminated to TPT participants via our publication outlets (i.e., email listserv, TPT Resource GitHub, and Zenodo) for database updates. Vertebrate lists were already reviewed by experts prior to integration with the arthropod lists and did not require a primary review step.

Reconciliation Analysis

After the taxonomic lists were quality checked and reviewed, they were compared to data held by GBIF to check for completeness. GBIF was selected as the external resource because it integrates many other resources (e.g., CoL, ITIS) and is used by many consumers of taxonomic information. The GBIF Backbone Taxonomy was downloaded (GBIF.org, Access Date: 13 Apr 2021) and imported into a SQL database alongside the TPT taxonomic lists. A series of SQL queries were used to align these 2 resources using the canonical names of each taxonomic record with species or subspecies ranks. Counts of unique and overlapping names were generated for each resource, including where names were considered valid, invalid, or in disagreement between the compared resources (Fig. 2). Concordance means names had the same taxonomic status (regardless of validity) and were present in both lists. Discordance between taxonomic resources is due to either disagreement in the status of names ("disagreement") or names being recorded in only 1 list ("uniqueness").

Publishing

A GitHub repository (https://github.com/njdowdy/tpt-taxonomy) was created to host the taxonomic lists produced by TPT and serve as a resource hub so materials would be available to anyone in the community. GitHub was used to create the resource hub because it is freely available, allows for multiple people managing and contributing to projects that can be openly sourced, and has the potential for sustainability beyond just individual interest. In October 2021 at the beginning of the third year of the TPT project, TPT taxonomic lists that were expert reviewed and/or quality checked as well as workflow protocols were made available via GitHub as TPT Taxonomic Resource v1. Additionally, TPT taxonomic lists were published as an open-source dataset on Zenodo (Dowdy et al. 2021) to increase visibility within the scientific community. The lists are accompanied by information on appropriate attribution, data providers, and version history for each taxon-specific list. Those who use our taxonomic lists as a resource are asked to use the Zenodo DOI to cite this work in addition to any specific taxonomic sources used. This methodology creates readily accessible taxonomic lists that are now available in other online platforms including Catalogue of Life ChecklistBank (Bánki et al. 2023), Global Biotic Interactions (GloBI; Poelen et al. 2014), Global Names (Pyle 2016), and BugFlow; (https://entcollnet.github.io/BugFlow/). TPT and unaffiliated data providers can access these taxonomic lists and modify them for their own databases or research platforms. These lists have also been added in part to the Arctos Source taxonomy

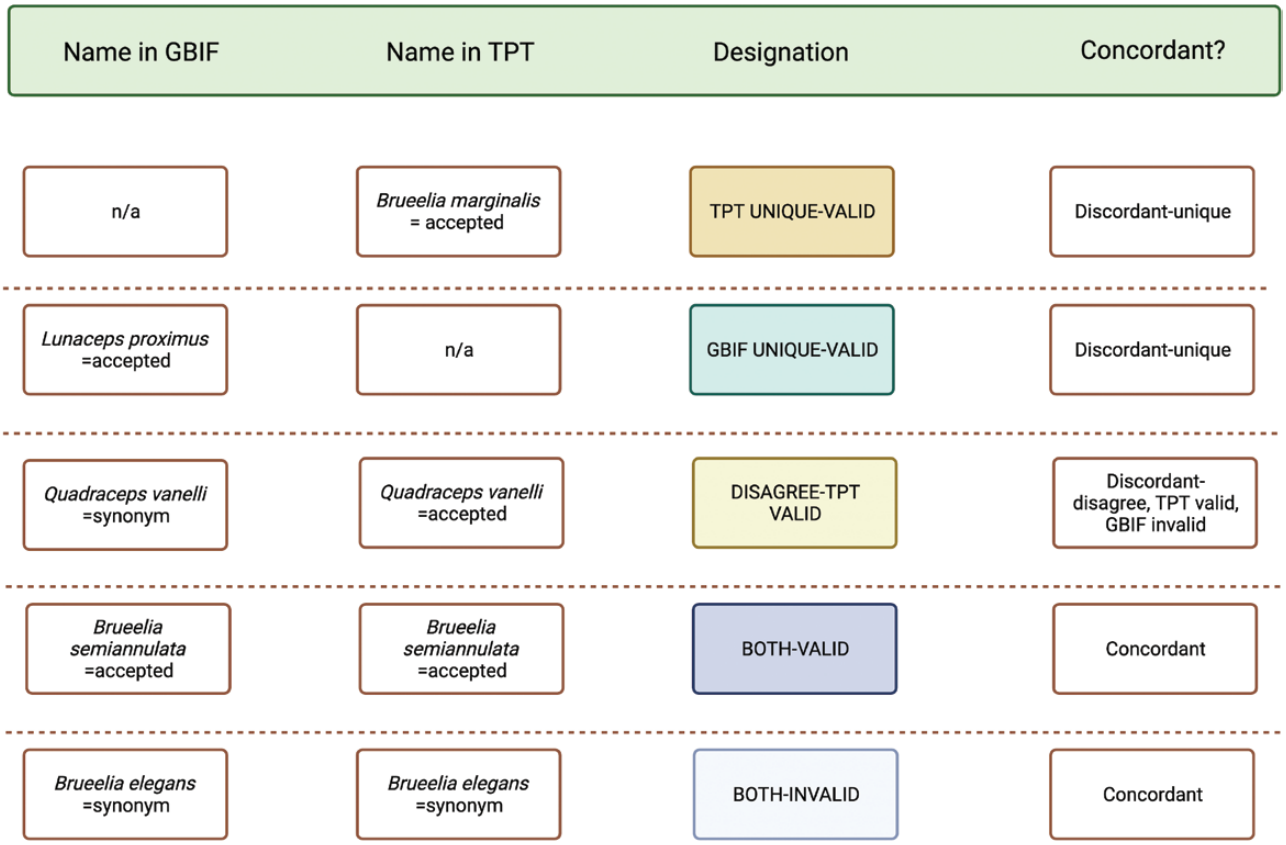

Fig. 2. Example for possible name designations in GBIF and TPT reconciliation analysis. Canonical names appearing in GBIF and/or TPT lists are designated as either valid or invalid, where they are found, and whether the designations are concordant or discordant between lists. Discordant is either categorized as "unique" or "disagreement," where concordant refers to whether the designations are the same, regardless of validity. Created with BioRender.com.

(https://arctos.database.museum/info/ctDocumentation.cfm?table=cttaxonomy_source) and all TPT data providers using Arctos for collection management can ensure names are updated via the TPT standard. We have also worked with SCAN to create additional versions of the name files that are formatted specifically to be compatible with Symbiota-based systems (available for download on GitHub as a tagged version). Symbiota also now has a tool for linking to CoL taxonomy which can be used for lists available on CoL. Custom versions for Ke Emu and Specify users in the TPT network were also made available for import into collections management systems.

Continuous Review

After the primary review and publishing phases were completed, new versions of taxonomic lists can now be updated and revised with future taxonomic changes. This largely consists of keeping up with literature to add, remove, and modify names as necessary, but can be done on more frequent intervals or as official releases on an expert's preferred timetable. Updates can be recorded via new publications, pull requests on GitHub, or continued personal communication from experts amidst the growing taxonomic resource user base. Each release published on Zenodo is citable with a DOI, and users can reference datasets from a specific point in time. Additionally, datasets remain available on GitHub as unique file and list versions that are uploaded and shared through the repository, which downstream users can continuously index.

Maintenance

The maintenance portion of this workflow consists of continued list review by experts, management by designated users, and publishing so this resource can remain useful and accessible. Maintenance of the GitHub repository during the project period was mediated by several TPT personnel who collaboratively manage the files but is now increasingly supported by the taxonomic communities. All individuals involved in the process of providing taxonomic data, managing the lists, and updating can be found on the Resource Hub. Management permissions of the repository can be assigned to additional users as appropriate, and collaboration is encouraged to assist in repository maintenance and growth.

Results

TPT Taxonomic Names Lists

Six taxonomic names lists and an additional 6 mirrored lists appended as *-symbiota are available as downloadable files via the GitHub repository and Zenodo publication (Table 1, Dowdy et al. 2021). The *-standardized lists contain the same information as other files but have regulated columns in a consistent arrangement for ease of aggregator ingestion. Since these lists are actively maintained (i.e., in continuous review and publishing), files may undergo additional rounds of expert review and be published at the discretion of the expert source. After additional expert review, new versions are added to the GitHub, but the old versions will remain available. The non-Ixodida Acari (mites), Ixodida (ticks), Siphonaptera (fleas), and files for vertebrate host mammals and birds have completed primary review and are available. The Phthiraptera file has also completed the primary review phase; however, another comprehensive review by the TPT louse experts is underway to include host association names and update the chewing louse list based on roughly

2000 publications related to louse taxonomy or host associations that have been published since the Price et al. (2003) checklist (Smith et al. 2023). The Diptera file is sourced from Systema Dipterorum (Evenhuis and Pape [eds] 2021) and filtered to only parasitic families: Carnidae, Ceratopogonidae, Culicidae, Fanniidae, Glossinidae, Hippoboscidae, Muscidae, Nycteribiidae, Oestridae, Psychodidae, Simuliidae, Streblidae, and Tabanidae. Direct links to each source are available in Supplementary File 1.

Reconciliation Analysis

We reconciled our original expert-vetted lists with names available in GBIF (Fig. 3, Supplementary Fig. S2). For some of these lists our expert source is also a contributing source for the GBIF Backbone Taxonomy, so the expected amount of overlap is higher (i.e., Systema Dipterorum). The degree of concordance for valid and invalid names between the TPT expert sources and GBIF is 58.9% (n = 115,151) across all names lists. The most concordant list is the parasitic dipteran families with 70.6% (n = 46,906) agreement, while the least concordant is Ixodida with 44.2% (n = 3,946). Overall, discordance is 41.1% (n = 115,151), with "uniqueness" contributing the most to discordance (86.5%; n = 47,297). Disagreement is absent in Ixodida, and most dominant in Diptera (7.5%; n = 46,906). The number of unique names is greater than disagreed upon names in each list. Many of our TPT expert lists are not comprehensive for synonyms and other invalid names compared to all other sources indexed by GBIF. This indicates that the largest difference between our lists and GBIF was not in valid names, but in maintaining records of invalid names. For example, 45.5% (n = 3,946) of the names for Ixodida are invalid in GBIF only, but the proportion of unique valid names for either GBIF or TPT is 10.3% (n = 3,946). Overall, the Disagreement category still comprises 5.5% (n = 115,151) of the total names found in our lists and GBIF, which amounts to 6,373

Table 1. TPT Taxonomic Names Lists from the TPT Taxonomic Resource v1 with number of total names, original expert sources, and steps completed in the TPT Taxonomy Workflow

Taxon list	Number of names	Sources	Steps completed in workflow
Non-Ixodida Acari	30,148	University of Michigan	1–4
Siphonaptera	3,794	Lewis World Species List; Brigham Young University	1–4, 5 in progress
Phthiraptera	7,728	TPT Louse Experts	1–4, 5 in progress
Ixodida	969	US National Tick Collection; Georgia Southern University	1–4
Diptera	~40,000	Systema Dipterorum/Bishop Museum, Adler Simuliidae Collection/CUAC, ITIS	1–3
Vertebrate Hosts	33,674 birds; 7,894 mammals	University of Florida	1–2,4 (outside of TPT taxonomy experts)

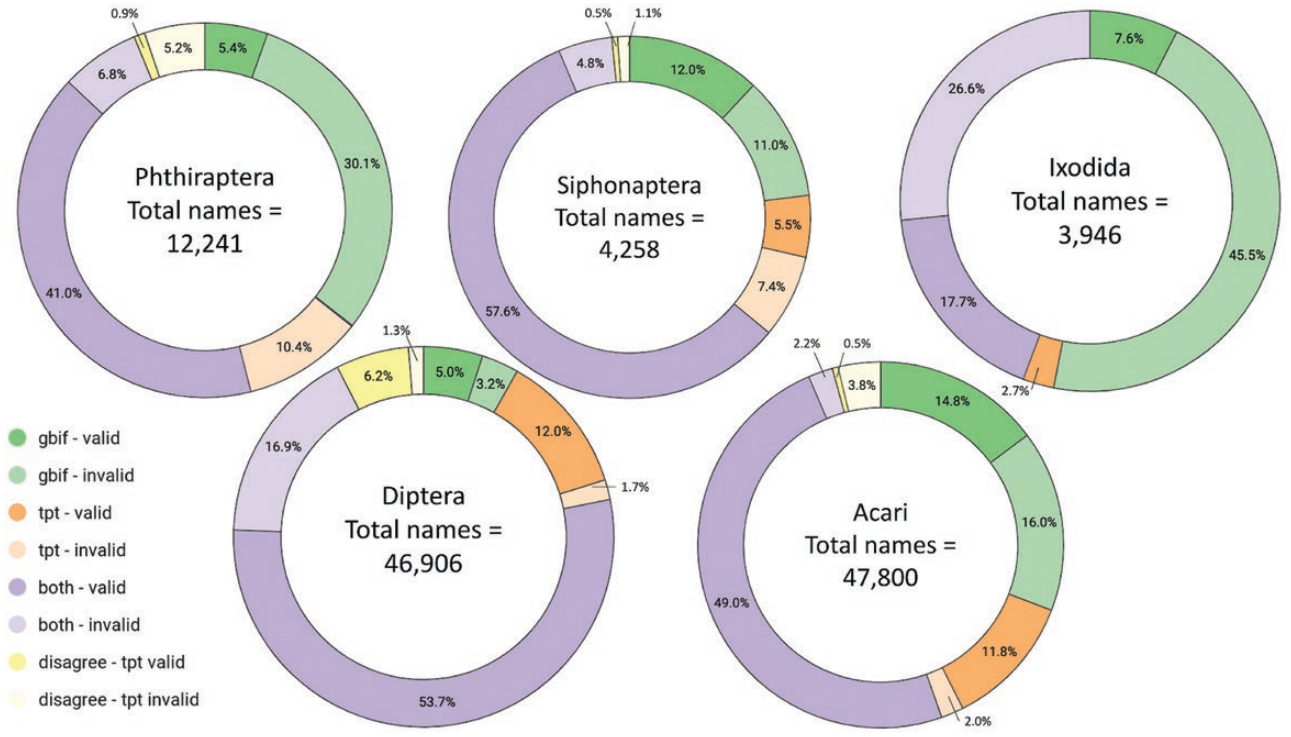

Fig. 3. Reconciliation analysis between names available on GBIF and TPT expert-verified lists. Each name present in these sources was either designated as unique to GBIF, unique to TPT, agreed in both, disagreed between both, and whether the name is valid or invalid per these sources. Percentage of total names is displayed for all categories. Total number of names for each group is listed in the center of each circle plot.

instances within our case study. These 6,373 disagreeing names are linked to over 100,000 specimen records at this time (GBIF.org, accessed 20 Aug 2023).

TPT Resource Hub

In October 2021, during the TPT project's third year, all available TPT taxonomy files were published via GitHub and Zenodo. The GitHub site serves as a general resource hub for names lists, outline of this workflow for assembling taxonomic names lists, scripts used to clean files and standardize formatting, and a record of Zenodo publications as new version history. Contact information is provided for the current TPT managers of these resources and GitHub administrators in addition to expert contacts for questions about the content of the names lists. This resource hub facilitates future maintenance and sustainability of these assets beyond the original TPT project.

Integration With Other Repositories

For some TPT names lists that were not previously serving as a source file for GBIF and CoL during the reconciliation analysis, we provided files for ChecklistBank (checklistbank.org) to integrate into the GBIF/CoL backbone. The Lewis World Species Flea List (Hastriter and Bossard 2023) was added while efforts to remove outdated sources are in progress. Additionally, a new version of Systema Dipterorum (v4) is published, which was available in advance to TPT providers and is now publicly available via ChecklistBank and the site diptera.org (Evenhuis and Pape [eds] 2023).

Discussion

Here, we present an agile method using lightweight tools as a lower cost solution for collaboratively assembling and maintaining taxonomic name lists. The workflow uses basic csv formatting and freely available platforms for taxonomic experts to manage publicly available name lists as they generate new knowledge. This method is complementary to other name catalog systems that may require a longer time commitment and great investment from the data providers (e.g., TaxonWorks) than is feasible on short-term grant funded projects. However, there are tradeoffs, as TaxonWorks and others have more sustainable funding, and software enhancements can provide quality control checks that may improve data quality in the long-term. The agility of this method produced a TPT workflow to rapidly collate and disseminate names lists that were not publicly available at the start of the TPT project, which provides greater visibility and accessibility to the taxonomies of parasitic arthropods.

This resource also demonstrates a methodology for assembling information about taxon names from an expert community and making it available for ingestion into name catalogs that have a much broader taxonomic scope. The utility of smaller taxon-centered name catalogs is powerful as updates of names in these larger systems may be more infrequent or lack relevant metadata about versions. Taxon-focused catalogs may also provide a community focused effort around taxon names, ensuring greater sustainability of resources. Smaller catalogs should be used as a primary data source for aligning taxon names for data analysis, while aggregator taxonomy systems can reuse these primary data sources with the aim to improve the taxonomic coverage of specialist groups. For example, name alignment tools like *nomer* (Salim and Poelen 2023), which indexes GBIF and the TPT Taxonomy in addition to the names generated from digitized TPT collections, can aid in synthesizing names in individual datasets.

In addition to publishing names via GitHub and Zenodo, we looked to integrate directly into other repositories that GBIF can harvest (i.e., Global Names Architecture, ChecklistBank). At this stage these are still manual direct deposits as there are limitations for harvesting updates from our TPT taxonomic name lists while still using these existing publishing platforms. Ideally a more robust database could house and serve updates automatically via API, but that vastly exceeds the scope of our project's main objective to produce digital specimen records. However, this integration at least ensures some long-term accessibility and visibility of the lists. Additionally, these improvements will accurately associate accepted taxonomic names with occurrence record data already available via GBIF, which impacts thousands if not millions of records.

It is essential that taxonomic resources from various sources are properly maintained to prevent confusion. The reconciliation analyses point to larger issues in traditional taxonomy maintenance where multiple sources exist, and it is unclear how to choose one when they disagree without additional expert input. It is especially critical to track invalidated names as this causes a large proportion of discordance between datasets. This is of particular interest to our network as invalid names may be associated with many specimens in collections, and improper tracking of those names adds to the already large time taken to curate records for digitization. Ongoing and future work will seek to add the invalid names often found across different sources to the TPT verified lists for completion. Although source discrepancy may not impact most names examined for our network, the number of names still reaches into the thousands, which has implications for countless other downstream analyses that use conflicting lists. Despite the overall low level of discordance due to disagreement, our results justify the cost and effort of undertaking the endeavor to harmonize lists, particularly for the taxonomic groups needed for TPT. As scientific names for species are hypotheses, true disagreement may exist between credible sources that cannot be resolved to 1 taxonomic concept and need to be aligned accordingly (Pyle 2004, Franz et al. 2015). We do not expect the discrepancy problem to ever be fully resolved in the larger landscape of taxonomy, but this workflow provides a method for tracking taxonomic viewpoints at a specific point in time with citations, as each versioned names list is associated with taxonomic authorities. The availability of these kinds of name lists creates an explicit link between biodiversity data and the taxonomic concepts being followed to format, resolve, and align the taxonomic names.

The sustainability of our effort is the greatest challenge, but digital infrastructure and data management systems improvements should increase persistence of these resources. Many digital platforms are available for sharing resources, but few allow for ease of update and/or collaborative work without technical expertise or dedicated support from data users. This process reveals the complex effort involved in gathering taxonomic lists and supporting those who do taxonomic research, and 1 solution likely will not fit for all. This workflow, though generally applicable to other projects, is difficult to repeat even with our own lists, and solutions for taxonomic name resolution can be further improved. There is still a lot of work needed to ensure future sustainability of the taxonomic resources described herein, but collaborations with existing organizations that have long-term funding show promise for the sustained utility of these resources across all disciplines seeking accurate taxonomic names lists. Long-term sustainability of ever-changing taxonomic lists is undoubtedly difficult, but the backbone for facilitating this type of taxonomic work is now constructed and available. Here we provide an infrastructural resource that will facilitate this task for taxonomists, data providers, and data managers working with any taxonomic group.

Supplementary material

Supplementary material is available at *Insect Systematics and Diversity* online.

Acknowledgments

We firstly wish to acknowledge Jim Boone for all his contributions to the TPT project and insect collections over his career who sadly passed away during this project period before this manuscript was completed. We would like to thank Rich Pyle (BPBM), David Pecor (WRBU), Michael Caterino (CUAC), Jason Weckstein (ANSP), Jessica Bird (NMNH), Joe Miller (GBIF), and Neil Cobb (BON) for providing database support and additional taxonomic resources. Thank you to all the TPT data providers and network members for input on handling taxonomic names and digitization protocols. We also thank 2 anonymous reviewers for their comments improving this manuscript.

Funding

This work was supported by National Science Foundation (US) grants DBI-1901932 to JMZ, DBI-1901926 to KCS, DBI-1902031 to JMA, DBI-1811897 to NJD, DBI-1902023 to SEB, DBI-1901916 to JEL, DBI-1902113 to BMO, DBI-1901928 to NLE. Additionally, DBI-1902048 supported MH and DBI-1901911 supported GRR.

Author Contributions

Kathryn Sullivan (Conceptualization [Equal], Data curation [Equal], Formal analysis [Lead], Investigation [Lead], Methodology [Equal], Project administration [Lead], Validation [Lead], Visualization [Lead], Writing—original draft [Lead], Writing—review & editing [Lead]), Erika Tucker (Conceptualization [Equal], Data curation [Equal], Methodology [Equal], Project administration [Equal], Resources [Equal], Writing—review & editing [Supporting]), Nicolas Dowdy (Conceptualization [Equal], Data curation [Equal], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Resources [Equal], Visualization [Supporting], Writing—review & editing [Supporting]), Julie Allen (Conceptualization [Equal], Data curation [Supporting], Resources [Supporting], Supervision [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Vijay Barve (Conceptualization [Equal], Data curation [Supporting], Methodology [Supporting], Project administration [Supporting], Resources [Equal], Software [Equal], Validation [Supporting], Writing—review & editing [Supporting]), Sarah Bush (Conceptualization [Equal], Data curation [Supporting], Methodology [Supporting], Resources [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Neal Evenhuis (Conceptualization [Equal], Data curation [Supporting], Resources [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Michael Hastriter (Conceptualization [Equal], Data curation [Supporting], Resources [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Jessica Light (Conceptualization [Equal], Investigation [Supporting], Methodology [Supporting], Writing—review & editing [Supporting]), Teresa Mayfield-Meyer (Conceptualization [Equal], Data curation [Equal], Investigation [Equal], Methodology [Equal], Writing—review & editing [Supporting]), Barry OConnor (Conceptualization [Equal], Data curation [Supporting], Resources [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Jorrit Poelen (Conceptualization [Equal], Data curation [Supporting], Investigation [Supporting], Methodology [Supporting], Software [Supporting], Writing—review & editing [Supporting]), Gabor Racz (Conceptualization [Equal], Data curation [Supporting], Resources [Supporting], Validation [Supporting], Writing—review & editing [Supporting]), Katja Seltmann (Conceptualization [Equal], Investigation [Supporting], Methodology [Supporting], Resources [Supporting], Writing—review & editing [Supporting]), and Jennifer Zaspel (Conceptualization [Equal], Funding acquisition [Lead], Project administration [Equal], Writing—review & editing [Supporting])

Data Availability

All taxonomic lists and previous versions that were used at any point for TPT digitization are found in the GitHub for the project (https://github.com/njdowdy/tpt-taxonomy). Metadata related to list generation and review are found in the README as well as contact information for current site managers. Scripts used to combine and clean lists are found in the repositories detailed in the Supplementary Material File S1 and may be found as links on GitHub. SQL queries and additional information on methods used in the reconciliation analysis can be found on GitHub at https://github.com/njdowdy/tpt-taxonomy/tree/main/sql%20queries. The raw numbers output from the analysis are in Supplementary File S2.

References

Agosti D, Johnson N. Antbase. World Wide Web electronic publication. antbase. org, version (05/2005). 2005.

Alvarez M, Luebert F. The taxlist package: managing plant taxonomic lists in R. Biodivers Data J. 2018:6:e23635. https://doi.org/10.3897/bdj.6.e23635

AmphibiaWeb. Berkeley (CA):University of California,; 2023. https://AmphibiaWeb.org

Ascher J, Pickering J. Discover life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila). 2020.

Bakker FT, Antonelli A, Clarke JA, Cook JA, Edwards SV, Ericson PGP, Faurby S, Ferrand N, Gelang M, Gillespie RG, et al. The Global Museum: natural history collections and the future of evolutionary science and public education. PeerJ. 2020:8:e8225. https://doi.org/10.7717/peerj.8225

Ball-Damerow JE, Brenskelle L, Barve N, Soltis PS, Sierwald P, Bieler R, LaFrance R, Ariño AH, Guralnick RP. Research applications of primary biodiversity databases in the digital age. PLoS One. 2019:14(9):e0215794. https://doi.org/10.1371/journal.pone.0215794

Bánki O, Hobern D, Döring M, Remsen D. Catalogue of Life Plus: a collaborative project to complete the checklist of the world's species. Biodivers Inf Sci Stand 2019:3:e37652.

Bánki O, Roskov Y, Döring M, Ower G, Hernández Robles DR, Plata Corredor CA, Stjernegaard Jeppesen T, Örn A, Vandepitte L, Hobern D, et al. Catalogue of life checklist. Annual checklist 2023. Catalogue of life; 2023. https://doi.org/10.48580/dfsr

Barve V. taxotools: tools to handle taxonomic lists. R package version 0.0.79; 2021. https://CRAN.R-project.org/package=taxotools

Chapman AD. Numbers of living species in Australia and the world. Report for the Australian Biological Resources Study. Australian Government, Department of the Environment, Water, Heritage, and the Arts; 2009.

Croft J, Cross N, Hinchcliffe S, Lughadha EN, Stevens P, West J, Whitbread G. Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. Taxon. 1999:48:317–324.

Dobson A, Lafferty KD, Kuris AM, Hechinger RF, Jetz W. Homage to Linnaeus: how many parasites? How many hosts? Proc Natl Acad Sci USA. 2008:105(supplement_1):11482–11489. https://doi.org/10.1073/pnas.0803232105

Dowdy NJ, Barve V, Mayfield-Meyer T, Sullivan K, Zaspel JM. TPT taxonomic resource v1.0.3. Zenodo. 2021.

Evenhuis NL, Pape T, editors. Systema Dipterorum, version 3.1; 2021. http://diptera.org/

Evenhuis NL, Pape T, editors. Systema Dipterorum, version 4.0; 2023. http://diptera.org/

Favret C. Cybertaxonomy to accomplish big things in aphid systematics: cybertaxonomy in aphid systematics. Insect Sci. 2014:21(3):392–399. https://doi.org/10.1111/1744-7917.12088

Franz NM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B. Reasoning over taxonomic change: exploring alignments for the Perelleschus use case. PLoS One. 2015:10(2):e0118247. https://doi.org/10.1371/journal.pone.0118247

Garnett ST, Christidis L, Conix S, Costello MJ, Zachos FE, Bánki OS, Bao Y, Barik SK, Buckeridge JS, Hobern D, et al. Principles for creating a single authoritative list of the world's species. PLoS Biol. 2020:18(7):e3000736. https://doi.org/10.1371/journal.pbio.3000736

GBIF Secretariat. GBIF Backbone Taxonomy Checklist dataset; 2022. https://doi.org/10.15468/39omei

GBIF: The Global Biodiversity Information Facility. What is GBIF? 2023. https://www.gbif.org/what-is-gbif

Grenié M, Berti E, Carvajal-Quintero J, Dädlow GML, Sagouis A, Winter M. Harmonizing taxon names in biodiversity data: a review of tools, databases and best practices. Methods Ecol Evol. 2023:14:12–25.

Gubler DJ. Resurgent vector-borne diseases as a global health problem. Emerg Infect Dis. 1998:4(3): 442–450. https://doi.org/10.3201/eid0403.980326

Hastriter MW, Bossard RL. Robert E. Lewis world species flea (Siphonaptera) list. Version 2023-06-22; 2023.

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci USA. 2015:112(41):12764–12769. https://doi.org/10.1073/pnas.1423041112

Hobern D, Barik SK, Christidis L, Garnett ST, Kirk P, Orrell TM, Pape T, Pyle RL, Thiele KR, Zachos FE, et al. Towards a global list of accepted species VI: the catalogue of life checklist. Org Divers Evol. 2021:21:677–690.

Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu Rev Ecol Evol Syst. 2015:46(1):523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Integrated Taxonomic Information System (ITIS). Integrated Taxonomic Information System (ITIS); 2023. https://www.itis.gov/

Jansen F, Dengler J. Plant names in vegetation databases – a neglected source of bias. J Veg Sci. 2010:21(6):1179–1186. https://doi.org/10.1111/j.1654-1103.2010.01209.x

Kennedy DJ, Hyam R, Kukla R, Paterson T. Standard data model representation for taxonomic information; 2006. https://home.liebertpub.com/omi. https://www.liebertpub.com/doi/10.1089/omi.2006.10.220.

Mason Jr S, Betancourt I, Gelhaus J. Importance of building a digital species index (spindex) for entomology collections: a case study, results and recommendations. BDJ. 2020:8:e58310.

Miralles A, Bruy T, Wolcott K, Scherz MD, Begerow D, Beszteri B, Bonkowski M, Felden J, Gemeinholzer B, Glaw F, et al. Repositories for taxonomic data: where we are and what is missing. Syst Biol. 2020:69(6):1231–1253. https://doi.org/10.1093/sysbio/syaa026

Nelson G, Ellis S. The history and impact of digitization and digital data mobilization on biodiversity research. Philos Trans R Soc B. 2019:374:20170391.

Parr CS, Wilson N, Leary P, Schulz K, Lans K, Walley L, Hammock J, Goddard A, Rice J, Studer M, et al. The Encyclopedia of Life v2: providing global access to knowledge about life on earth. Biodivers Data J. 2014:2:e1079.

Peterson AT, Soberón J, Krishtalka L. A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecol. 2015:15:15. https://doi.org/10.1186/s12898-015-0046-8

Poelen JH, Simons JD, Mungall CJ. Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. Ecol Inf. 2014:24:148–159. https://doi.org/10.1016/j.ecoinf.2014.08.005

Price RD, Hellenthal RA, Palma RL, Johnson KP, Clayton DH. The Chewing Lice: world checklist and biological overview. Illinois Natural History Survey Special Publication; 2003.

Pyle RL. Taxonomer: a relational data model for managing information relevant to taxonomic research. Phyloinformatics. 2004:1:1–54.

Pyle RL. Towards a Global Names Architecture: the future of indexing scientific names. ZooKeys. 2016:550:261–281.

Salim JA, Poelen J. globalbioticinteractions/nomer:0.4.11. Zenodo; 2023. https://doi.org/10.15468/39omei

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022:50(D1):D20–D26. https://doi.org/10.1093/nar/gkab1112

SCAN. Symbiota collection of arthropods network; 2018. www.scan-bugs.org

Smith VS, Broom Y, Dalgleish R. Phthiraptera.info (web resource); 2023 [accessed 2023 Apr 5]. http://phthiraptera.myspecies.info/

Soltis PS. Digitization of herbaria enables novel research. Am J Bot. 2017:104(9):1281–1284. https://doi.org/10.3732/ajb.1700281

TaxonWorks Community. TaxonWorks [software and supporting resources], https://taxonworks.org. Species File Group (maintainers, https://speciesfilegroup.org); 2022. Accessible at: https://github.com/SpeciesFileGroup/taxonworks

Troudet J, Vignes-Lebbe R, Grandcolas P, Legendre F. The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it? Syst Biol. 2018:67(6):1110–1119. https://doi.org/10.1093/sysbio/syy044

Tulig M, Tarnowsky N, Bevans M, Kirchgessner A, Thiers B. Increasing the efficiency of digitization workflows for herbarium specimens. ZooKeys. 2012:209:103–113. https://doi.org/10.3897/zookeys.209.3125

Uetz P, Koo MS, Aguilar R. A quarter century of reptile and amphibian databases. 2021.

Vogel Ely C, Bordignon SA. de L, Trevisan R, Boldrini II. Implications of poor taxonomy in conservation. J Nat Conser. 2017:36:10–13. https://doi.org/10.1016/j.jnc.2017.01.003

Vollmar A, Macklin JA, Ford L. Natural history specimen digitization: challenges and concerns. Biodivers Inf. 2010:7(2):93–112.

Weinstein SB, Kuris AM. Independent origins of parasitism in Animalia. Biol Lett. 2016:12(7):20160324. https://doi.org/10.1098/rsbl.2016.0324

Weissenböck H, Hubálek Z, Bakonyi T, Nowotny N. Zoonotic mosquito-borne flaviviruses: worldwide presence of agents with proven pathogenicity and potential candidates of future emerging diseases. Vet Microbiol. 2010:140(3-4):271–280. https://doi.org/10.1016/j.vetmic.2009.08.025

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. Darwin core: an evolving community-developed biodiversity data standard. PLoS One. 2012:7(1):e29715. https://doi.org/10.1371/journal.pone.0029715

World Spider Catalog. World Spider Catalog. Version 24.0. Natural History Museum Bern; 2023. http://wsc.nmbe.ch

WoRMS Editorial Board. World Register of Marine Species (WoRMS). https://www.marinespecies.org at VLIZ; 2023 [accessed 2023 Aug 28]. https://doi.org/10.14284/170

Zermoglio PF, Guralnick RP, Wieczorek JR. A standardized reference data set for vertebrate taxon name resolution. PLoS One. 2016:11(1):e0146894. https://doi.org/10.1371/journal.pone.0146894