BeeDC: An R package and globally synthesised and flagged bee occurrence dataset

Authors

James B. Dorey[1], Erica E. Fischer[2], Paige R. Chesshire[3], Angela Nava-Bolaños[4], Robert L. O'Reilly[1], Silas Bossert[5,6], Shannon M. Collins[7], Elinor M. Lichtenberg[7], Erika M. Tucker[8], Allan Smith-Pardo[9], Armando Falcon-Brindis[10], Diego A. Guevara[11], Bruno Ribeiro[12], Diego de Pedro[13], John Pickering[14], Keng-Lou James Hung[15], Katherine A. Parys[16], Lindsie M. McCabe[17], Matthew S. Rogan[18,19], Robert L. Minckley[20], Santiago J.E. Velazco[21], Terry Griswold[17], Tracy A. Zarrillo[22], Walter Jetz[18,19], Yanina V. Sica[18,19], Michael C. Orr[23,24†], Laura Melissa Guzman[25†], John A. Ascher[26†], Alice C. Hughes[27†], Neil S. Cobb[2,7†]

[†]*These authors jointly supervised this work*

Affiliations

1. College of Science and Engineering, Flinders University, Sturt Rd, Bedford Park, 5042, SA, Australia
2. Centre for the History of Science, Technology, and Medicine, Department of History, King's College London, Strand, WC2R 2LS, London, United Kingdom
3. Department of Biological Sciences, Northern Arizona University, S Beaver St, Flagstaff, 86011, AZ, USA
4. Unidad Multidisciplinaria de Docencia e Investigación, Facultad de Ciencias, Campus Juriquilla, Universidad Nacional Autónoma de México, Boulevard Juriquilla, Jurica La Mesa, Juriquilla, 76230, Querétaro, México
5. Department of Entomology, Washington State University, Dairy Rd, Pullman, 99164-6382, WA, USA
6. Department of Entomology, National Museum of Natural History, Smithsonian Institution, 10th and Constitution Avenue, Washington, 20560, DC, USA
7. Department of Biological Sciences and Advanced Environmental Research Institute, University of North Texas, W Mulberry St, Denton, 76201, TX, USA
8. Biodiversity Outreach Network, W Silver Spruce Ave, Flagstaff, 86001, AZ, USA
9. Animal Plant Health Inspection Service (APHIS); Plant Protection and Quarantine (PPQ); Science and Technology (S&T); Pest Identification Technology laboratory (PITL) United States Department of Agriculture (USDA), St. Suite, Sacramento, CA. 95814, USA
10. Department of Entomology, Research and Education Center, University of Kentucky, University Dr, Princeton, KY 42445
11. Departamento de Biología, Universidad Nacionalde Colombia, Bogotá, Cra 45 #268-5, D.C., Colombia
12. Programa de Pós-graduação em Ecologia e Evolução, Universidade Federal de Goiás, Goiânia, Av. Esperança, 74690-900, GO, Brazil
13. Ensenada Center for Scientific Research and Higher Education, Carr. Tijuana-Ensenada, Zona Playitas, 22860, Ensenada, Baja California, Mexico
14. Discover Life, Blue Heron Drive, Athens, GA 30605
15. Oklahoma Biological Survey, University of Oklahoma, Chesapeake St, Norman, 73019, OK, USA
16. USDA ARS Pollinator Health in Southern Crop Ecosystems Research Unit, Experiment Station Rd, Stoneville, 38776, MS, USA
17. USDA-ARS Pollinating Insects-Research Unit, Old Main Hill, Logan, 84322, UT, USA
18. Center for Biodiversity and Global Change, Yale University, Prospect St, New Haven, 06511, CT, USA
19. Department of Ecology & Evolutionary Biology, Yale University, Prospect St, New Haven, 06511, CT, USA

20. Department of Biology, University of Rochester, Rochester, 14620, NY, USA
21. Instituto de Biología Subtropical, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Misiones, Puerto Iguazú, Misiones, Argentina.
22. The Connecticut Agricultural Experiment Station, Huntington St, New Haven, 06511, CT, USA
23. Entomologie, Staatliches Museum für Naturkunde Stuttgart, Rosenstein, Stuttgart, 70191, Baden-Württemberg, Germany
24. Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beichen West Road, Beijing, 100101, China
25. Marine and Environmental Biology, Department of Biological Sciences, University of Southern California, Trousdale Pkwy, Los Angeles, 90089-0371, CA, USA
26. Department of Biological Sciences, National University of Singapore, Science Dr, 117558, Singapore
27. School of Biological Sciences, University of Hong Kong, Pok Fu Lam Rd, Lung Fu Shan, Hong Kong

corresponding author: James Dorey (jbdorey@me.com)

NOTE: This manuscript and the package behind it are still undergoing tests and development. Once these are complete and a final version is accepted we will update the input data, package versions, and re-run all queries (values will change). Please contact James for further queries of collaborations in the meantime.

Abstract

Species occurrence data are foundational for research, conservation, and science communication. But the limited availability and accessibility of reliable data represents a major obstacle, particularly for insects, which face mounting pressures. We present *BeeDC*, a new R package, and a global bee occurrence dataset to address this issue. We combined >17.7 million bee occurrence records from multiple public repositories (GBIF, SCAN, iDigBio, USGS, ALA) and smaller datasets, then standardised, flagged, deduplicated, and cleaned the data using the reproducible *BeeDC R*-workflow. Specifically, we harmonised species names following established global taxonomy, country, and collection date and we added record-level flags for a series of potential quality issues. These data are provided in two formats, "completely-cleaned" and "flagged-but-uncleaned". Our data cleaning process is open and documented for transparency and reproducibility. The BeeDC package and *R* Markdown are provided, and will be improved and updated regularly. By publishing reproducible R workflows and globally cleaned datasets we can increase the accessibility and reliability of downstream analyses. This workflow can be implemented for other taxa to support research and conservation.

Background & Summary

Occurrence datasets are increasingly critical for scientific research, conservation, and communication worldwide. From foundational systematic[1,2], life-history[3], and conservation studies[4,5], to continental or global macroecological[6,7] or macroevolutionary analyses[8], occurrence data are used to understand the natural world and form the basis of research, policy, and management. Spatiotemporal occurrence records (e.g., from specimens, images, or observations) are being delivered to, and produced by, community scientists in quantities that were unthinkable just a decade ago, with ever-increasing identification support by professional experts[9]. Online archives, such as the Global Biodiversity Information Facility (GBIF) and Symbiota Collection of Arthropod Network (SCAN), are invaluable data aggregators and tools in supporting the mobilisation of arthropod occurrence data to help both researchers and general audiences understand biodiversity. Concurrently and complementarily, initiatives like iNaturalist and QuestaGame (Australia) mobilise community and professional scientists to generate occurrence data for research, science communication, and more[9]. Occurrence data are being generated daily, making new analyses possible for previously data poor taxa. These data are key to developing the National Strategic Biodiversity and Action plans, upon which the monitoring framework of the Post-2020 Global Biodiversity Framework will depend[10]. Therefore, developing methods to aggregate and standardise such data is essential.

Despite the great utility of occurrence data, many issues can inhibit their successful fitness for use or mobilisation. Firstly, whilst "open data" is now mandated in many journals, data standards are implemented infrequently. This means that these data are increasingly fragmented and thus still inaccessible. Occurrence records can be rendered partially or critically incomplete at many different phases between the initial point of data collection and point of data publication and sharing[11]. For example, poor data collection practices at the point of production (in the field) can make downstream recovery efforts impossible. Data can often be omitted or altered during transcription, digitisation, or geo-referencing, making occurrences unusable or misleading. For example, only 2.7 out of 31 million (9%) Brazilian plant occurrences were considered high-quality following data cleaning and validation by Ribeiro, et al. [12]. Secondly, data can be of excellent quality, digitised, and used in publications, but they may not be made available in accessible repositories. Whatever the reason, such data are effectively unavailable for — or in the case of misleading data detrimental to — further research, conservation, and management[13]. Funding institutions to curate and digitise collections might fill in some of the major data gaps revealed in recent global analyses of public data[6].

Mobilising massive datasets with numerous potential issues represents a major roadblock for researchers and other users. A careful balance between discarding too much data and accepting too many problematic records must be maintained. Ideally, problematic records can instead be flagged (allowing users the choice of treatment) or fixed using a combination of automated pipelines and expert validation[14]. The barriers are such that the data cleaning process is often built from scratch or nearly so for each new project (re-inventing the wheel)[6,7]. Constantly rebuilding the data cleaning process can result in haphazard application of cleaning and data quality and is a major hurdle for many researchers and conservation practitioners that might preclude them from undertaking robust and reproducible analyses. This represents a major accessibility problem that stalls research, management, and conservation — particularly for research groups that lack sufficient support, expertise, facilities, or time to develop or implement cleaning workflows — that could otherwise produce excellent research. The result is major gaps in our knowledge of species distributions and ecological niches, especially in developing economies and over extended timescales[15]. Until now, very few major efforts have been made to combine, clean/flag, and make accessible the world's occurrence datasets; this is particularly true for diverse and ecologically important arthropods[7,16]. We tackle this issue of data cleaning and reproducibility using a major group that is taxonomically extensive (>20,000 valid species), ecologically and economically important, and for which research is booming globally: bees.

Bees (Hymenoptera: Anthophila) are a keystone taxon of terrestrial ecosystems across the globe and the most significant group of pollinating animals in both agricultural and natural settings[17,18]. As such, great efforts are made to understand their spatial distributions[6]. It is critical for research on this flagship invertebrate taxon to be robust and correctly analysed; if studies are found to be unreliable, it can reduce public and policy enthusiasm for funding the necessary research and conservation of bees as well as less-studied insect groups. Despite the urgent need for quality occurrence data, there are many barriers to their use, particularly accessibility and reliability. Removing these hurdles will encourage new research and enable best practices at greatly reduced effort and financial cost. At the time of download and to our knowledge, there are ~17.6 million bee occurrence records that are publicly available online and more that are private or inaccessible. Here we collate and process several bee datasets for open-use by all interested parties. We also produce a machine-readable version of the global bee taxonomy and country-level checklist available on the Discover Life website[19]. Specifically, we curated multiple raw datasets, combined curated records, taxonomically harmonized species names and annotated potentially problematic records. Herein, we provide (i) a completely-cleaned and standardized dataset (4.7 million occurrences), (ii) a flagged-but-uncleaned dataset (for users to filter based on our flagging columns; 17.7 million occurrences), (iii) all R-scripts and input files needed to produce both datasets (Fig. 1), and (iv) summary figures and tables highlighting the state of global bee data. All data and scripts are accessible, documented, and open access. Importantly, we plan to periodically update and publish new versions of the BeeDC package and bee data. This will leave room for improved versions and functionality, especially because new occurrence data are consistently being added to online databases. Despite being developed specifically for bee datasets, many of our functions are generic and can be applied to various taxa for which similar issues exist.

Methods

Data sources

We sourced data from publicly available online data repositories (Supplementary Table 1) as well as select non-public, private, or otherwise publicly inaccessible sources that were willing to share their data for this study. We use the term "data source" in the manuscript, but

recognise that these sources can be very distinct in function. For example, aggregators serve data that is created and curated by multiple distinct data providers. The primary repositories that we sourced data from were (i) the Global Biodiversity Information Facility (GBIF), (ii) Symbiota Collections of Arthropods Network (SCAN), (iii) Integrated Digitized Biocollections (iDigBio), (iv) the United States Geological Survey (USGS), and (v) the Atlas of Living Australia (ALA). Additionally, we have sought out and incorporated smaller public or private data sources from (i) Allan Smith-Pardo (ASP), (ii) Robert Minckley (BMin), (iii) Elle Pollination Ecology Lab (EPEL[20]), (iv) *Bombus* Montana (Bmont; Casey Delphia[21]), (v) Ecdysis[22] (Ecd), (vi) Gaiarsa, et al. [23] (Gai), (vii) the Connecticut Agricultural Experiment Station (CAES), (viii), USDA ARS Southeastern USA (Parys), (ix) Eastern Colorado (Arathi Seshadri), (xi) Florida State Collection of Arthropods (FSCA), (xi) Armando Falcon-Brindis (Arm), and (xii) publicly available bee literature (SMC, Bal, Lic, Dor)[3,4,24-30].

Global Biodiversity Information Facility data were downloaded via the online portal on the 31st of August 2022[31-37] on a per-family basis. SCAN data were downloaded on the 31st of August 2022[38-47] on a per-family basis and by subsets of collections if the one million record limit was surpassed (the maximum capacity for a single download). Integrated Digitized Biocollections data were downloaded on the 31st of August 2022[48-54] by selecting all Apoidea records and filtering to the seven bee families (Andrenidae, Apidae, Colletidae, Halictidae, Megachilidae, Melittidae, and Stenotritidae). United States Geological Survey data were provided directly by Sam Droege on the 19th of November 2022. Finally, ALA data were downloaded on the 5th of October 2022[55] directly via their *R*-package, *galah* version 1.4.0[56].

We sourced taxonomic name and checklist data from information hosted on the Discover Life website[19]. Our bee taxonomy is current as of the 11th of November 2022 and the bee country checklist as of the 16th of January 2023. Both of these datasets are kept updated by Ascher and Pickering [19] and are critical sources to ensure occurrence quality.

Data-cleaning workflow

Analyses were undertaken in *R* version 4.2.1[57] and *R-Studio* "Spotted Wakerobin" Release. The script relied heavily on several *R*-packages. Much of the workflow script used and built upon the *bdc* (biodiversity data cleaner) version 1.1.1 package[12], tidyverse packages — particularly *dplyr* v. 1.0.9[58], *magrittr* v. 2.0.3[59], *tibble* v. 3.1.8[60], *stringr* v. 1.4.0[61], *tidyselect* v. 1.1.2[62], *readxl* v. 1.4.0[63], *ggplot2* v. 3.3.6[64], *forcats* v. 0.5.1[65], *tidyr* v. 1.2.0[66], *rlang* v. 1.0.4[67], *xml2* v. 1.3.3[68], *readr* v. 2.1.2[69], and *lubridate* v. 1.8.0[70], — and *CoordinateCleaner* v. 2.0–20[71]. Additionally, we used *R.utils* v. 2.12.0[72], *rlist* v. 0.4.6.2[73], *galah* v. 1.4.0[56], *EML* v. 2.0.6.1[74], *emld* v. 0.5.1[75], *rnaturalearth* v. 0.1.0[76], *rnaturalearthdata* v. 0.1.0[77], *countrycode* v. 1.4.0[78], *rangeBuilder* v. 1.6[79], *rworldmap* v. 1.3–6[80], *hexbin* v. 1.28.2[81], *cowplot* v. 1.1.1[82], *ggspatial* v. 1.1.6[83], *renv* v. 0.15.5[84], *chorddiag* v. 0.1.3[85], *igraph* v. 1.3.5[86], *sf* v. 1.0–7[87], and *terra* v. 1.5–21[88]. Our script was both built and run on a 2019 MacBook Pro with a 2.4 GHz 8-core Intel i9 with 64GB 2667 MHz DDR4 RAM. Hence, while it can be run on personal computers — and this accessibility is intentional — we are aware that memory (RAM) and physical storage could inhibit some devices from processing the full dataset (i.e., 17.7 million occurrences and ~120 columns) .

The core workflow, R Markdown, and functions are freely available on GitHub (*available in the supplementary SubmissionGit folder prior to acceptance;* Link will be posted upon acceptance). The workflow is clearly numbered and labelled using the *R-Studio* document outline to allow quick navigation of the script. For clarity and continuity, we list the sections below according to our script, including (0.x) script preparation, (1.x) data merge, (2.x) data preparation, (3.x) initial flags, (4.x) taxonomy, (5.x) space, (6.x) time, (7.x) de-duplication, (8.x) data filtering, and (9.x) summary figures. Additionally, most of our functions attempt to provide extensive and

informative user-outputs for quality assurance. Our script focuses on the specifics of bee data; however, it provides the templates needed to integrate the specifics of other taxon groups and input data.

Most functions in our script aim to identify and flag potentially problematic occurrence records based on specific tests and user provided thresholds (flagging functions). However, several functions modify the occurrence records themselves when errors are identified (carpentry functions). We also provide summary and filtering functions that allow users to explore data issues and export useable datasets. While most functions can be implemented anywhere in the workflow, a small subset relies on columns produced from earlier functions. Below, we explain the sections and functions found in our script and the logic behind their implementation. We do not explain the optional steps documented in the script as they were not implemented in our dataset.

0. Script preparation

To ensure accessibility and ease of use, we ask users to set the root file path in which the script should look for or create certain datasets or functions. Our custom *dirMaker* function should locate and, if missing, create the file structure sought by the rest of the script and the *bdc* package. Packages are then installed and loaded. Here the script initialises and stores the *renv* files. The *renv* package files keep information about package versions to encourage further reproducibility of scripts and ease of publishing for users (see *renv* Supplementary Files).

1. Data merge

The data merge section of the script reads very large datasets from the major data repositories — GBIF, iDigBio, USGS, SCAN, and ALA. It then (i) unifies the data (selecting certain columns set by the custom *ColTypeR* function) to DarwinCore format[89], (ii) merges them into a single dataset, and (iii) creates a metadata file that both accompanies the R-object and is saved locally (Supplementary Table 1). These functions were built to accommodate the entirety of the large repositories above; however, they can be slow on computers with limited RAM.

2. Data preparation

Here, we provide users with two options to re-import the data saved during *1. Data merge*. For general applications, the *bdc* package provides excellent import functionality (2.1a in script). For users who want to run data from the above major repositories, we provide seamless functionality to read those data (2.1b in script). For users aiming to update the bee datasets from online repositories, they may incorporate the documented manual data edits (2.2 in script) and re-incorporate several privately curated datasets (2.3–2.4 in script). Once a dataset is standardised to DarwinCore format and contains the necessary columns, users can begin flagging or cleaning each occurrence record for data integrity. Our total dataset comprised 17,732,164 uncleaned bee occurrence records.

3. Initial flags

To perform initial data-quality flags and checks, we used a combination of custom and *bdc* functions, which we will list below. These include data flagging and carpentry functions, with the latter supporting the former.

3.1 Scientific name flags. We used the *bdc* function, *bdc_scientificName_empty,* to flag records with no scientific name provided (i.e., an empty *scientificName* cell). This step flagged 162,155 (1%) occurrences.

3.2 Missing Coordinates. We used the *bdc* function, *bdc_coordinates_empty,* to flag records with no coordinates provided (empty *decimalLatitude* or *decimalLongitude* columns). This step flagged 3,070,932 (17%) occurrences.

3.3 Coordinates out of range. We used the *bdc* function, *bdc_coordinates_outOfRange,* to flag records that were not on the map (i.e., not between -90 and 90 for latitude or not between -180 and 180 for longitude). This step flagged 138 (<1%) occurrences.

3.4 Poor record source. We used the *bdc* function, *bdc_basisOfRecords_notStandard,* to flag occurrences that did not meet our criteria regarding their basis of records. Broadly, we kept all event, human observation, living specimen, material/preserved specimen, occurrence, and literature data. For this step, we were relatively liberal in what we allowed to remain in the dataset and only excluded fossil specimens; however, some users might prefer to also remove human observations. This step flagged 28,555 (<1%) occurrences.

3.5 Country name. We provide a custom function, *countryNameCleanR*, to match GBIF ISO2 country codes to country names from a static Wikipedia ISO2 to country name table within the function. Users can also input a data frame of problem names and fixed names to replace them in the dataset. We then wrapped the *bdc* function *bdc_country_from_coordinates* into a chunking function (to work on smaller portions of the dataset at a time) where the user can specify chunk-sizes to best manage the RAM-intensive *bdc* function. This function assigns country names to those occurrences that were missing them but that have valid coordinates that correspond to a country. This step assigned 285,806 (2%) country names.

3.6 Standardise country names. We used the *bdc* function, *bdc_country_standardized,* to attempt to further standardise country names for consistency. This step standardised 9,684,242 (55%) country names.

3.7 Transposed coordinates. We used a modified, wrapped, and chunked (as above) *bdc* function, *bdc_coordinates_transposed* that flags and corrects coordinates for which the latitude and longitude are transposed. This step corrected 2,644 (<1%) transposed coordinates.

3.8 Coordinates and country inconsistent. Here, we created a similar, but custom, function to *bdc*'s called *jbd_coordCountryInconsistent*. We re-built the *bdc* version for memory efficiency and to accommodate the processing of our dataset, which is much larger than the *bdc* test dataset. Our function flags occurrences for which the coordinates and country do not match while allowing for a user input map buffer (in decimal degrees). This step flagged 306,362 (2%) occurrences.

3.9 Georeference issue. We ran the *bdc* function, *bdc_coordinates_from_locality,* which highlights occurrences for which there is no latitude or longitude directly associated with the data, but for which sufficient locality data may be used to extrapolate relatively accurate geological coordinates. This step flagged 2,564,391 (14%) occurrences.

3.10 Absent records. We provide a custom function to flag occurrence records that are marked as "ABSENT" to remove absence data. Users may not be aware that many thousands of records are often provided as absent. These records might occur, for example, due to repeated sampling programmes not finding a certain species on a particular date or at a particular site. This step flagged 1,091,462 (6%) occurrences.

3.11 Restricted licenses. We provide a custom function that aims to flag records that are not licensed for use. While such records should not be on public data repositories, some few are. Therefore, users should be aware not to use these protected data points. This step flagged 514 (<1%) occurrences.

3.12 GBIF issues. We provide a custom function to flag occurrences already flagged for user-specified GBIF issues. For example, we flagged GBIF occurrences marked as "COORDINATE_INVALID" and "ZERO_COORDINATE". This step flagged 137 (<1%) occurrences.

At the end of the initial flags section, we provide a function to save the flags for all occurrences as a separate file. Users may then use *bdc* functions to create summaries, reports, and some figures.

4. Taxonomic cleaning

While *bdc* provides great functionality for taxonomic cleaning of some groups, it does not provide access to the most-current bee taxonomy. Hence, we used the online global bee taxonomy available through Discover Life[19], which lists approximately 21,000 valid names and 31,000 synonyms or orthographic variants. Discover Life is the leading source of bee names and is peer-reviewed by global experts on an ongoing basis. We flagged this list for ambiguous names that, to varying degrees, could not be matched due to their complicated taxonomic histories. Records with ambiguous names were either processed differently or excluded from the below function and hence the final dataset. We also made some manual corrections as taxonomic issues (mostly orthographic variants) were identified (Supplementary Table 2). The final taxonomy is available in the supplementary data (for the R workflow) and as Supplementary Table 3.

We first cleaned our species names using the *bdc* function *bdc_clean_names,* which attempts to clean names and unify writing-styles (e.g., capitalization and punctuation). It also removes family names, qualifiers (this is flagged), infraspecific terms, separates authors, dates, and annotations. This function also adds a taxonomic uncertainty flag, *.uncer_terms*, which flagged 102,940 (<1%) records. We then used our custom function, *HarmoniseR,* to harmonise occurrence names with our combined Discover Life taxonomy. This function completes steps to (i) compare occurrence names directly with valid names (genus species authority), (ii) combine the *bdc*-cleaned name and occurrence authority to match against the valid name, (iii) match species names with canonical flagged names (i.e., scientific names with flags from Discover Life), (iv) directly compare scientific names, (v) combine the occurrence scientific name and occurrence authority to match against the valid name, (vi) match occurrence's scientific name with the taxonomy's valid name with subgenus removed from both, and (vii) as above but using canonical. Between each step, the function filters out already matched names. These steps are first taken for the non-ambiguous names (according to our taxonomy) and then applied again for the ambiguous names, but also matching the authority ("author year", all lowercase and with no punctuation). The function then moves the provided name to the *verbatimScientificName* column and updates the *scientificName, species, family, subfamily, genus, subgenus, specificEpithet, infraspecificEpithet*, and *scientificNameAuthorship* columns where better matches are found. Occurrence records that

do not match the taxonomies are flagged accordingly. The data produced from each step are merged and an updated object is output. This step matched valid names to 15,798,589 (89%) and flagged 1,971,443 (11%) occurrences, respectively.

5. Space flagging

We further flagged our data for spatial issues using a combination of custom functions, *bdc*, and *CoordinateCleaner*. We outline these steps below.

5.1 Coordinate precision. We used a custom function, *jbd_coordinates_precision*, to flag occurrence records with latitude and longitudes below a threshold of two decimal places (~1.1 km at the equator). This step differed from the bdc function, *bdc_coordinates_precision*, by only flagging occurrences where both latitude *and* longitude were rounded. This step flagged 3,661,235 (21%) occurrences.

5.2 Common spatial issues. We then used the *CoordinateCleaner* function, *clean_coordinates,* which runs several tests to flag potentially erroneous data. We flagged records that were within (i) 1 km of capital cities, (ii) 500 m of province or country centroids, (iii) 1 km of GBIF headquarters in Copenhagen, Denmark, or (iv) 100 m of biodiversity institutions; or that have (v) equal latitude and longitude coordinates, or (vi) zero as latitude and longitude. For example, these issues can arise when occurrences were labelled only with the city, province, country, institution, or repository are georeferenced to those exact locations. The latter two could arise from copy errors or incomplete data. We did not flag points in the ocean as they are flagged with a small buffer by *5.5 Country Checklist* (below). This step flagged (i) 14,257; (ii) 16,451; (iii) 7; (iv) 64,047; (v) 10,640; and (vi) 10,467 (each <1%) occurrences, respective to the above criteria.

5.3 Fill-down errors and gridded datasets. Firstly, we provide a custom function, *diagonAlley*, that uses a sliding window to attempt and flag potential fill-down errors in the latitude and longitude columns. The function removes any empty values and then groups data by event date and collector. It then arranges latitude and longitude, removes identical values, and flags sequences of latitudes or longitudes where the differences between records are exactly equal. The user defines a minimum number of repeats required for a flag (we used four for our datasets) and a minimum number of decimal places required to consider an occurrence (using 5.1 jbd_coordinates_precision; we used three). Secondly, we then implemented the *cd_round* function from *CoordinateCleaner* to identify datasets (using the *datasetName* column) that have their latitudes or longitudes potentially gridded, using the default values. This step flagged 208 and 113,680 (each <1%) occurrences for fill-down and gridded datasets, respectively.

5.4 Coordinate uncertainty. We provide a custom function, *coordUncerFlagR*, that flags records by a user-defined threshold for coordinate uncertainty — usually provided in the Darwin Core *coordinateUncertaintyInMeters* column. This step flagged 2,431,777 (14%) occurrences.

5.5 Country Checklist. We provide a custom function, *countryOutlieRs*, that uses the country-level checklist available on the Discover Life website (Supplementary Table 4). The function checks all of the harmonised species names against the checklist and the country in which the occurrence falls (using overlap with *rnaturalearth*). Points that don't align with *rnaturalearth* (e.g., they are on the coast) can be buffered by a user-specified amount, in degrees, to attempt a match. In our case, we used 0.05 degrees (~5.6 km). The function produces three columns. The first column, countryMatch, summarises the occurrence-level

result: where the species is not known to occur in that country (*noMatch*), it is known from a bordering country (*neighbour*), or it is known to occur in that country (*exact*). The second column's output depends on user input. If the user wants to keep occurrences that are either exact matches or in adjacent (bordering) countries to those in the checklist (*keepAdjacentCountry* = TRUE) then the filtering column *.countryOutlier* will be TRUE for these cases and FALSE only for those with *noMatch*. A *keepAdjacentCountry* = FALSE argument will only flag exact country checklist matches as TRUE. For our dataset, we chose the former. The third column, .sea, flags the points which don't align with *rnaturalearth* or its buffer — they are identified as being in the ocean. The *countryOutlier* column flagged 1,542,506 (9%) occurrences and the .sea column flagged 473,027 (3%).

Users may then produce maps, reports, and figures using the *bdc* package. They can also append the saved flag file using the custom *flagRecorder* function.

6. Time flagging

The next major sequence of functions aims to recover or filter occurrence records with date-related issues. We outline the steps taken below.

6.1 Recover missing event date. We provide the custom function, *dateFindR*, that seeks to recover occurrence records that would otherwise be removed due to missing event date. The function first removes unreasonable dates based on a user-defined year-range (we removed dates prior to 1700 and after the present year). The function then seeks to harmonize date formats in a step-wise manner by first (i) combining dates from the year, month, and day columns into the *eventDate* column, then (ii) taking just the year column where it is provided without event date. Subsequently, the function looks for dates in the *verbatimEventDate*, *fieldNotes*, and *locationRemarks* columns by looking in sequence for unambiguous date strings in the following formats: (iii) year-month-day, (iv) day-month-year (where day is identifiable — i.e., days >12; or months are identifiable — in written or Roman-numeral formats), (v) month-day-year (where days or months are identifiable as above), and (vi) month-year (where month is identifiable as written or Roman-numeral formats). The function then finds ambiguous date strings in the *verbatimEventDate*, *fieldNotes*, and *locationRemarks* columns and keeps only the year value. All date values are then returned in the standardised year-month-day-hours-minutes-seconds format. This step rescued dates for 1,084,628 (6%) occurrences.

6.2 Missing event date. We then used the *bdc* function, *bdc_eventDate_empty*, to flag records that do not have date in the *eventDate* column. This step flagged 1,856,354 (10%) occurrences.

6.3 Year from event date. We included the *bdc_year_from_eventDate* function from *bdc* to find year from the *eventDate* column when it is missing. This step extracted four-digit year from 15,875,815 (90%) occurrences.

6.4 Old records. We used the *bdc* function, *bdc_year_outOfRange*, to flag occurrence records that are likely too old for use in species distribution modelling. Here, we flagged occurrences from before 1950. Users may want to consider changing this value or disregarding this flag as it applies, or does not, to the question(s) that they want to ask. This step flagged 1,622,687 (9%) occurrences.

Users may then create reports and figures on the time filters using the *bdc* package. They can also append the saved flag file using the custom *flagRecorder* function.

445 ***7. Duplicate records***

446 Duplicate records frequently arise between or within repositories. These records can be
447 difficult to discern particularly where single specimens have been assigned multiple, or worse,
448 no unique identifiers, or where the same locality has been georeferenced independently by
449 multiple institutions. We provide a custom function called *dupeSummary*, that iteratively
450 searches occurrences for duplication using multiple column-sets. Users can choose to identify
451 duplicates based on identifier columns, collection information, or both. They may also define
452 any number of custom column sets by which to identify duplicates.

453 Some identifier columns might contain codes that are too simple. The function allows users to
454 set thresholds to ignore those occurrences when checking the relevant columns for duplicates.
455 For example, the *catalogNumber* might be "145" or "174a", which could result in over-
456 matching of duplicates. Users may set a *characterThreshold* and *numberThreshold* which
457 would ignore codes which don't pass both. Users may also set a *numberOnlyThreshold*, which
458 will check codes above that threshold, irrespective of the *characterThreshold*. We use the
459 defaults of two, three, and five, respectively. Hence, minimum passing codes could include
460 "AG194" and "390174". This can be entirely turned off by setting all values to zero or ignored
461 for selected column sets using *CustomComparisonsRAW*.

462 The function identifies duplicates based on collection information where it iteratively
463 compares user-defined sets of columns. The function first compares custom column sets and
464 can then compare generic, but customisable, column sets. For our bee data, we used several
465 steps to identify duplicates. Firstly, we compared (i) catalogNumber and *institutionCode* using
466 *CustomComparisonsRAW*. Secondly, we identified duplicates based on the *scientificName* with
467 the (ii) *catalogNumber* and *institutionCode*, (iii) *gbifID*, (iv) *occurrenceID*, (v) *recordId*, and (vi)
468 *id* columns using CustomComparisons. Finally, we identified duplicates using the generic
469 column sets of *decimalLatitude*, *decimalLongitude*, *scientificName*, *eventDate*, and *recordedBy*
470 columns at the same time as the (vii) *catalogNumber* and (viii) *otherCatalogNumbers* columns.

471 Using the *scientificName* column with the identifier columns allows different species with the
472 same identifier to be maintained. This is important where an event identifier (multiple
473 specimens from one collection event) is placed in the wrong column. For the occurrences
474 where the taxonomy did not match (e.g., because of an ambiguous or incomplete
475 identification) duplicates won't be identified; in these cases they will be removed by
476 *HarmoniseR* (or another flag). At the same time, occurrences where species identifications
477 have been updated in only some datasets will not be identified as duplicates. Users may
478 choose their own input parameters while weighing the costs and benefits.

479 The function arranges the data by (i) a user-defined list of input sources (where the first data
480 sources are preferred over later ones; i.e., GBIF > SCAN > iDigBio, etc.), then, (ii) completeness
481 by user-defined columns, and (iii) by the summary column (to keep clean occurrences). For
482 point three above, duplicate occurrences with more of these completeness columns will be
483 preferred over those with fewer; i.e., the most-complete record is chosen. Where public and
484 private data were duplicated, we gave preference to private data providers over the public
485 data aggregators under the assumption that data providers have the most recent information.
486 We also preferred manually cleaned occurrence records from Chesshire, et al. [90] over those
487 sourced directly from data aggregators. All pairwise duplicates were clustered where they
488 overlapped and a single best occurrence was kept using the above arrangement. This step
489 flagged 7,425,881 (42%) occurrences.

490 ***8. Final filter***

We first removed 181 occurrences that experts manually identified as misidentifications or outliers. We then provide the code for users to produce a cleaned dataset to their own specifications. Users may download our flagged dataset and begin at this step to choose the flags that are important for them to remove. They can also re-run specific functions on an already flagged dataset using custom specifications. For our cleaned dataset, we removed several filtering columns that we decided were not critical to data integrity. These were the grid flags (*.gridSummary*, as well as *.lonFlag* and *.latFlag*) and the taxonomic uncertainty flag (*.uncer_terms*). We also did not filter for high coordinate uncertainty (*.uncertaintyThreshold*), because this filtering level might be too specific to a user's intended analysis and can remove ~1 million otherwise clean records at the 1 km level. We then selected all flag columns (those that begin with a full-stop), re-produced the *.summary* column using our custom *summaryFun* function to include any records that are flagged in at least one flagging column, and then filtered to keep only unflagged records. This step removed 11,920,090 (67%) occurrences and left 5,811,564 (33%) cleaned occurrences. Most of these duplicates arose between data sources; however, within-source duplicates were quite prominent in SCAN (Fig. 2). Finally, we removed all filtering columns, leaving only the data columns (Supplementary File 1). We also provide the unfiltered dataset with all filtering columns included so that users may selectively implement our filtering steps (Supplementary File 2).

9. Summary figures, tables, and outputs

Beyond the figures produced throughout the cleaning process by the *bdc* package, we also provide several unique custom figure functions.

9.1 Duplicate chordDiagrams. We provide a function, *chordDiagramR*, that wraps the circlize v. 0.4.15[91], ComplexHeatmap v. 2.12.1[92,93], and paletteer v. 1.3.0[94] packages to build a chord diagram that visualises the linkages between duplicated occurrence data sources (Fig. 2).

9.2 Duplicate histogram. We provide a function, *dupePlotR*, to visualise duplicates by source, breaking them down into (i) discarded duplicates, (ii) kept duplicates (as chosen in *7.x Duplicate records*), and (iii) unique records (Fig. 3). This is displayed as both a total number of records and the proportion within each data source.

9.3 Flags by source. We provide another function, *plotFlagSummary*, that produces a compound bar plot that, for each data source, indicates the proportion of records that pass or fail, or cannot be assessed for each flag (Fig. 4). We built additional functionality into this function that allows users to provide a species name (and the column in which that name is found) and (i) save the occurrence data, (ii) produce a map coloured by a filtering column of choice, and (iii) the compound bar plot for that specific species. This can be an excellent function to quickly examine potential issues (see SubmissionGit > BasicWorkflow.R in Supplementary Materials). Additionally, users can choose to output the table used to make the figure using saveTable = TRUE (Supplementary Table 5).

9.4 Maps. We provide a function, *summaryMaps*, that uses the cleaned data to show the number of species and number of occurrences per country (Fig. 5). For the latter function, we broke up these values using classes and "fisher" intervals[95]; however, any style from the classInt v. 0.4–8[96] package can be entered. Additionally, we provide a function, *interactiveMapR*, that iteratively makes and saves interactive html maps for any number of species. These interactive maps colour occurrences if they pass or fail any flags, if they are country outliers, or expert outliers. Collection and flag information are provided in pop-ups for each point.

9.5 Data providers. We provide a function, *dataProvTables*, that produces a table of data providers with the number of cleaned occurrences and species for each. The function also attempts to identify the data provider using other columns where it is omitted from the *institutionCode* column.

9.6 Flag summary. We built a function, *flagSummary*, that will produce a summary table of the total number of failed occurrences per flag and per species (or any other categorical column of interest).

9.7 Taxonomic and country checklist queries. Users may be interested in querying BeeDC for species names, validities, and in which countries they can be found. Our function, *BeeDCQuery*, allows users to enter one or more species name as a character vector and it will report on (i) the validity of the name, (ii) all synonyms, and (iii) the countries in which it is found. This information will be returned as a list of tables.

10. Package data

10.1 Bee taxonomy and checklist. We provide a modified version of the global bee taxonomy and country-level checkliss produced by John Ascher and John Pickering and hosted on Discover Life[19] as part of BeeDC. These datasets are names *beesTaxonomy* and *beesChecklist*, respectively.

10.2 Test datasets. We further provide three test datasets with BeeDC. The dataset *bees3sp* includes 105 flagged points from three haphazardly chosen species; *beesFlagged* contains 100 randomly chosen flagged occurrence records; and *beesRaw* contains 100 randomly chosen unflagged occurrence records.

Data Records

Our dataset and the data used in our analyses are available for download in our repository (Link will be posted upon acceptance). Occurrence datasets are provided in standard DarwinCore format and saved as .csv files. The Supplementary Files are organised into the following folders:

1. Figures. Contains primary figures and additional *bdc* figures.
2. OutputData. Contains the "completely-cleaned" (SupplementaryFile1_05_cleaned_database.csv) and "flagged-but-uncleaned" (SupplementaryFile205_unCleaned_database.csv) datasets, reports, and the R console outputs from the script (SupplementaryFile3_CleaningRun_10Dec22.txt).
3. InputData. Contains the major repository downloads, additional input datasets, custom Chesshire, et al. [90] data files, and data from the DiscoverLife website (taxonomy and checklist).
4. SubmissionGit. Contains the files and functions available on the GitHub at the time of submission.
5. The *Renv* folder and lockfile.
6. SupplementaryTables. Contains all tables mentioned above (Supplementary Tables 1–5) and a small example "flagged-but-uncleaned" dataset with every 1000th row retained (Supplementary Table 6).

Technical Validation

For our current data version, we outline the data quality according to our flags. We provide summaries of the deduplication process, including the duplication linkages within and between datasets (Fig. 2), and the number and proportion of duplicates, kept duplicates, and unique occurrences per dataset (Fig. 3). Importantly, we provide a figure summary of the

proportion of records flagged for all data sources (Fig. 4). This figure indicates data quality across the entire dataset and where each data source might focus their cleaning efforts. We demonstrate the country-level patterns of both species and occurrences on a global map (Fig. 5). Finally, all *bdc* summary and data quality figures and maps are provided in the Supplementary Figures folder (Link will be posted upon acceptance).

We found that data quality was highly variable between sources (Fig. 4). Data duplication was often the most prominent flag, comprising 42% of the total uncleaned dataset. There was substantial duplication between and even within both major and minor data repositories (Figs. 2 and 3). Importantly, the assumption that "all data make their way to GBIF" is incorrect (at least at the time of downloading); rather, each repository holds unique data (Fig. 3). It is also possible that poor data quality might preclude successful duplicate matching. Regardless, this highlights the importance of sourcing occurrence data from a variety of repositories. However, in terms of actual input data quality and completeness, many occurrences (i) lacked coordinates, (ii) did not have a scientific name or that name did not match known taxonomy, (iii) had low-resolution coordinates, (iv) had high coordinate uncertainty (where this was estimated), (v) did not match the country checklist on the Discover Life website, (vi) lacked collection date, and/or (vii) were collected prior to 1950 (Fig. 4). These quality issues might preclude data use in downstream works and require further efforts to repair. In worst case scenarios, unfit data might pass flagging steps and cause misleading results. The presented workflow enables researchers to prepare data that can support robust analyses. To the best of our knowledge, this is now the most thoroughly curated global occurrence dataset for bees, and is potentially leading for any terrestrial invertebrate group of this size.

The mismatch between bee species richness according to occurrence data and country checklists has already been highlighted by Orr, et al. [6]. We also highlight the geographical, taxonomic, and collection biases in the global bee data (Fig. 5). We show that the number of cleaned species occurrences in most of Africa and large parts of Asia are in the lowest class counts, where classes are defined using a Fisher-Jenks algorithm[95] (Fig. 5a). We also show that the number of occurrences in Africa, Asia, and even South America are often in the lowest class counts and to a greater degree than species counts (Fig. 5b). Some countries in Africa (Guinea, Equatorial Guinea, Djibouti, and parts of Western Sahara) do not have a single cleaned occurrence record (Fig. 5). There are also large mismatches between the number of species and number of occurrence records. These disparities are hugely concerning and highlight ongoing inequalities in both taxonomic and sampling efforts. We hope that this contribution provides a foundation for a more formal recognition and prioritization of these important efforts.

Usage Notes

All of our occurrence datasets are cleaned for (i) taxonomic information according to the Discover Life website, (ii) country name and ISO2 code, and (iii) recoverable event dates from other columns. When using our data or parts of our workflow users should also cite the data sources where they are relevant.. Our occurrence data are provided in two ways:

Completely-cleaned. We provide a dataset that is completely cleaned of records that failed any of our filtering steps with the exception of: (i) *.gridSummary,* (ii) *.lonFlag*, (iii) *.latFlag*, and (iv) *.uncertaintyThreshold*. All filtering columns are also removed from this dataset to reduce file size.

Flagged-but-uncleaned. We provide a dataset that includes all of the above filters as flags, in addition to the otherwise unfiltered flags above. This dataset is provided with a column for each of these flags where occurrences that failed for a filtering step are flagged as

"FALSE" and those that passed as "TRUE". This convention is used to maintain continuity with functions from the *bdc* package. Users are able to use our script to read in and filter their data or manually filter these data in a way that is appropriate for their use. Our filtering script should also provide users with the necessary template to reduce or modify their data into their desired size and structure. In particular, users may use the *dontFilterThese* command in the *summaryFun* function to exclude certain filters that do not relate to their research question (see SubmissionGit > BasicWorkflow.R in Supplementary Materials).

Users may incorporate our data pulls or conduct their own. However, we do note that ALA, through the *galah* package, is the only major data provider with a convenient download protocol for large datasets. Other Application Programming Interfaces (APIs) and even web protocols make downloading a limiting and lengthy process; for example, the *rgbif* package has a limit of 100,000 occurrences. Making similar API methods available for other major data repositories would greatly increase their accessibility. Usage of our workflow should be accessible to most users with basic R knowledge. Once the datasets and folder structure are determined the script should require minimal user input. We highlight again that this workflow was tested on a machine with 64 GB of RAM and that users with less RAM might have prolonged run times and need to reduce chunk sizes for some functions. In the end, this work is aimed to facilitate opensource data for both scientific and applied usage.

Code Availability

We provide our *R* code in several formats. Firstly, we provide a full R Markdown file that demonstrates the entire functionality of our script with the input data (*available in the supplementary SubmissionGit folder prior to acceptance;* Link will be posted upon acceptance). Secondly, we provide this code as an annotated *R*-script to run from start to finish (Link will be posted upon acceptance). Finally, we provide a "basic workflow" *R*-script for users that simply wish to download our flagged but unfiltered dataset and apply (i) manual filters based on our filtering columns or further filter to only include certain (ii) countries, (iii) date ranges, or (iv) uncertainty levels (Link will be posted upon acceptance). This latter script can also be added to the filtering steps from the prior scripts to re-run or modify the arguments of existing flag columns. Our scripts and related files can be found on our GitHub page (In Supplementary Data for submission).

Acknowledgements

We acknowledge the wonderful data providers such as ALA, GBIF, SCAN, and iDigBio, who maintain and make available datasets for public use. We also sincerely thank Sam Droege for organising and contributing the substantial USGS bee dataset. We additionally acknowledge and thank individuals and organisations that do the same, including those individually used here produced by: Elizabeth Elle, Casey Delphia, Ecdysis, Gaiarsa et al., Arathi Seshadri, and the Florida State Collection of Arthropods. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer. JB Dorey was partly funded by the Biodiversity Outreach Network. MS Rogan, YV Sica, and W Jetz were partly funded by Burt's Bees. A Nava-Bolaños is grateful to Consejo Nacional de Ciencia y Tecnología for the postdoctoral fellow for the project "Polinizadores: actores clave en la seguridad alimentaria". EM Lichtenberg and SM Collins were partially funded through the Texas State Wildlife Grants program grant CA-0002506 in cooperation with the U.S. Fish and Wildlife Service, Wildlife and Sport Fish Restoration Program. EE Fischer was supported by the Hans Rausing Scholarship in the History of Science. TA Zarrillo was additionally funded by Hatch funds from the Connecticut Agricultural Experiment Station and the Connecticut Department of Energy and Environmental Protection

Wildlife Division and the federal State Wildlife Grants Program. Additional partial funding was provided by the iDigBees TCN NSF award# 2216927.

Author contributions

The project was **conceptualised** by JB Dorey and KLJ Hung. The project design was guided by JB Dorey, MS Rogan, YV Sica, MC Orr, LM Guzman, JS Ascher, AC Hughes, and NS Cobb. **Data were provided and/or curated** by JB Dorey, PR Chesshire, A Nava-Bolaños, SM Collins, EM Lichtenberg, EM Tucker, A Smith-Pardo, A Falcon-Brindis, DE de Pedro, KA Parys, RL Minckley, T Griswold, T Zarillo, and JA Ascher. The **code was drafted** by JB Dorey, tested by JB Dorey and EE Fischer, and input was provided by RL O'Reilly, S Bossert, and SM Collins. R Markdown was generated by R O'Reilly with input from JB Dorey. Figures were created by JB Dorey and S Bossert. The **manuscript was drafted** by JB Dorey and all authors provided critical feedback and reviewed the final manuscript.

Competing interests

The authors declare no conflicts of interest.

Figures

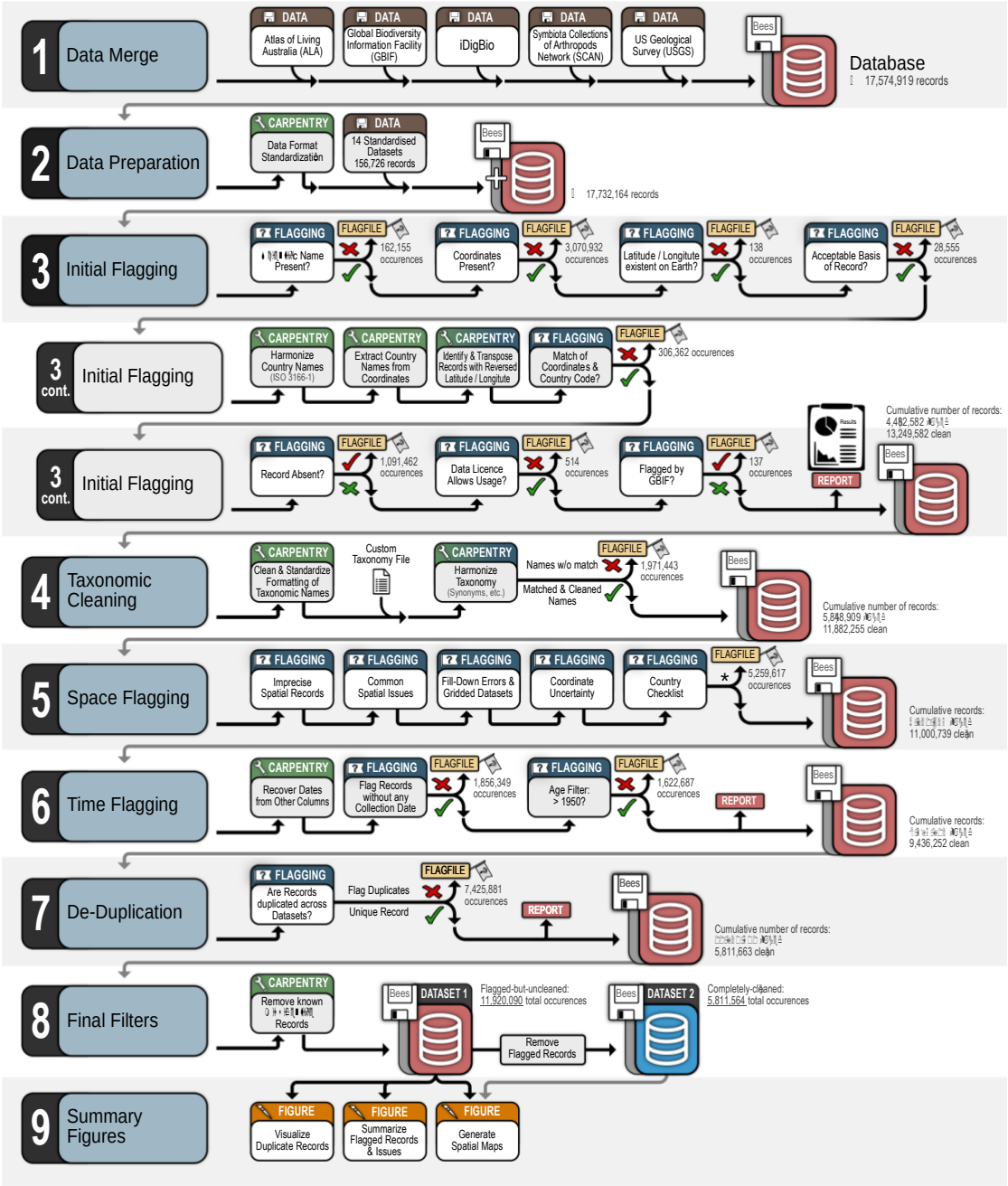

Figure 1.

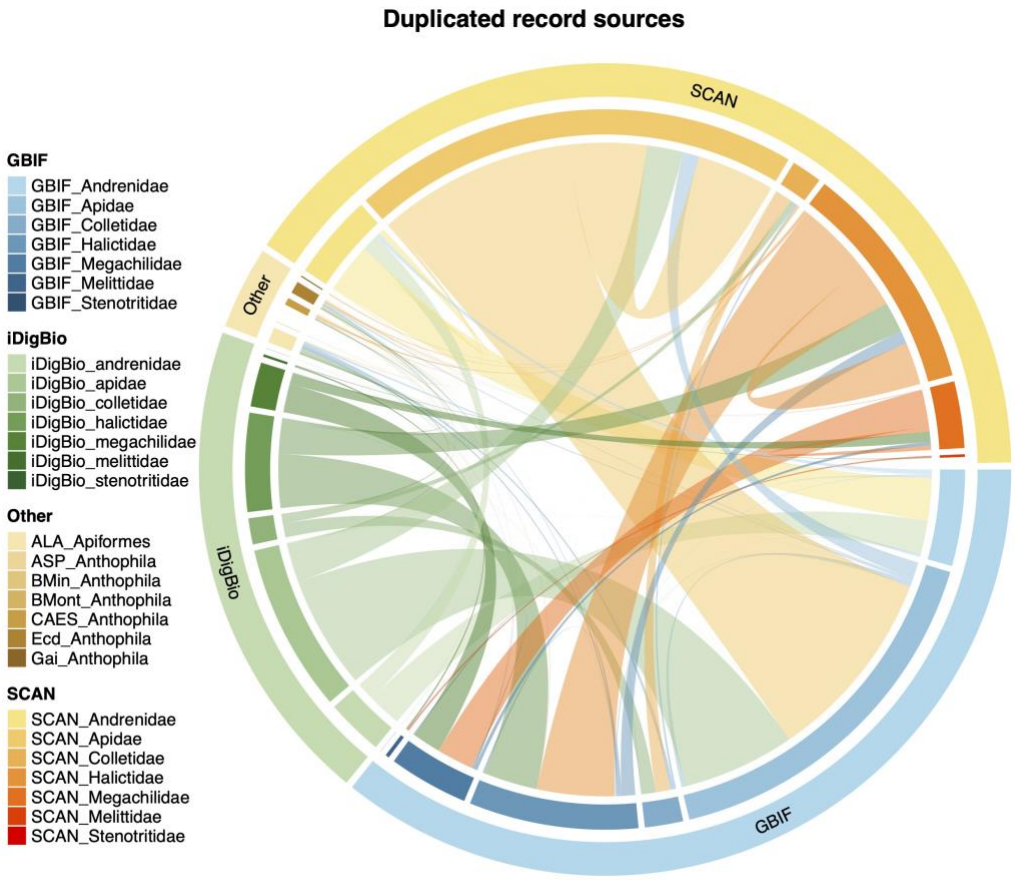
Duplicated record sources
GBIF
GBIF_Andrenidae
GBIF_Apidae
GBIF_Colletidae
GBIF_Halictidae
GBIF_Megachilidae
GBIF_Melittidae
GBIF_Stenotritidae
iDigBio
iDigBio_andrenidae
iDigBio_apidae
iDigBio_colletidae
iDigBio_halictidae
iDigBio_megachilidae
iDigBio_melittidae
iDigBio_stenotritidae
Other
ALA_Apiformes
ASP_Anthophila
BMin_Anthophila
BMont_Anthophila
CAES_Anthophila
Ecd_Anthophila
Gai_Anthophila
SCAN
SCAN_Andrenidae
SCAN_Apidae
SCAN_Colletidae
SCAN_Halictidae
SCAN_Megachilidae
SCAN_Melittidae
SCAN_Stenotritidae
SCAN
Other
iDigBio
GBIF

Figure 2.

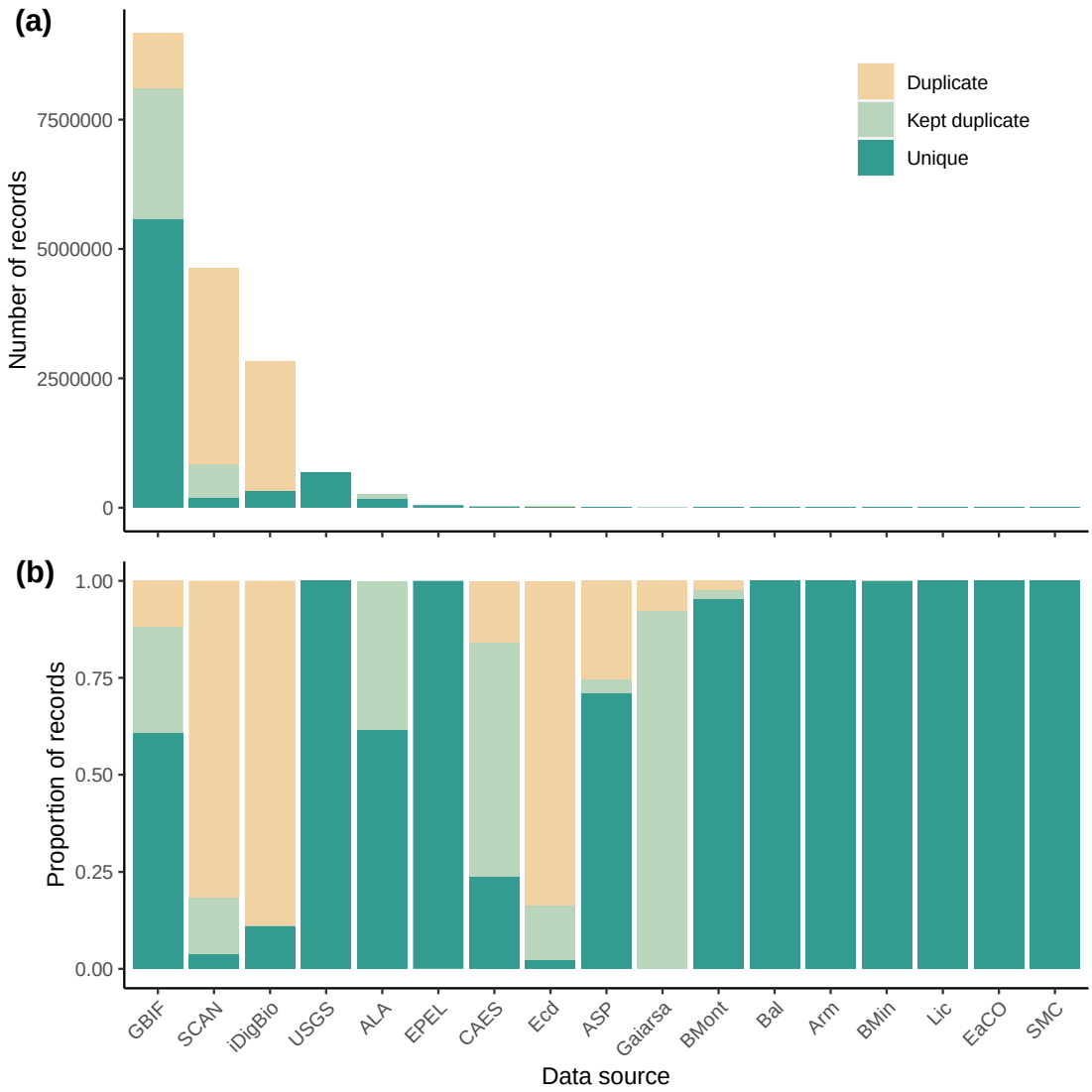

Figure 3.

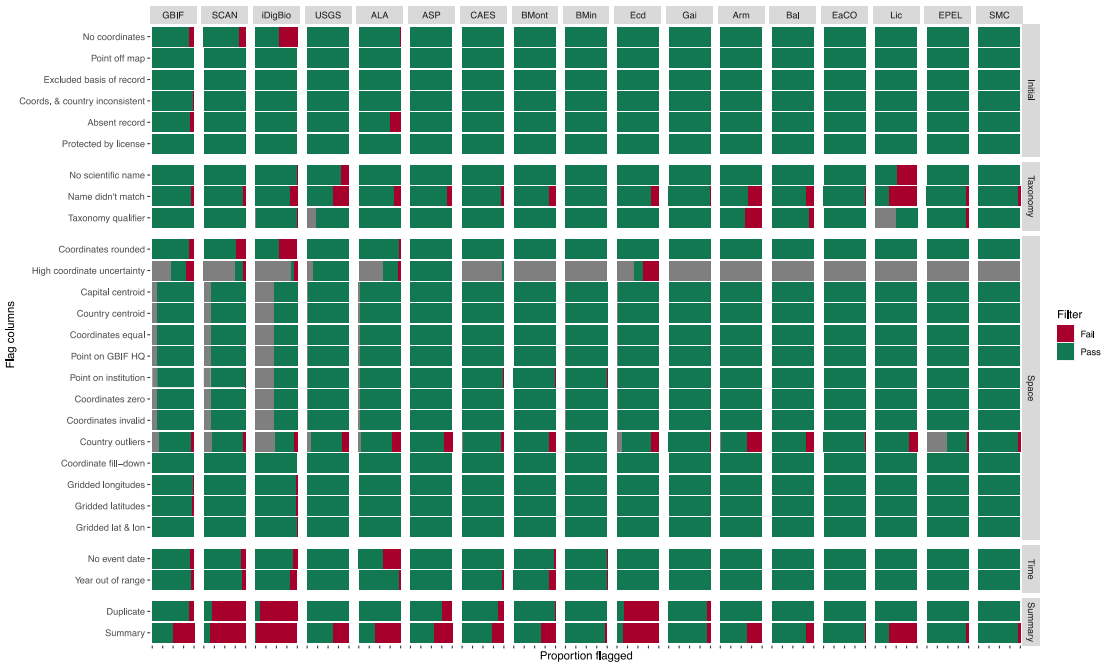

Figure 4.

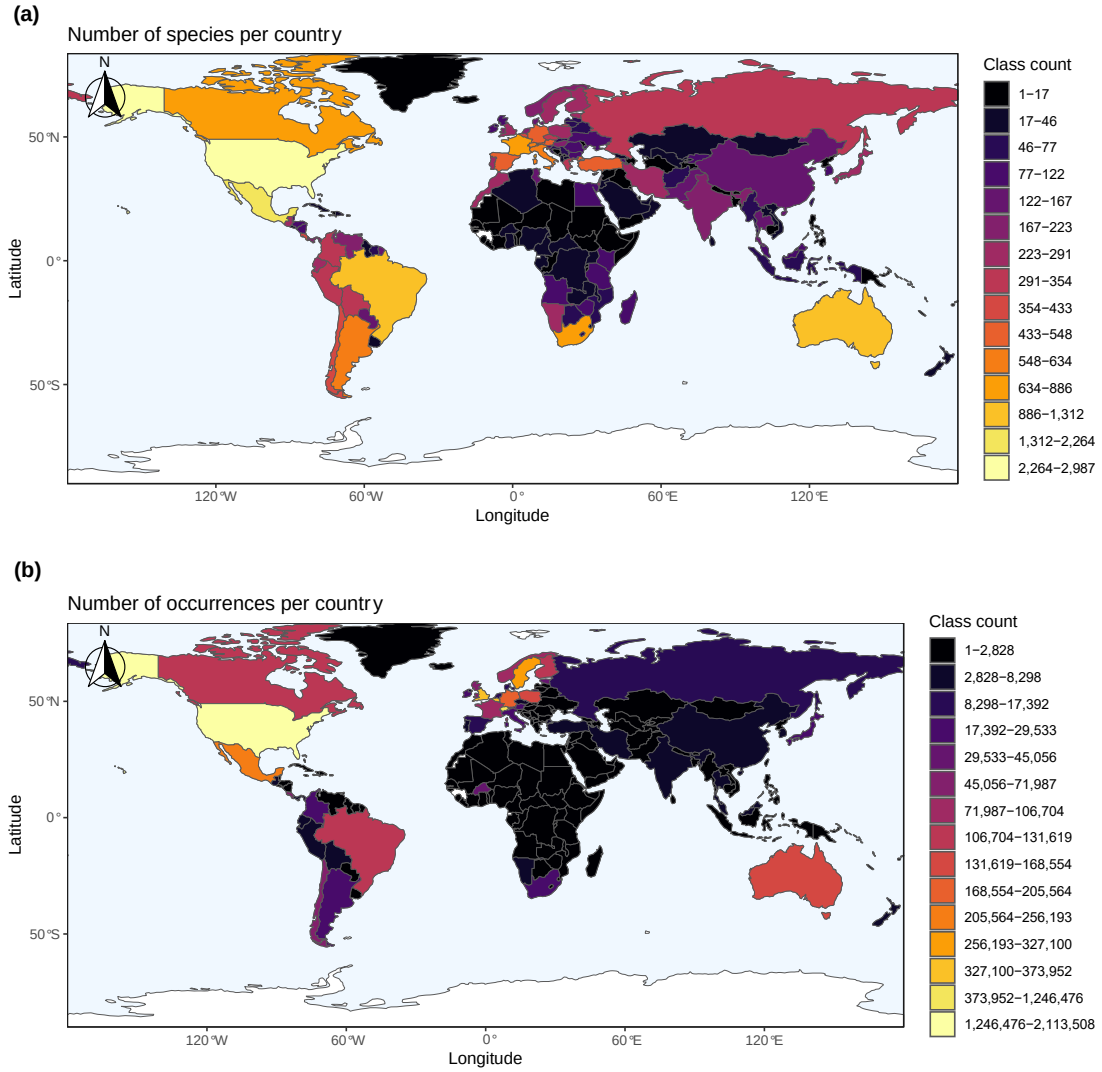

Figure 5.

Figure Legends

Figure 1. Visual summary of our workflow for compiling, cleaning, flagging, and summarizing the most extensive, publicly available data set of native bee occurrences. Asterisk indicates that the number of records flagged in (5) Space Flagging includes all steps.

Figure 2. A chord diagram showing duplications between major data sources. The major data repositories, the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), and Symbiota Collections of Arthropods Network (SCAN) are indicated on the diagram's outer ring with bee families indicated on the associated inner ring. Smaller collections are indicated under 'Other' on the outer ring and the associated inner ring indicates individual datasets. The 'Other' collections are the Atlas of Living Australia (ALA), Allan Smith-Pardo (ASP), Bob Minckley (BMin), *Bombus* Montana (BMont), The Connecticut Agricultural Experiment Station (CAES), Ecdysis (Ecd), Gaiarsia (Gai), Elinor Lichtenberg (Lic)[29], and the United States Geological Survey (USGS) data. The size of each ring and inner linkage (chord) is relative to the number of occurrences. The chords link occurrences are that are duplicated between data sources[85].

Figure 3. Duplicate occurrence summary. Two bar plots showing (a) the total number of records and (b) the proportion of records in each dataset that were duplicates (sand), kept

duplicates (light green), and unique (dark green). Duplicates are occurrences that were identified to have a match in another or the same dataset and that were thus flagged or discarded. Kept duplicates are the same as duplicates, except they are the version of the occurrence records that were kept. Unique occurrences are those that were not matched to any other occurrences and were also kept. The included datasets are the Global Biodiversity Information Facility (GBIF), Symbiota Collections of Arthropods Network (SCAN), Integrated Digitized Biocollections (iDigBio), the United States Geological Survey (USGS), the Atlas of Living Australia (ALA), Elle Pollination Ecology Lab (EPEL), the Connecticut Agricultural Experiment Station (CAES), Ecdysis (Ecd), Allan Smith-Pardo (ASP), Gaiarsia (Gai), *Bombus* Montana (BMont), Ballare, et al. [28] (Bal), Armando (Arm), Bob Minckley (BMin), Elinor Lichtenberg (Lic)[29], Eastern Colarado (EaCO), and Texas literature data (SMC).

Figure 4. Flag summary. A compound bar plot showing the proportion of all occurrences within each data source (x-axis) that were flagged for each filtering step (y-axis). Green indicates the proportion of occurrences that passed a filter and red indicates the proportion that failed a filter. The summary row indicates the total proportion of passed or failed occurrences for each dataset based on those that were chosen to be filtered. In this instance, (i) taxonomic qualifier, (ii) gridded longitudes, (iii) gridded latitudes, and (iv) gridded latitude and longitude were not included in the summary row. The included datasets are the Global Biodiversity Information Facility (GBIF), Symbiota Collections of Arthropods Network (SCAN), Integrated Digitized Biocollections (iDigBio), the United States Geological Survey (USGS), the Atlas of Living Australia (ALA), Allan Smith-Pardo (ASP), the Connecticut Agricultural Experiment Station (CAES), *Bombus* Montana (BMont), Bob Minckley (BMin), Ecdysis (Ecd), Gaiarsia (Gai), Elle Pollination Ecology Lab (EPEL), Armando (Arm), Ballare, et al. [28] (Bal), Eastern Colarado (EaCO), Elinor Lichtenberg (Lic)[29], and Texas literature data (SMC).

Figure 5. Occurrence-country summary maps created using the cleaned data indicating the (a) number of species per country and (b) number of occurrences per country from the filtered data. Colours indicate the number of (a) species or (b) occurrences where dark colours are low and yellow colours are high. Class intervals were defined using a “fisher” method.

Tables

Supplementary Table 1. Download and citation information for the major data sources, the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), and Symbiota Collections of Arthropods Network (SCAN), the Atlas of Living Australia (ALA), and the United States Geological Survey (USGS). This table includes the (i) alternate identifier; (ii) data title; (iii) publication date; (iv) download date stamp; (v) doi; and where provided, (vi) download link; (vii) download citation; and (viii) usage rights.

Supplementary Table 2. The custom taxonomy list that identifies (i) incorrect names (*Original*; often orthographic variants), (ii) their valid match (*Correct*; with or without an authority), and (iii) the point at which they were identified (*Source*).

Supplementary Table 3. The master taxonomy list used for bee species. The taxonomy list was generated using the Discover Life website that is maintained by Ascher and Pickering [19] and manual error-checking (Supplementary Table 2). The table includes (i) flags; (ii) taxonomic status; (iii) name source; (iv) the accepted name id (zero for accepted names); (v) the unique name id number; (vi) higher taxonomy (kingdom-subtribe); (vii) the valid name with authorship, without authorship, and without authorship but with DiscoverLife flags; (viii) lower taxonomy (genus-infraspecies); (ix) scientific name authorship; (x) taxonomic rank of name; (xi) TRUE/FALSE column to indicate valid names; and (xii) notes on the name.

Supplementary Table 4. The country checklist as generated using the Discover Life website that is maintained by Ascher and Pickering [19]. The checklist contains the species name (*ValidName*) and country (*ISO*) information which is then matched against rNaturalEarth[76] data (including a per-country count; *Freq*) and harmonised with the master taxonomy (Supplementary Table 3).

Supplementary Table 5. The table used to create Fig. 4. It contains the database name, the flag name, the flag value (where TRUE = passed, FALSE = failed, and NA = unable to assess), the occurrence count, and the flag type.

Supplementary Table 6. A test version of SupplementaryFile2_05_unCleaned_database.csv where every 1000th occurrence (row) has been kept and the intervening occurrences removed.

References

1 Tu, D. V. *et al.* Taxonomy notes and new occurrence data of four species of atyid shrimp (Crustacea: Decapoda: *Atyidae*) in Vietnam, all described from China. *Biodivers. Data J.* **9**, e70289 (2021). https://doi.org:https://doi.org/10.3897%2FBDJ.9.e70289

2 Dar, G. H., Khuroo, A. A., Reddy, C. S. & Malik, A. H. Impediment to taxonomy and its impact on biodiversity science: an Indian perspective. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences* **82**, 235-240 (2012). https://doi.org:https://doi.org/10.1007/s40011-012-0031-3

3 Dorey, J. B., Fagan-Jeffries, E. P., Stevens, M. I. & Schwarz, M. P. Morphometric comparisons and novel observations of diurnal and low-light-foraging bees. *J. Hymenoptera Res.* **79**, 117–144 (2020). https://doi.org:https://doi.org/10.3897/jhr.79.57308

4 Dorey, J. B. Missing for almost 100 years: the rare and potentially threatened bee *Pharohylaeus lactiferus* (Hymenoptera, Colltidae). *J. Hymenoptera Res.* **81**, 165-180 (2021). https://doi.org: https://doi.org/10.3897/jhr.81.59365

5 Dorey, J. B. *et al.* Continental risk assessment for understudied taxa post catastrophic wildfire indicates severe impacts on the Australian bee fauna. *Global Change Biol.* **27**, 6551-6567 (2021). https://doi.org:https://doi.org/10.1111/gcb.15879

6 Orr, M. C. *et al.* Global Patterns and Drivers of Bee Distribution. *Curr. Biol.* **31**, 451-458.e454 (2021). https://doi.org:https://doi.org/10.1016/j.cub.2020.10.053

7 Kass, J. M. *et al.* The global distribution of known and undiscovered ant biodiversity. *Science Advances* **8**, eabp9908 (2022). https://doi.org:https://doi.org/10.1126/sciadv.abp9908

8 Murray, E. A. *et al.* Phylogeny, Phenology, and Foraging Breadth of *Ashmeadiella* (Hymenoptera: Megachilidae). *Insect Systematics and Diversity* **5** (2021). https://doi.org:https://doi.org/10.1093/isd/ixab010

9 Callaghan, C. T. *et al.* The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLoS Biol.* **20**, e3001843 (2022). https://doi.org:https://doi.org/10.1371/journal.pbio.3001843

10 CDB. First draft of the post-2020 global biodiversity framework by the Convention on Biological Diversity. (2021).

11 Chapman, A. Principles of Data Quality. (Copenhagen, 2005).

12 Ribeiro, B. R. *et al. bdc*: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods Ecol. Evol.* **13**, 1421-1428 (2022). https://doi.org:https://doi.org/10.1111/2041-210X.13868

13 Peterson, A. T., Asase, A., Canhos, D. A. L., de Souza, S. & Wieczorek, J. Data Leakage and Loss in Biodiversity Informatics. *Biodivers. Data J.* **6**, e26826 (2018).

14 Chapman, A. Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data version 1.0. (Copenhagen, 2005).

15 Boakes, E. H. *et al.* Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* **8**, e1000385 (2010). https://doi.org:https://doi.org/10.1371/journal.pbio.1000385

16 Pinkert, S., Barve, V., Guralnick, R. & Jetz, W. Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Global Ecol. Biogeogr.* **31**, 830-839 (2022). https://doi.org:https://doi.org/10.1111/geb.13475

17 Ollerton, J. Pollinator diversity: distribution, ecological function, and conservation. *Annual Review of Ecology, Evolution and Systematics* **48** (2017). https://doi.org:https://doi.org/10.1146/annurev-ecolsys-110316-022919

18 Danforth, B. N., Minckley, R. L., Neff, J. L. & Fawcett, F. *The solitary bees: biology, evolution, conservation*. (Princeton University Press, 2019).

19 Ascher, J. S. & Pickering, J. *Discover Life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila).* *http://www.discoverlife.org/mp/20q?guide=Apoidea_species*, 2020).

20 Guzman, L. M., Kelly, T. & Elle, E. A data set for pollinator diversity and their interactions with plants in the Pacific Northwest. *Ecology* **n/a**, e3927 (2022). https://doi.org:https://doi.org/10.1002/ecy.3927

21 Delphia, C. M. Bumble bees of Montana. https://www.mtent.org/projects/Bumble_Bees/bombus_species.html. (2022)

22 Ecdysis. *Ecdysis: a portal for live-data arthropod collections*, <https://serv.biokic.asu.edu/ecdysis/index.php> (2022).

23 Gaiarsa, M. P., Kremen, C. & Ponisio, L. C. Pollinator interaction flexibility across scales affects patch colonization and occupancy. *Nature Ecology & Evolution* **5**, 787-793 (2021). https://doi.org:https://doi.org/10.1038/s41559-021-01434-y

24 Auerbach, E. S., Johnson, W. P., Smith, J. R. & McIntyre, N. E. Wildlife refuges support high bee diversity on the Southern Great Plains. *Environ. Entomol.* **48**, 968-976 (2019). https://doi.org:https://doi.org/10.1093/ee/nvz063

25 Angela, B., Lisa, M. O., Loren, M. S. & Scott, T. M. A survey of the insects of the Southern High Plains (Llano Estacado) of Texas, with particular reference to pollinators and other anthophiles. *J. Kans. Entomol. Soc.* **91**, 255-309 (2019). https://doi.org:https://doi.org/10.2317/0022-8567-91.4.255

26 Cate, C. A. *Monitoring, assessing and evaluating the pollinator species (Hymenoptera: Apoidea) found on a native brush site, a revegetated site and an urban garden* Doctoral dissertation thesis, Texas A&M University., (2007).

27 Cusser, S., Neff, J. L. & Jha, S. Land-use history drives contemporary pollinator community similarity. *Landscape Ecol.* **33**, 1335-1351 (2018). https://doi.org:https://doi.org/10.1007/s10980-018-0668-2

28 Ballare, K. M., Neff, J. L., Ruppel, R. & Jha, S. Multi-scalar drivers of biodiversity: local management mediates wild bee community response to regional urbanization. *Ecol. Appl.* **29**, e01869 (2019). https://doi.org:https://doi.org/10.1002/eap.1869

29 Lichtenberg, E. M., Milosavljević, I., Campbell, A. J. & Crowder, D. W. Differential effects of soil conservation practices on arthropods and crop yield. *bioRxiv*, https://doi.org/10.1101/2021.1112.1106.471474 (2022). https://doi.org:10.1101/2021.12.06.471474

30 Dorey, J. B., Schwarz, M. P. & Stevens, M. I. Review of the bee genus *Homalictus* Cockerell (Hymenoptera: Halictidae) from Fiji with description of nine new species. *Zootaxa* **4674**, 1–46 (2019). https://doi.org:https://doi.org/10.11646/zootaxa.4674.1.1

31 GBIF.org. GBIF occurrence download, Stenotritidae. https://doi.org/10.15468/dl.kt889x. (2022)

32 GBIF.org. GBIF occurrence download, Apidae. https://doi.org/10.15468/dl.e8ykkp. (2022)

33 GBIF.org. GBIF occurrence download, Megachilidae. https://doi.org/10.15468/dl.qzc4ea. (2022)

34 GBIF.org. GBIF occurrence download, Colletidae. https://doi.org/10.15468/dl.8cw4p6. (2022)

35 GBIF.org. GBIF occurrence download, Halictidae. https://doi.org/10.15468/dl.pa7uwy. (2022)

36 GBIF.org. GBIF occurrence download, Andrenidae. https://doi.org/10.15468/dl.95kq7s. (2022)

37 GBIF.org. GBIF occurrence download, Melittidae. https://doi.org/10.15468/dl.chhawx. (2022)

38 SCAN. SCAN-Bugs occurrence download, Apidae (7a7eeb61-46b8-47b0-a9fa-2bb84a8abf5a). (2022).
39 SCAN. SCAN-Bugs occurrence download, Apidae (c73679eb-751b-48c2-a6f4-afc697d187c4). (2022).
40 SCAN. SCAN-Bugs occurrence download, Andrenidae (425bd610-789e-4b80-b89c-593c302978c3). (2022).
41 SCAN. SCAN-Bugs occurrence download, Colletidae (8a324774-4591-4e1a-8139-928b280702f9). (2022).
42 SCAN. SCAN-Bugs occurrence download, Megachilidae (adce8010-1fc1-47dd-b7ff-935ed476e399). (2022).
43 SCAN. SCAN-Bugs occurrence download, Melittidae (f8f1d45e-e2ed-433f-b256-277bef40339f). (2022).
44 SCAN. SCAN-Bugs occurrence download, Halictidae (45bcd0ad-dd69-4c80-b2f9-78868fded1b9). (2022).
45 SCAN. SCAN-Bugs occurrence download, Halictidae (2a2dfc93-69ec-4980-bcb8-a1d169baacba). (2022).
46 SCAN. SCAN-Bugs occurrence download, Stenotritidae (ebc56df1-6d45-4df5-8ec4-79cf60676c4a). (2022).
47 SCAN. SCAN-Bugs occurrence download, Apidae (80f8cd1d-2168-4bac-b5d2-37f00ae1ee3e). (2022).
48 iDigBio.org. iDigBio occurrence download, Halictidae. http://s.idigbio.org/idigbio-downloads/0dcb19ee-20d5-4f31-9378-ace17d4e648f.zip. (2022)
49 iDigBio.org. iDigBio occurrence download, Andrenidae. http://s.idigbio.org/idigbio-downloads/1628cae7-9c45-4c6f-8030-4b85e55fd8a3.zip. (2022)
50 iDigBio.org. iDigBio occurrence download, Colletidae. http://s.idigbio.org/idigbio-downloads/33d3eb54-aa6f-48bd-ab3e-0edd199ce87c.zip. (2022)
51 iDigBio.org. iDigBio occurrence download, Stenotritidae. http://s.idigbio.org/idigbio-downloads/5aa5abe1-62e0-4d8c-bebf-4ac13bd9e56f.zip. (2022)
52 iDigBio.org. iDigBio occurrence download, Melittidae. http://s.idigbio.org/idigbio-downloads/a1f9d87e-b68d-4152-b2bc-f5b22fa861b4.zip. (2022)
53 iDigBio.org. iDigBio occurrence download, Megachilidae. http://s.idigbio.org/idigbio-downloads/b9105de9-40e2-45db-ab94-50f0bd532eb1.zip. (2022)
54 iDigBio.org. iDigBio occurrence download, Apidea. http://s.idigbio.org/idigbio-downloads/cb6ec734-00b1-47d7-811e-90ec9ce9ebb7.zip. (2022)
55 ALA.org.au. ALA occurrence download. https://doi.org/10.26197/ala.cdb5a16c-5f19-4dee-b584-5f2f40196bd9. (2022)
56 Stevenson, M., Westgate, M. & Newman, P. *galah*: Atlas of Living Australia (ALA) data and resources in *R* v. *R* package version 1.4.0 (2022).
57 R Development Core Team. R: a language and environment for statistical computing (Vienna, Austria, 2019).
58 Wickham, H., François, R., Henry, L. & Müller, K. *dplyr*: a grammar of data manipulation v. *R* package version 1.0.9 (2022).
59 Bache, S. M. & Wickham, H. *magrittr*: a forward-pipe operator for *R* v. *R* package version 2.0.3 (2022).
60 Müller, K. & Wickham, H. *tibble*: simple data frames v. *R* package version 3.1.8 (2022).
61 Wickham, H. *stringr*: simple, consistent wrappers for common string operations v. *R* package version 1.4.0 (2019).
62 Henry, L. & Wickham, H. *tidyselect*: select from a set of strings v. R package version 1.1.2 (2022).
63 Wickham, H. & Bryan, J. *readxl*: read excel files v. *R* package version 1.4.0 (2022).
64 Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer-Verlag, 2016).

65 Wickham, H. *forcats*: tools for working with categorical variables (factors) v. *R* package version 0.5.1 (2021).
66 Wickham, H. & Girlich, M. *tidyr*: tidy messy data v. *R* package version 1.2.0 (2022).
67 Henry, L. & Wickham, H. *rlang*: functions for base types and core *R* and '*tidyverse*' features v. *R* package version 1.0.4 (2022).
68 Wickham, H., Hester, J. & Ooms, J. *xml2*: parse XML v. *R* package version 1.3.3 (2021).
69 Wickham, H., Hester, J. & Bryan, J. *readr*: read rectangular text data v. *R* package version 2.1.2 (2022).
70 Grolemund, G. & Wickham, H. Dates and times made easy with *lubridate*. *Journal of Statistical Software* **40**, 1–25 (2011). https://doi.org:https://doi.org/10.18637/jss.v040.i03
71 Zizka, A. *et al. CoordinateCleaner*: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10**, 744-751 (2019). https://doi.org:https://doi.org/10.1111/2041-210X.13152
72 Bengtsson, H. *R.utils*: various programming utilities v. R package version 2.12.0 (2022).
73 Ren, K. *rlist*: a toolbox for non-tabular data manipulation v. *R* package version 0.4.6.2 (2021).
74 Boettigeros, C. & Jones, M. B. *EML*: read and write ecological metadata language files v. R package version 2.0.6.1 (2022).
75 Boettiger, C. Ecological metadata as linked data. *The Journal of Open Source Software* **4**, 1276 (2019). https://doi.org:https://doi.org/10.21105/joss.01276
76 South, A. *rnaturalearth*: world map data from Natural Earth v. *R* package version 0.1.0 (2017).
77 South, A. *rnaturalearthdata*: world vector map data from Natural Earth used in '*rnaturalearth*' v. *R* package version 0.1.0 (2017).
78 Arel-Bundock, V., Enevoldsen, N. & Yetman, C. *countrycode*: An *R* package to convert country names and country codes. *Journal of Open Source Software* **3**, 848 (2018). https://doi.org:https://doi.org/10.21105/joss.00848
79 Davis-Rabosky, A. *et al.* Coral snakes predict the evolution of mimicry across New World snakes. *Nat. Commun.* **7**, 11484 (2017).
80 South, A. *rworldmap*: a new *R* package for mapping global data. *The R Journal* **3**, 35–43 (2011). https://doi.org:https://doi.org/10.32614/RJ-2011-006
81 Carr, D., Lewin-Koh, N., Maechler, M. & Sarkar, D. *hexbin*: hexagonalbBinning routines v. *R* package version 1.28.2 (2021).
82 Wilke, C. O. *cowplot*: streamlined plot theme and plot annotations for 'ggplot2' v. R package version 1.0.0 (2019).
83 Dunnington, D. ggspatial: spatial data framework for ggplot2 (2021).
84 Ushey, K. *renv*: project environments v. *R* package version 0.15.5 (2022).
85 Flor, M. *chorddiag*: interactive chord diagrams v. *R* package version 0.1.3 (2022).
86 Csardi, G. & Nepusz, T. The *igraph* software package for complex network research. *InterJournal*, 1695 (2006).
87 Pebesma, E. Simple features for *R*: standardized support for spatial vector data. *The R Journal* **10**, 439–446 (2018). https://doi.org:https://doi.org/10.32614/RJ-2018-009
88 Hijmans, R. J. *terra*: spatial Data analysis v. *R* package version 1.5-21 (2022).
89 Wieczorek, J. *et al.* Darwin Core: an evolving community-developed biodiversity data standard. *PLOS ONE* **7**, e29715 (2012). https://doi.org:https://doi.org/10.1371/journal.pone.0029715
90 Chesshire, P. R. *et al.* Completeness analysis for over 3000 United States bee species identifies persistent data gaps. *Ecography* (In Press).
91 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. *circlize* Implements and enhances circular visualization in *R*. *Bioinformatics* **30**, 2811-2812 (2014). https://doi.org:https://doi.org/10.1093/bioinformatics/btu393

92 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016). https://doi.org:https://doi.org/10.1093/bioinformatics/btw313

93 Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022). https://doi.org:https://doi.org/10.1002/imt2.43

94 Hvitfeldt, E. *paletteer*: Comprehensive Collection of Color Palettes v. 1.3.0 (2021).

95 Fisher, W. D. On grouping for maximum homogeneity. *Journal of the American Statistical Association* **53**, 789–798 (1958). https://doi.org:https://doi.org/10.1080/01621459.1958.10501479

96 Bivand, R. *classInt*: Choose Univariate Class Intervals v. 0.4–8 (2022).