Applications in Plant Sciences

florabr: An R package to explore and spatialize species distribution using Flora e Funga do Brasil

Weverton C. F. Trindade [iD]

Laboratory of Plant Ecology, Federal University of Paraná, Curitiba, Paraná, Brazil

Correspondence

Weverton C. F. Trindade, Laboratory of Plant Ecology, Federal University of Paraná, Curitiba, Paraná, Brazil.
Email: wevertonf1993@gmail.com

Abstract

Premise: The Flora e Funga do Brasil project is the most comprehensive effort to reliably document Brazilian plant and fungal diversity. It involves the collaborative work of hundreds of taxonomists, integrating detailed and standardized morphological descriptions, nomenclatural status, and geographic distribution information of plants, algae, and fungi collected throughout Brazil. Despite the extensive information available, managing the information from the Flora e Funga do Brasil website poses certain challenges.

Methods and Results: florabr is an R package developed to facilitate the exploration and geographical analysis of species information derived from the Flora e Funga do Brasil. Unique to florabr is its ability to interact with the latest, or any other version of the dataset, which undergoes weekly updates. I illustrate the practical application of florabr in common tasks in biogeography and conservation studies.

Conclusions: florabr is anticipated to be of significant interest to biogeographers, ecologists, curators of biological collections, and taxonomists actively contributing to the Flora e Funga do Brasil.

KEYWORDS

biodiversity dataset, data cleaning, online flora, plant distribution, plant diversity

One of the primary goals outlined by the Global Strategy for Plant Conservation (GSPC), adopted by the Parties to the Convention on Biological Diversity (2012), is to understand, document, and recognize plant diversity. GSPC's Target 1 for 2020 called for the creation of an online flora encompassing all known plants. Brazil, boasting the highest biodiversity and one of the highest levels of endemism on Earth, achieved this objective through the Flora do Brasil 2020 project (Brazil Flora Group, 2021). This initiative resulted in the creation of a comprehensive list of species freely available online (Brazil Flora Group, 2018). This digital biodiversity database involved the effort of hundreds of taxonomists, integrating data from plants, algae, and fungi collected in Brazil during the past two centuries (Brazil Flora Group, 2018), and is the most comprehensive work to reliably document Brazilian plant, algae, and fungal diversity. The database includes detailed and standardized morphological descriptions, illustrations, nomenclatural

data, geographic distribution, and keys for the identification of all native and non-native plants, algae, and fungi found in Brazil (Brazil Flora Group, 2021). In 2021, the project transitioned into a new phase, now titled Flora e Funga do Brasil, with a continued mission to enhance our understanding of Brazil's plant, algae, and fungal diversity. This phase particularly emphasizes improving the representation of algae and fungi, which are underrepresented in the database (Brazil Flora Group, 2021).

Digital biodiversity databases documenting the distribution of species are a central element in efforts to understand and conserve biodiversity (Whittaker et al., 2005; Peterson et al., 2018). The confluence of rapid technological advancements and the escalating biodiversity crisis has motivated the collection and availability of a massive volume of digital biodiversity data (Nelson and Ellis, 2019), opening new pathways for discovering complex biodiversity patterns (Soltis and Soltis, 2016; Folk and Siniscalchi, 2021).

Nevertheless, our capacity to efficiently provide relevant information addressing contemporary societal challenges depends on our ability to access, share, integrate, and compare data across diverse databases (Hampton et al., 2013).

Here, I introduce florabr, an R package designed to facilitate the exploration and downstream spatial analysis of species information derived from the Flora e Funga do Brasil, which will serve as a valuable tool for common tasks in taxonomy, biogeography, and conservation studies, especially ecological niche modeling. These tasks include resolving taxonomic issues, selecting target species based on their traits and distribution, and identifying erroneous occurrence records that deviate from the natural ranges of the species. florabr is anticipated to be of significant interest to biogeographers, ecologists, conservationists, curators of biological collections, and taxonomists actively contributing to the Flora e Funga do Brasil initiative. The florabr package is implemented in R (R Core Team, 2023).

METHODS AND RESULTS

The package is accompanied by a proposed workflow that can be applied to common tasks in biogeography and conservation studies (Figure 1). As an illustrative example, I employed the package to select all species of endemic shrubs with accepted names and confirmed occurrences in the Atlantic Forest (Appendix S1). Subsequently, I verified the validity and status of taxonomic names from a species list and retrieved information about these species from the Flora e Funga do Brasil. Additionally, I flagged occurrence records of the endemic shrub *Abatia americana* (Gardner) Eichler that fall outside its natural range, as determined by specialists in Flora e Funga do Brasil.

Loading the Flora e Funga do Brasil dataset

All data included in the Flora e Funga do Brasil website (https://floradobrasil.jbrj.gov.br/consulta/) are stored in seven separate Darwin Core Archive datasets that are updated weekly. The *get_florabr()* function downloads these datasets, translates the information from Portuguese to English, and merges them into a single file. The function requires a few seconds to download the data and a few minutes to merge the datasets. By default, the function downloads the most recent version available, but there is an option to download any older version as well. The *check_version()* function verifies the presence of any dataset version in the specified directory and checks if it includes the latest version available.

To utilize the other functions of the package, it is necessary to load the data into the R environment using the *load_florabr()* function. Users can also specify a particular version using the *data_version* parameter. A link to access available previous versions can be found in Appendix S1

FIGURE 1 The four main steps of the workflow proposed here for using florabr to help with common tasks in biogeography and conservation studies. Each of the four steps is represented by brown boxes, while their main florabr functions are illustrated in off-white boxes.

(in the section "Get data from Brazilian Flora"). Additionally, users can choose between two versions of the data: the "short" version and the "complete" version. The short version contains 23 essential columns required for running the package's functions. These columns include information on taxonomy (species, scientific name, accepted name, kingdom, group, subgroup, phylum, class, order, family, genus, taxonomic status, nomenclatural status, vernacular name, taxon rank, and ID number), distribution (states, biomes, vegetation types with confirmed occurrences of the species, origin, and endemism), and characteristics (life form and habitat). Conversely, the complete version encompasses all 39 original columns of the dataset. This includes comprehensive taxonomic details such as the author and year of publication of the scientific name, additional ID numbers, bibliographic citation, and reference link.

In the original dataset, discrepancies may arise concerning species and subspecies/varieties information. For instance, in version 399.401 of the Flora e Funga do Brasil, the species *Acianthera ochreata* (Lindl.) Pridgeon & M. W. Chase is listed in the database as having confirmed occurrences in two biomes: Caatinga and Cerrado. However, there are two subspecies of *A. ochreata*: *A. ochreata* subsp. *cylindrifolia* (Borba & Semir) Borba, with confirmed occurrences in the Cerrado, and *A. ochreata* subsp. *ochreata*, with confirmed occurrences in the Caatinga, Cerrado, and Atlantic Forest. Through the *solve_discrepancies()* function, we can retrieve information from subspecies and varieties with accepted names to update the information of the related species. For example, after applying the *solve_discrepancies()* function to the loaded dataset, the biomes with confirmed occurrences of the species *A. ochreata* are updated to Caatinga, Cerrado, and Atlantic Forest (the latter retrieved from the subspecies). This function to resolve discrepancies can be applied to the data loaded with the *load_florabr()* function. Additionally, discrepancies can be resolved during data acquisition by setting "solve_discrepancy" to TRUE in the *get_florabr()* function, resulting in the saving of a final dataset with resolved discrepancies.

Select species

Selecting a list of target species is a common task in macroecological and conservation studies. One of the primary objectives of florabr is to assist in selecting a species list based on taxonomic classification (kingdom, group, family, and genus), characteristics (life form and habitat), and distribution (federal states, biomes, vegetation types, origin, and endemism). The *get_attributes()* function displays all the options available for filtering species by characteristics (Appendix S2).

The *select_species()* function allows using these attributes to filter and select species with determined characteristics and distribution. The function returns a subset of the dataset loaded by the *load_florabr()* function,

encompassing detailed information on the identity, taxonomy, distribution, and characteristics of the selected species. In the example provided in Appendix S1, I selected all native and endemic shrubs with accepted names and confirmed occurrences in the Atlantic Forest biome. As I employed the filters *filter_LifeForm* = "in" and *filter_Biome* = "in", the selection included shrub species that could also have other life forms and occur in other biomes. *Myrcia eriopus* DC., for example, is classified as a shrub but can also exhibit a subshrub life form and can be found in the Cerrado biome. To refine the filter to exclusively include species that are solely shrubs and are found only in the Atlantic Forest, we can set both *filter_LifeForm* and *filter_Biome* to "only". Apart from "in" and "only", the filter_* arguments accept two more values: "not_in" to select species not matching the specified criteria, and "and" to exclusively select species matching all specified conditions.

Check names and subset species

In some cases, users may already have a list of species and wish to retrieve information about these species from Flora e Funga do Brasil. Before proceeding, it is essential to ensure that the list contains only binomial names (genus + specific epithet). To accomplish this, we can utilize the *get_binomial()* function to extract the binomial name from a full scientific name. I strongly recommend that users employ the *check_names()* function to verify the spelling, nomenclatural status, and taxonomic status of species. This ensures that the species names correspond exactly to those in the Flora e Funga do Brasil. In the example provided (Appendix S1), the function returns a dataframe indicating that *Abatia americana* and *Solanum restingae* S. Knapp are spelled correctly and have accepted names. Although *Araucaria brasiliana* is correctly spelled, it is identified as a synonym of *Araucaria angustifolia* (Bertol.) Kuntze. The spelling of *Butia cattarinensis* appears to be potentially incorrect because the name was not found in the database; however, a similar name, *Butia catarinensis* Noblick & Lorenzi, is suggested. Given that *Homo sapiens* is not a plant, the spelling of *Homo sapiens* was flagged as "Not_found" as the name was not found in the database and no comparable names were available. With the binomial names checked, we can safely subset the species and retrieve their information provided in the Flora e Funga do Brasil.

The package also features the *select_by_vernacular()* function, enabling the search for species based on their vernacular names. The function can perform searches for exact or fuzzy matches. For instance, one might be interested in all species popularly known as "pimenta" (pepper). If exact = TRUE, the function returns species with the exact vernacular name "pimenta." Conversely, if exact = FALSE, the function returns species encompassing other names like "pimenta-longa" (*Psittacanthus cucullaris* (Lam.) G. Don) and "ouro-pimenta" (*Ocotea canaliculata* (Rich.) Mez).

Spatialize distribution and flag records

Georeferencing errors in online species records can introduce significant bias into ecological and biogeographical research findings (Maldonado et al., 2015). The Flora e Funga do Brasil provides descriptive information based on expert knowledge on the endemism level, federal states, and biomes with confirmed occurrences of the species in Brazil, allowing users to validate occurrence records obtained from online databases (Colli-Silva et al., 2020; Reginato and Michelangeli, 2020), such as the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) and SpeciesLink (https://specieslink.net/). The *filter_florabr()* function facilitates the verification of records obtained from these databases, ensuring that the coordinates sourced from the database align with Flora e Funga do Brasil regarding endemism, states, and biomes with confirmed occurrences. For example, *Abatia americana* is a shrub endemic to Brazil with confirmed occurrences in two biomes (Atlantic Forest and Cerrado) and four states (Espírito Santo, Minas Gerais, São Paulo, and Rio de Janeiro) (Figure 2). The occurrences of *A. americana* were retrieved from GBIF (GBIF, 2024) and integrated as data examples in the package. As detailed in Appendix S1, the *filter_florabr()* function reveals that out of the 115 records of *A. americana* from GBIF, 111 align with the biomes and states reported in Flora e Funga do Brasil. Of the four records that do not agree, two fall inside Brazil, but in states and biomes not reported by specialists in Flora e Funga

do Brasil, while the other two records, in Ecuador and Belarus, have been flagged due to the species' reported endemism to Brazil. Thus, the *filter_florabr()* function can be seamlessly integrated into a workflow alongside functions from other packages that assist in flagging common spatial errors in biological collection data, such as CoordinateCleaner (Zizka et al., 2019), plantR (de Lima et al., 2023), and bdc (Ribeiro et al., 2022).

Creating a presence-absence matrix of a species

One way to organize biodiversity data is through the utilization of presence-absence matrices (PAMs). In a PAM, a value of one signifies the presence of species j in cell i, while a value of zero indicates its absence. PAMs enable the estimation of various metrics that summarize biodiversity patterns, including richness, range size, and community composition (Soberon and Cavner, 2015; Soberón et al., 2021). To retrieve the distribution information of a species and convert it into a PAM, we can employ the *get_pam()* function. Each site in the PAM can represent a biome, a state, a vegetation type, or any combination thereof. For instance, if sites are designated by both biome and state, biomes will be subdivided into several sites based on their intersection with states. To illustrate this, we will convert the dataset containing endemic shrubs with accepted names and confirmed occurrences in the Atlantic
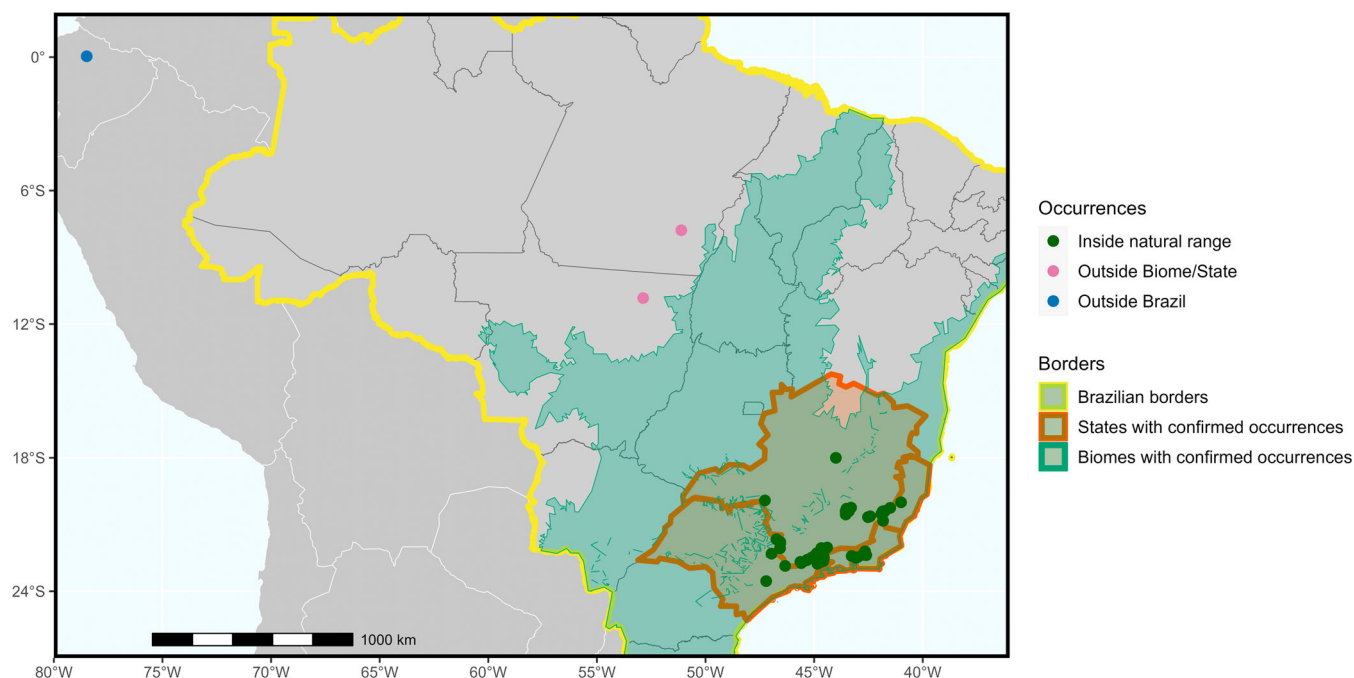

FIGURE 2 Map of flagged records of *Abatia americana* using the *filter_florabr()* function. The green-shaded area represents the biomes (Atlantic Forest and Cerrado) with confirmed occurrences of *A. americana* in Brazil, while the orange line delineates the states (Espírito Santo, Minas Gerais, São Paulo, and Rio de Janeiro) with confirmed occurrences. The yellow line marks the boundary of Brazil. Green dots indicate records aligned with the distribution provided by specialists in Flora e Funga do Brasil. Pink dots represent records flagged as potentially incorrect due to falling outside the confirmed biomes and states. A blue dot in Ecuador represents a flagged record due to the species being endemic to Brazil. Notably, there is another blue dot in Belarus (not visible on the map).
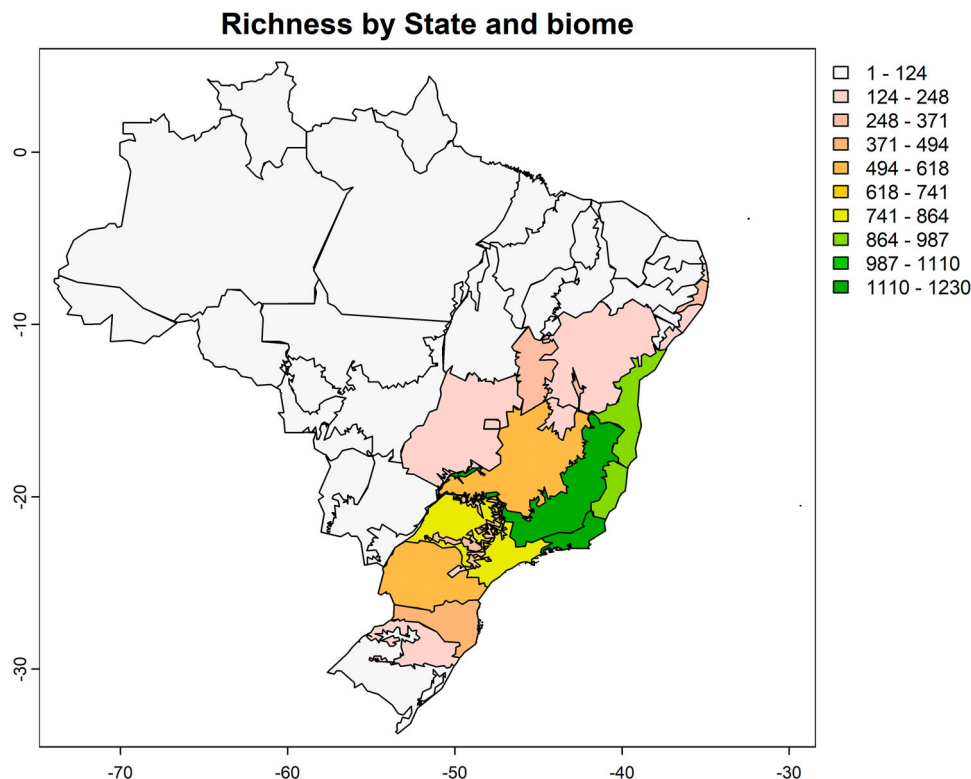
Richness by State and biome

FIGURE 3 Plot returned by the *get_pam()* function indicating the number of species in each site. Here, each site represents a combination of states and biomes.

Forest (*shrubs_in_af*) into a PAM. This conversion will utilize the combination of biomes and states as sites. The function returns a list with the PAM, a table, and a spatial object with the number of species in each site (Figure 3).

CONCLUSIONS

The florabr R package is a versatile toolkit designed to facilitate the exploration, filtering, and spatial analysis of information sourced from Flora e Funga do Brasil, the most comprehensive work to reliably document Brazilian plant and fungal diversity. Here, I illustrated the practical application of florabr in common tasks related to biogeography and conservation studies. The package's applicability has the potential to span Neotropical and global studies encompassing Brazil, a country renowned for hosting the most diverse flora on the planet (Forzza et al., 2012; Brazil Flora Group., 2021; Brummitt et al., 2021). For example, some global studies employ presence-absence data at the scale of "botanical countries" (Tietje et al., 2022), which subdivides Brazil into states or regional boundaries according to the World Geographic Scheme for Recording Plant Distributions (Brummitt, 2001). In this broader context, the florabr package proves valuable by facilitating the extraction of the most recent information to refine plant distributions in Brazil at regional, state, or biome levels.

Other R packages offer tools comparable to those found in florabr. WorldFlora (Kindt, 2020) and U.Taxonstand

(Zhang and Qian, 2023) allow matching and standardizing lists of plant names with the species names sourced from World Flora Online. CoordinateCleaner (Zizka et al., 2019) helps in flagging common spatial errors in biological collection data, while plantR (de Lima et al., 2023) and bdc (Ribeiro et al., 2022) help to integrate and validate biodiversity data. rWCVP (Brown et al., 2023) and expowo (Zuanny et al., 2024) connect to the World Checklist of Vascular Plants and offer tools to standardize species names and integrate geospatial data. However, no single solution exists to check and enhance the quality of data sourced from online databases (Zizka et al., 2020). I emphasize that the goal of florabr is not to replace the functionalities of these packages but rather to offer an easy way to select target species, spatialize their distributions, and flag records based on the expert knowledge available in the Flora e Funga do Brasil.

I am aware of only one package that connects to the Flora e Funga do Brasil database in R, namely flora (Carvalho, 2020). This package also enables the collection of taxonomic and distribution information from the Flora e Funga do Brasil, but it operates based on a fixed version of the dataset. The differential of florabr relies on its capability to interact with the latest, or any other version of the Flora e Funga do Brasil dataset, which undergoes weekly updates. Moreover, florabr provides a single function for selecting target species based on multiple criteria, as well as the option to use distribution information to obtain a PAM and flag occurrence records based on expert knowledge. In this

regard, I encourage and welcome users' support to address issues and suggest new features.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

The version of the package described in this paper (version 1.2.0) is archived in Zenodo at https://doi.org/10.5281/zenodo.11188692. The florabr R package source code is hosted at https://github.com/wevertonbio/florabr/. The package is available on CRAN at https://cran.r-project.org/package=florabr. Extensive documentation on the usage of all functions is available at https://wevertonbio.github.io/florabr/.

ORCID

Weverton C. F. Trindade 🄳 http://orcid.org/0000-0003-2045-4555

REFERENCES

Brazil Flora Group. 2018. Brazilian Flora 2020: Innovation and collaboration to meet Target 1 of the Global Strategy for Plant Conservation (GSPC). *Rodriguésia* 69: 1513–1527.

Brazil Flora Group. 2021. Brazilian Flora 2020: Leveraging the power of a collaborative scientific network. *Taxon* 71: 178–198.

Brown, M. J. M., B. E. Walker, N. Black, R. H. A. Govaerts, I. Ondo, R. Turner, and E. Nic Lughadha. 2023. rWCVP: A companion R package for the World Checklist of Vascular Plants. *New Phytologist* 240: 1355–1365.

Brummitt, N., A. C. Araújo, and T. Harris. 2021. Areas of plant diversity— What do we know? *Plants, People, Planet* 3: 33–44.

Brummitt, R. K. 2001. World Geographic Scheme for Recording Plant Distributions. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Carvalho, G. 2020. *flora: Tools for Interacting with the Brazilian Flora 2020.* R package version 0.3.4. Website: https://CRAN.R-project.org/package=flora [accessed 26 July 2024].

Colli-Silva, M., M. Reginato, A. Cabral, R. C. Forzza, J. R. Pirani, and T.N. da C. Vasconcelos. 2020. Evaluating shortfalls and spatial accuracy of biodiversity documentation in the Atlantic Forest, the most diverse and threatened Brazilian phytogeographic domain. *Taxon* 69: 567–577.

Convention on Biological Diversity. 2012. The global strategy for plant conservation: 2011–2020. Botanic Gardens Conservation International, Richmond, United Kingdom.

Folk, R. A., and C. M. Siniscalchi. 2021. Biodiversity at the global scale: The synthesis continues. *American Journal of Botany* 108: 912–924.

Forzza, R. C., J. F. A. Baumgratz, C. E. M. Bicudo, D. A. L. Canhos, A. A. Carvalho, M. A. N. Coelho, A. F. Costa, et al. 2012. New Brazilian floristic list highlights conservation challenges. *BioScience* 62: 39–45.

GBIF. 2024. florabr R package: Records of plant species. https://doi.org/10.15468/DD.QPGEB7

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11: 156–162.

Kindt, R. 2020. WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. *Applications in Plant Sciences* 8: e11388.

de Lima, R. A. F., A. Sánchez-Tapia, S. R. Mortara, H. ter Steege, and M. F. de Siqueira. 2023. plantR: An R package and workflow for managing species records from biological collections. *Methods in Ecology and Evolution* 14: 332–339.

Maldonado, C., C. I. Molina, A. Zizka, C. Persson, C. M. Taylor, J. Albán, E. Chilquillo, et al. 2015. Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? *Global Ecology and Biogeography* 24: 973–984.

Nelson, G., and S. Ellis. 2019. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374: 2–10.

Peterson, A. T., A. Asase, D. A. L. Canhos, S. de Souza, and J. Wieczorek. 2018. Data leakage and loss in biodiversity informatics. *Biodiversity Data Journal* 6: e26826.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website http://www.R-project.org/ [accessed 26 July 2024].

Reginato, M., and F. A. Michelangeli. 2020. Bioregions of Eastern Brazil, based on vascular plant occurrence data. *In* V. Rull and A. C. Carnaval [eds.], Neotropical Diversification: Patterns and Processes, 475–494. Springer, Cham, Switzerland.

Ribeiro, B. R., S. J. E. Velazco, K. Guidoni-Martins, G. Tessarolo, L. Jardim, S. P. Bachman, and R. Loyola. 2022. bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution* 13: 1421–1428.

Soberon, J., and J. Cavner. 2015. Indices of biodiversity pattern based on presence-absence matrices: A GIS implementation. *Biodiversity Informatics* 10: 22–34. https://doi.org/10.17161/bi.v10i0.4801

Soberón, J., M. E. Cobos, and C. Nuñez-Penichet. 2021. Visualizing species richness and site similarity from presence-absence matrices. *Biodiversity Informatics* 16: 20–27.

Soltis, D. E., and P. S. Soltis. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity* 38: 264–270.

Tietje, M., A. Antonelli, W. J. Baker, R. Govaerts, S. A. Smith, and W. L. Eiserhardt. 2022. Global variation in diversification rate and species richness are unlinked in plants. *Proceedings of the National Academy of Sciences, USA* 119: e2120662119.

Whittaker, R. J., M. B. Araujo, P. Jepson, R. J. Ladle, J. E. M. Watson, and K. J. Willis. 2005. Conservation biogeography: Assessment and prospect. *Diversity and Distributions* 11: 3–23.

Zhang, J., and H. Qian. 2023. U.Taxonstand: An R package for standardizing scientific names of plants and animals. *Plant Diversity* 45: 1–5.

Zizka, A., D. Silvestro, T. Andermann, J. Azevedo, C. Duarte Ritter, D. Edler, H. Farooq, et al. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10: 744–751.

Zizka, A., F. A. Carvalho, A. Calvente, M. R. Baez-Lizarazo, A. Cabral, J. F. Ramos Coelho, M. Colli-Silva, et al. 2020. No one-size-fits-all solution to clean GBIF. *PeerJ* 8: e9916.

Zuanny, D. C., B. Vilela, P. W. Moonlight, T. E. Särkinen, and D. Cardoso. 2024. expowo: An R package for mining global plant diversity and distribution data. *Applications in Plant Sciences* 12: e11609.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Code to reproduce the examples provided in the paper.

Appendix S2. The table returned by the *get_attributes()* function in florabr.