

 Open access • Posted Content • DOI:10.1101/2021.04.06.437754

plantR: An R package and workflow for managing species records from biological collections — [Source link](#)

[Renato A. F. de Lima](#), [Renato A. F. de Lima](#), [Andrea Sánchez-Tapia](#), [Sara Ribeiro Mortara](#) ...+3 more authors

Institutions: [Naturalis](#), [University of São Paulo](#), [International Institute of Minnesota](#), [VU University Amsterdam](#)

Published on: 08 Apr 2021 - [bioRxiv](#) (Cold Spring Harbor Laboratory)

Topics: [Data editing](#), [Workflow](#), [Data retrieval](#) and [Data quality](#)

Related papers:

- [rMyCoPortal](#) - an R package to interface with the Mycology Collections Portal.
- [A workflow for standardising and integrating alien species distribution data](#)
- [Data storage and data re-use in taxonomy—the need for improved storage and accessibility of heterogeneous data](#)
- [Taxamat: Automated biodiversity data management tool - Implications for microbiome studies.](#)
- [The application of interactive multimedia software in taxonomy and biological diversity studies](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/plantr-an-r-package-and-workflow-for-managing-species-33popzoqf0>

plantR: An R package and workflow for managing species records from biological collections

Renato A. F. de Lima^{1,2} | Andrea Sánchez-Tapia³ | Sara R. Mortara^{3,4} | Hans ter Steege^{1,5} | Marinez F. de Siqueira³

¹Tropical Botany, Naturalis Biodiversity Center, Leiden, The Netherlands

²Departamento de Ecologia, Universidade de São Paulo, São Paulo, Brazil

³Instituto de Pesquisas Jardim Botânico do Rio de Janeiro, Rio de Janeiro, Brazil

⁴International Institute for Sustainability, Rio de Janeiro, Brazil

⁵Systems Ecology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence

Renato A. F. de Lima
Email: raflima@usp.br

Funding information

European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 795114; Coordination for the Improvement of Higher Education Personnel - CAPES (process 88887.145924/2017-00); Instituto Nacional da Mata Atlântica - INMA

Abstract

1. Species records from biological collections are becoming increasingly available online. This unprecedented availability of records has largely supported recent studies in taxonomy, biogeography, macroecology, and biodiversity conservation. Biological collections vary in their documentation and notation standards, which have changed through time. For different reasons, neither collections nor data repositories perform the editing, formatting, and standardization of the data, leaving these tasks to the final users of the species records (e.g. taxonomists, ecologists and conservationists). These tasks are challenging, particularly when working with millions of records from hundreds of biological collections.
2. To help collection curators and final users perform those tasks, we introduce `plantR`, an open-source package that provides a comprehensive toolbox to manage species records from biological collections. The package is accompanied by the proposal of a reproducible workflow to manage this type of data in taxonomy, ecology, and biodiversity conservation. It is implemented in R and designed to handle relatively large data sets as fast as possible. Initially designed to handle plant species records, many of the `plantR` features also apply to other groups of organisms, given that the data structure is similar.
3. The `plantR` workflow includes tools to (1) download records from different data repositories, (2) standardize typical fields associated with species records, (3) validate the locality, geographical coordinates, taxonomic nomenclature, and species identifications, including the retrieval of duplicates across collections, and (4) summarize and export records, including the construction of species checklists with vouchers.
4. Other R packages provide tools to tackle some of the workflow steps described above. But in addition to the new features and resources related to the data editing and validation, the greatest strength of `plantR` is to provide a comprehensive and user-friendly workflow in one single environment, performing all tasks from data retrieval to export. Thus, `plantR` can help researchers better assess data quality and avoid data leakage in a wide variety of studies using species records.

KEYWORDS

biodiversity, data cleaning, data download, duplicate records, gazetteer, GBIF, herbarium, taxonomic validation

1 | INTRODUCTION

Biological collections (e.g. museums and herbaria) are essential for studying biodiversity (Graham et al., 2004). Taxonomists use these collections to describe new species, produce taxonomic revisions and species checklists, among other important uses (Funk, 2003; Bebbler et al., 2010; Besnard et al., 2018). In macroecology, biogeography, and conservation, biological collections are often the main source of species records, which are used to study spatial patterns of biodiversity, species ecological niches, endemism levels, and conservation status (Graham et al., 2004; Dauby et al., 2017; Ulloa et al., 2017; Lima et al., 2020). Biological collections are increasingly making their electronic databases available in online databases, such as the Global Biodiversity Information Facility (GBIF). This growing availability of information has catalyzed many syntheses of our biodiversity knowledge (e.g. Antonelli et al. 2018), highlighting the importance of biological collections even more.

The increasing availability of biological collections databases has also exposed the wide variation of the documentation standards within and between collections (Willemse et al., 2008). Within collections, specimens collected by different people or in different periods may vary in their notation standards. The international documentation standards themselves are constantly evolving (www.tdwg.org/standards). Moreover, older records tend to have less associated information (e.g. missing geographical coordinates) and may contain names of localities that no longer exist (i.e. changing toponyms). Between collections, differences may emerge from different choices of documentation standards, on how to enter specimen information in the electronic databases, and on which fields should be entered first in the face of limited resources. The staff of biological collections often have little time to update the information that has been already entered in their databases or to correct data entry errors (e.g. typographical errors). These tasks become more challenging as the number of records in the collection increases.

Despite the global efforts to standardize the documentation of biodiversity information (e.g. Darwin Core

standards), there is still much variation within fields associated with species records. This variation is likely to remain for years to come because biological collections are often underfunded, undervalued, and understaffed (de Gasper et al., 2020). Online databases, such as GBIF, gather, store, flag, and check some but not all the information provided by the data providers. This means that, although highly valuable, the available databases from biological collections are not always ready for use (Peterson et al., 2018). So, the final users of species records (e.g. taxonomists, ecologists, and conservationists) often have to decide between performing those procedures themselves or trusting the data available without knowing exactly the level of data quality. This is problematic because variation in data quality can impact the outcomes of studies in taxonomy, ecology, and conservation (Graham et al., 2004; Zizka et al., 2019; Rodrigues et al., 2020). Thus, we still need comprehensive and reproducible tools to manage species records from biological collections, particularly regarding notation standards, species identifications, duplicate records, and fine-scale validation of the geographical coordinates.

2 | OVERVIEW

We present `plantR`, a new R package for managing species records from biological collections. As a general approach, `plantR` does not edit the original information; it stores the standardized information in new columns to assist collection curators in comparing original and edited information. Much of the new functionalities depend on gazetteers, maps, lists of taxonomists, and plant collections, which are provided with the package. As its name suggests, `plantR` was initially designed to manage plant records from herbaria, with some functionalities being currently exclusive to plants. However, if the input data has the required fields and data format, many `plantR` features should work for any group of organisms. `plantR` should interest taxonomists, biogeographers, ecologists, and conservationists, as well as curators of biological collections. The package is implemented in R (R Core Team, 2020) and details on

83 its implementation and functionalities can be found at
 84 <https://github.com/LimaRAF/plantR>.

85 3 | THE PLANTR WORKFLOW

86 `plantR` is accompanied by the proposal of a workflow to
 87 process the information associated with species records
 88 (Fig. 1). Here, we present the steps of this workflow and
 89 the main `plantR` features to apply it. They are presented
 90 in the order that the workflow should be applied. This
 91 order aims to maximize the edition and validation of the
 92 available information, although many `plantR` functional-
 93 ities work independently from the previous steps of the
 94 workflow.

95 3.1 | Data entry

96 Users can download species records directly from R,
 97 which is currently done from the Centro de Refer-
 98 ência em Informação Ambiental (CRIA, www.cria.org.br) and GBIF (www.gbif.org), using functions
 99 `rspecieslink()` and `rgbif2()`, respectively. The func-
 100 tion `rgbif2()` performs a search based on scientific
 101 names using the `rgbif` package, but with a standard-
 102 ized output to enter the `plantR` workflow. The func-
 103 tion `rspeciesLink()` is more flexible allowing the user
 104 to search by scientific name or any other taxonomic
 105 level, collection, and locality. Since these two sources of
 106 species records return different fields, a function is pro-
 107 vided to guarantee their correspondence with the DwC
 108 standards (function `formatDwc()`). Users can also load
 109 their own data, which can be converted to the Darwin
 110 Core (DwC) standards (<https://dwc.tdwg.org>) using
 111 the function `formatDwc()`. Alternatively, users can im-
 112 port data from zipped DwC-Archive files from a local
 113 directory or from a link for data download provided by
 114 GBIF (function `readData()`).
 115

116 3.2 | Data editing

117 Data standardization is particularly important when
 118 combining records from multiple collections, because

they not always follow the same documentation stan- 119
 dards. `plantR` provides tools to edit and standardize the 120
 notation of the information associated with the records, 121
 which are very important for validating locality informa- 122
 tion, assessing the confidence level of species identifica- 123
 tions and searching duplicate records across collections 124
 (see 3.3 Data validation). 125

126 3.2.1 | People's names and collection 127 information

128 The first edits performed by `plantR` regards the name 128
 of collector and identifiers, collector's number and 129
 collection year (function `formatOcc()`). By default, 130
 people's names are returned in the Biodiversity Infor- 131
 mation Standards format (www.tdwg.org/standards/hispid3/), which is: last name + comma + initials sep- 132
 arated by points (e.g. Gentry, A.H.). Name formatting 133
 takes into account generational suffixes (e.g. Junior), 134
 prepositions (e.g. da, dos, von), compound last names 135
 (e.g. Saint-Hilaire), some titles (e.g. Dr., Profa.) and 136
 multiple collector names. `plantR` also standardizes the 137
 collection codes using a database of over 5000 plant 138
 collection names and their respective Index Herbariorum 139
 or Index Xylariorum codes (function `getCode()`). 140
 141

142 3.2.2 | Locality and spatial information

143 One of the innovations of `plantR` is the standardiza- 143
 tion of records' locality information (i.e the DwC fields 144
 "country", "stateProvince", "municipality" and "locality"; 145
 function `formatLoc()`). For instance, names are trans- 146
 formed to English (e.g. Brasil or Brésil become Brazil) 147
 and their notation is standardized (e.g. BR or BRA be- 148
 come Brazil). In the case of missing locality information, 149
`plantR` performs some text mining aiming to retrieve 150
 them from other fields. To make sure that the original 151
 or retrieved locality information does exist, the package 152
 cross-checks the locality information of records with a 153
 gazetteer (function `getLoc()`). This cross-checking is 154
 based on a standard name-string that hierarchically com- 155
 bines the locality information at the best resolution avail- 156
 able, thus avoiding spurious matches of same locality 157

names in different countries or states/provinces (function `strLoc()`). The default `plantR` gazetteer currently contains entries at country level for all countries and at the lowest administrative level available at GDAM (<https://gadm.org>) for all Latin American countries and dependent territories (e.g. U.S. Virgin Islands). For Brazil, the gazetteer also contains information at the locality level (e.g. farms, forest fragments, parks). Most importantly, users can provide their regional or personal gazetteers.

The gazetteer includes some of the most common spelling variants and historical changes to locality names (currently biased for Brazil), which allows collection curators to trace back the most up-to-date locality names to improve their databases (function `getAdmin()`). Additionally, `plantR` assigns a geographical coordinate from the gazetteer to all valid localities (function `getCoord()`), which can be used as working coordinates in the case of missing or problematic original coordinates. Besides the automated assignment of missing coordinates, the package formats the original geographical coordinates to obtain non-zero, non-missing coordinates in decimal degrees (function `prepCoord()`).

3.2.3 | Taxonomic information

`plantR` offers tools to format scientific name notation, such as the isolation and removal of taxonomic rank (e.g. var., subsp.) and name modifiers (e.g. cf., aff.), which is important for records containing more raw taxonomic information (e.g. morpho-species, incomplete identifications). The package also standardizes the name of botanical families, using a list of valid family names and synonyms from the APG IV for angiosperms (Chase et al., 2016) and PPG I for lycophytes and ferns (Schuettpelez et al. 2016; function `formatFamily()`). If the family name is not found in the list, a search for a valid family name is performed based on the genus. Finally, the package can replace synonyms, orthographic variants and typographical errors in species names (function `formatSpecies()`, which is performed using functions from the packages `Taxonstand` (Cayuela et al., 2021) and `flora` (Carvalho, 2020). These packages per-

form exact and fuzzy name matching from The Plant List (www.theplantlist.org/) and the Brazilian Flora 2020 project (<http://floradobrasil.jbrj.gov.br/>), respectively.

3.3 | Data validation

3.3.1 | Locality and spatial information

`plantR` compares the precision of the original locality information with the one obtained by the cross-checking with a gazetteer (function `validateLoc()`). This comparison allows to flag possible typographical errors or unknown place names, which users can drop from the analyses or double-check themselves depending on their goals. Obtaining valid locality information is essential for the validation of geographical coordinates because they are validated by comparing the locality information of the record and the locality obtained by overlapping the coordinates with administrative maps (function `checkCoord()`). The package offers procedures for detecting the inversion and/or swap of coordinates (function `checkInverted()`), coordinates falling in the sea or bays, near the shoreline (`checkShore()`), and in neighbouring countries (`checkBorders()`). If after these procedures the locality information from the record and maps matches, the coordinate is flagged as validated, with an indication of the resolution of the validation (i.e. country, state, municipality or locality levels). As before, the validation of geographical coordinates is done using maps at the country level for the world and at the lowest administrative level available at GDAM for Latin America, but users can provide their own maps. Finally, `plantR` also provides tools to detect records from cultivated individuals (function `getCult()`) and spatial outliers (function `checkOut()`), i.e. coordinates too far away from the core distributions for a given taxon (Liu et al., 2018).

3.3.2 | Species identifications

One highlight of `plantR` is the classification of records according to the confidence in their species identifica-

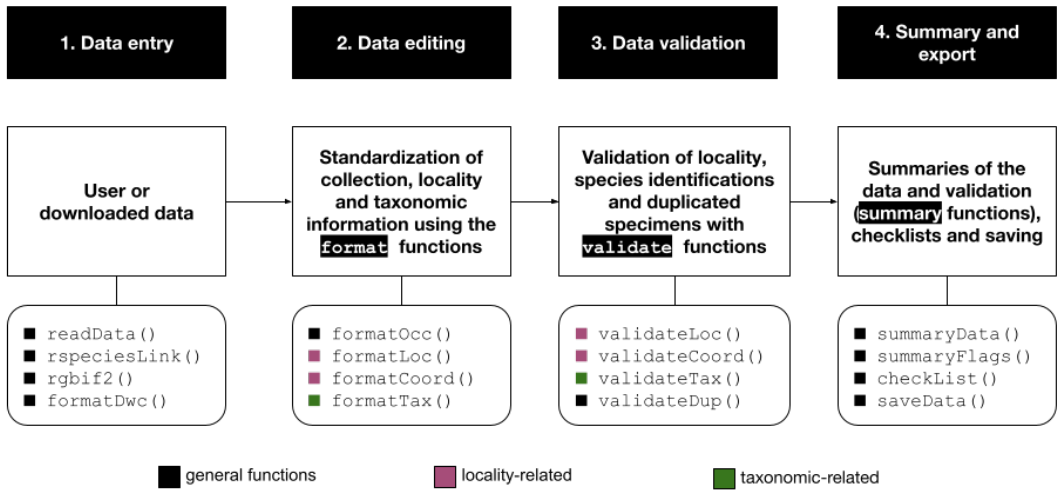


FIGURE 1 Chart illustrating the four main steps of the workflow proposed here to manage species records from biological collections for taxonomy, ecology, and biodiversity conservation. Black boxes represent each of the four steps, white boxes their description, and rounded boxes their main `plantR` functions.

237 tions (function `validateTax()`). This validation is based
 238 on a global list of ca. 8500 plant taxonomists names
 239 compiled from different sources (Lima et al., 2020). By
 240 default, this classification assigns the highest confidence
 241 level to three different cases: (i) type specimens (e.g. iso-
 242 types, holotypes), (ii) records identified by a specialist
 243 of the family, and (iii) records collected by the special-
 244 ist of the family but with the identifier field empty (case
 245 iii is optional). The confidence level of records without
 246 identifier information (including NA's) is flagged as 'un-
 247 known', while records identified by non-family special-
 248 ists it are flagged as 'low'. Users can provide their own
 249 list of taxonomists, as long as this list has the same gen-
 250 eral format as the default list provided by `plantR`. More-
 251 over, `validateTax()` returns the most frequent names
 252 of identifiers that are not in the taxonomist list, allowing
 253 users to provide missing taxonomist names.

254 3.3.3 | Duplicate records

255 Another novelty of `plantR` regards duplicates, i.e. sam-
 256 ples of the same specimen incorporated in two or more
 257 collections (function `validateDup()`). Sharing biologi-

258 cal material across collections is a common and encour- 258
 259 aged practice, and they can represent 25% or more of 259
 260 the records available for regional biotas (e.g. Lima et al., 260
 261 2020). The search for duplicates in `plantR` is executed 261
 262 by combining fields related to the taxonomy, collection 262
 263 and locality of the records (e.g., family + collector name 263
 264 + collector number + municipality). Because of the great 264
 265 variation in the notation and completeness of collec- 265
 266 tor's and localities names, the package allows the simulta- 266
 267 neous use of different combinations of these fields 267
 268 to search for duplicates (function `getDup()`). If two or 268
 269 more combinations are provided, the search of dupli- 269
 270 cates uses tools from network analysis to find both direct 270
 271 and indirect links between records. The retrieval 271
 272 of duplicates across collections performs well using rela- 272
 273 tively large data-sets (i.e. millions of records). How- 273
 274 ever, finding all existing duplicates requires that the 274
 275 databases of all collections are available and that all 275
 276 search fields are complete and filled in without typos 276
 277 using the same notation standards (or notations that 277
 278 `plantR` can standardize). This is rarely the case, so the 278
 279 list of duplicates returned should be considered incom- 279
 280 plete in many cases. 280

281 `plantR` provides not only tools to search for dupli-
 282 cates, but also to homogenize information within the
 283 groups of duplicates found, such as species, locality
 284 and/or spatial information (function `mergeDup()`). This
 285 homogenization allows retrieving the best information
 286 available within duplicates, which is particularly useful
 287 when collections vary in the number and completeness
 288 of the digitized fields. After this homogenization, users
 289 can choose to remove or not the duplicates from the
 290 data. See Lima et al. (2020) for more details on the
 291 search and merge of duplicates implemented here.

292 3.4 | Data summary and export

293 As a final step of the workflow, `plantR` can help users
 294 to summarize their data (e.g. number of occurrences,
 295 collections and species; function `summaryData()`) and
 296 the flags of the validation process (i.e. localities,
 297 coordinates, identifications and duplicates; function
 298 `summaryFlags()`). The package also provides species
 299 checklists with user-defined numbers of voucher speci-
 300 mens and the export of records by groups (e.g. families,
 301 countries, collections).

302 4 | IMPLEMENTATION

303 4.1 | Example of usage

304 The `plantR` workflow can be implemented using few
 305 command lines and wrapper functions (see Table 1 for
 306 details). Here, we provide a simple example using only
 307 one species. A detailed tutorial of the package is pro-
 308 vided at <https://github.com/LimaRAF/plantR>.

```
309
310 # Installing plantR
311 remotes::install_github("LimaRAF/plantR")
312 library("plantR")
313
314 # Data download
315 occs_splink <- rspeciesLink(species =
316                             "Euterpe edulis")
317 occs_gbif <- rgbif2(species =
318                     "Euterpe edulis")
```

```
occs <- formatDwc(splink_data = 319
                  occs_splink , 320
                  gbif_data = 321
                  occs_gbif) 322
323
```

324 # Data editing

```
occs <- formatOcc(occs) 325
occs <- formatLoc(occs) 326
occs <- formatCoord(occs) 327
occs <- formatTax(occs) 328
329
```

330 # Data validation

```
occs <- validateLoc(occs) 331
occs <- validateCoord(occs) 332
occs <- validateTax(occs) 333
occs <- validateDup(occs) 334
335
```

336 # Data summary

```
summs <- summaryData(occs) 337
flags <- summaryFlags(occs) 338
checklist <- checklist(occs) 339
```

340 4.2 | Dependencies on other packages

341 Some of `plantR`'s features depend on other R pack- 341
 342 ages (Table 1). Function `rgbif2()` uses package 342
 343 `rgbif` (Chamberlain et al., 2021) for downloading GBIF 343
 344 data. The management of strings, countries names, 344
 345 and spatial data use packages `stringr` (Wickham, 345
 346 2019), `countrycode` (Arel-Bundock et al., 2018), and 346
 347 `sf`, (Pebesma, 2018), respectively. As mentioned above, 347
 348 function `formatSpecies()` uses `Taxonstand` (Cayuela 348
 349 et al., 2021) and `flora` (Carvalho, 2020). The search 349
 350 of duplicates uses package `igraph` (Csardi and Nepusz, 350
 351 2006) to perform indirect string search. Finally, many 351
 352 functions use `data.table` (Dowle and Srinivasan, 2020), 352
 353 which provides fast table manipulation, reading and sav- 353
 354 ing. 354

TABLE 1 List of the main functions per type of information and per step of the proposed workflow. We also present the wrappers of the main functions for each step (if present) and the other R packages necessary to execute them.

Workflow step	Type of information	Main functions	Wrapper	Dependencies
1 - Data Entry	Species records	readData, rgbif2, rspeciesLink, formatDwc	-	rgbif, data.table
2 - Data Editing	Names, numbers, etc	prepName, colNumber, getYear, getCode	formatOcc	stringr
	Localities	fixLoc, strLoc, prepLoc, getLoc	formatLoc	countrycode, stringr
	Coordinates	prepCoord, getCoord	formatCoord	-
	Taxonomy	fixSpecies, prepSpecies, prepFamily	formatTax	flora, Taxonstand, data.table
3 - Data Validation	Localities	validateLoc	-	-
	Coordinates	checkCoord, checkBorders, checkShore, checkInverted, getCult, checkOut	validateCoord	sf, robustbase, data.table
	Species identification	validateTax	-	-
	Duplicate records	prepDup, getDup, mergeDup, rmDup	validateDup	data.table, igraph
4 - Summary and Export	Summaries	summaryData, summaryFlags, checklist	-	data.table, stringr
	Export	saveData	-	data.table

5 | DISCUSSION

5.1 | Comparison with other R packages

Other R packages already provide spelling and synonym checks of species names (Chamberlain and Szöcs 2013; Cayuela et al. 2021; Carvalho 2020; Kindt 2020), so there was no need to ‘reinvent the wheel’ and their functionalities were (or will be) integrated in `plantR`. `CoordinateCleaner` (Zizka et al., 2019) provides a great toolbox to work with geographical coordinates and we suggest this package for more advanced editing of geographical coordinates. The differential of `plantR` lies in providing both locality and coordinate validation, the

automatic retrieval of valid coordinates for missing or problematic coordinates, and the coordinate validation at the county level. However, because these validations depend on the package `gazetteer`, these innovations current apply only to Latin America. `plantR` also provides an approach to find cultivated specimens (i.e. `getCult()`), which is based on the fields ‘locality’ or ‘occurrenceRemarks’ and thus different from the approach used by `CoordinateCleaner`.

We found only one package that validates the species identifications, `naturaList` (Rodrigues et al., 2020), which also uses the field ‘identifiedBy’, but classifies the confidence level of records other than preserved speci-

380 mens (vouchers) and require a user-provided list of tax-
 381 onomists. The differential of `plantR` relies on the pro-
 382 vision of a large database of plant taxonomists, besides
 383 the possibility of the user providing an extra list of spe-
 384 cialist names. In addition, `plantR` also relies on the field
 385 ‘`typeStatus`’ and it performs the validation at the family-
 386 level. We are not aware of other R packages that per-
 387 form (i) the edition of people names, (ii) the validation of
 388 locality information and (iii) the search/merge of dupli-
 389 cates.

390 5.2 | Limitations and future 391 developments

392 The variation in the notation of names, numbers and
 393 dates associated with species records across biological
 394 collections is huge; `plantR` handles most but not all of
 395 them. We envisage having a dictionary of common col-
 396 lectors’ names, but today some double-checking is still
 397 necessary. As mentioned before, locality and county-
 398 level geographical validation are currently biased to-
 399 wards Latin America. Therefore, users must be aware
 400 that the package does not provide solutions to all prob-
 401 lems related to species records information. Some im-
 402 provements predicted to be implemented in the future
 403 include the download from other data repositories (e.g.
 404 JABOT, <http://jabot.jbrj.gov.br>), the expansion of the
 405 package gazetteer and county-level maps and the val-
 406 idation of species names against databases that have
 407 wider geographical and taxonomic coverage (e.g. ITIS,
 408 <https://itis.gov/>). We also plan to include simple
 409 functions that prepare records to enter the workflow of
 410 other R packages (e.g. `modleR` or `ConR` - Sánchez-Tapia
 411 et al. 2020; Dauby et al. 2017), that facilitate the citation
 412 of collections (e.g. `occCite` - Owens et al. 2021) and
 413 that collect provenance (e.g. `rdt` - Lerner et al. 2018).
 414 Moreover, the gazetteer, list of taxonomists, maps, and
 415 collections are constantly being improved; we are happy
 416 to receive and incorporate missing or more regional in-
 417 formation to make them more complete.

6 | CONCLUDING REMARKS

418
 419 The number of collection databases made available on-
 420 line has greatly increased in the last decades and will
 421 probably continue to increase in the years to come (Gra-
 422 ham et al., 2004; Sweeney et al., 2018). Therefore, hav-
 423 ing tools to assess and improve the quality of the in-
 424 formation associated with species record is a pressing
 425 issue in biodiversity research. `plantR` provides these
 426 tools, some of them being presented for the first time.
 427 Although there are packages that provide similar tools,
 428 the greatest strength of `plantR` is to provide a compre-
 429 hensive toolbox and a user-friendly workflow to pro-
 430 cess species records from beginning to end within a sin-
 431 gle environment. Thus, we expect that `plantR` can im-
 432 prove the reproducibility of taxonomic, ecological and
 433 conservation studies. But more importantly, we hope
 434 that `plantR` can assist collection curators to flag pos-
 435 sible issues that need attention, thus saving their time
 436 while conducting the important task of maintaining bio-
 437 logical collections.

ACKNOWLEDGEMENTS

438
 439 This package was supported by the European Union’s
 440 Horizon 2020 research and innovation program un-
 441 der the Marie Skłodowska-Curie grant agreement No
 442 795114. M.F.S., A.S.-T. and S.R.M. were supported by
 443 the Coordination for the Improvement of Higher Educa-
 444 tion Personnel - CAPES (process 88887.145924/2017-
 445 00), the PNPD/CAPES program and the PCI program
 446 of the ‘Instituto Nacional da Mata Atlântica’ (INMA), re-
 447 spectively. We thank Sidnei Souza from CRIA for his
 448 help with the web API. We also thank CNCFlora and
 449 the TreeCo database for providing localities used to con-
 450 struct the gazetteer, and Vinicius C. Souza (ESALQ/USP)
 451 who helped to curate the list of plant taxonomists.

AUTHORS’ CONTRIBUTIONS

452
 453 R.A.F.L. conceived the idea and R.A.F.L., A.S.-T., S.R.M.
 454 and M.F.S. designed methodology. R.A.F.L. constructed
 455 the list of taxonomists, collections, and families, while

456 R.A.F.L., A.S.-T., S.R.M. constructed the gazetteer and
 457 maps. R.A.F.L., A.S.-T., S.R.M. and H.t.S. wrote the codes
 458 and package documentations. R.A.F.L. led the writing of
 459 the manuscript, with contributions from A.S.-T. All au-
 460 thors contributed critically to the manuscript and gave
 461 final approval for publication.

462 DATA AVAILABILITY STATEMENT

463 The R package plantR is available at <https://github.com/LimaRAF/plantR>. The version of the package described in this paper (version 0.1.1) is archived at [link to be included].

467 references

468 Antonelli, A., Zizka, A., Carvalho, F. A., Scharn, R., Bacon,
 469 C. D., Silvestro, D. and Condamine, F. L. (2018) Amazonia is the primary source of Neotropical biodiversity. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 6034–6039.

473 Arel-Bundock, V., Enevoldsen, N. and Yetman, C. (2018)
 474 countrycode: An R package to convert country names
 475 and country codes. *Journal of Open Source Software*, **3**,
 476 848. URL: <http://joss.theoj.org/papers/10.21105/joss.00848>.

478 Bebbler, D. P., Carine, M. A., Wood, J. R. I., Wortley, A. H.,
 479 Harris, D. J., Prance, G. T., Davidse, G., Paige, J., Pen-
 480 ington, T. D., Robson, N. K. B. and Scotland, R. W.
 481 (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences*, **107**, 22169–22171. URL: <http://www.pnas.org/cgi/doi/10.1073/pnas.1011841108>.

485 Besnard, G., Gaudeul, M., Lavergne, S., Muller, S., Rouhan,
 486 G., Sukhorukov, A. P., Vanderpoorten, A. and Jabbour,
 487 F. (2018) Herbarium-based science in the twenty-first
 488 century. *Botany Letters*, **165**, 323–327. URL: <https://doi.org/10.1080/23818107.2018.1482783>.

490 Carvalho, G. (2020) flora: Tools for Interacting with the
 491 Brazilian Flora 2020. *R package version 0.3.4*. URL:
 492 <https://cran.r-project.org/package=flora>.

493 Cayuela, L., Stein, A. and Oksanen, J. (2021) Taxonstand:
 494 Taxonomic Standardization of Plant Species Names. *R*
 495 *package version 2.3*. URL: <https://cran.r-project.org/package=Taxonstand>.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet,
 P., Geffert, L. and Ram, K. (2021) rgbif: Interface to
 the Global Biodiversity Information Facility API. *R pack-*
age version 3.5.2. URL: <https://cran.r-project.org/package=rgbif>.

Chamberlain, S. A. and Szöcs, E. (2013) taxize: taxonomic
 search and retrieval in R. *F1000Research*, **2**, 191. URL:
<https://f1000research.com/articles/2-191/v1>.

Chase, M. W., Christenhusz, M. J., Fay, M. F., Byng, J. W.,
 Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N.,
 Soltis, P. S., Stevens, P. F., Briggs, B., Brockington, S.,
 Chautems, A., Clark, J. C., Conran, J., Haston, E., Möller,
 M., Moore, M., Olmstead, R., Perret, M. et al. (2016) An
 update of the Angiosperm Phylogeny Group classifica-
 tion for the orders and families of flowering plants: APG
 IV. *Botanical Journal of the Linnean Society*, **181**, 1–20.
 URL: <https://academic.oup.com/botlinnean/article-lookup/doi/10.1111/boj.12385>.

Csardi, G. and Nepusz, T. (2006) The igraph software pack-
 age for complex network research. *InterJournal Complex*
Systems, 1695. URL: <https://igraph.org>.

Dauby, G., Stévant, T., Droissart, V., Cosiaux, A., Deblauwe,
 V., Simo-Droissart, M., Sosef, M. S., Lowry, P. P., Schatz,
 G. E., Gereau, R. E. and Couvreur, T. L. (2017) ConR: An
 R package to assist large-scale multispecies preliminary
 conservation assessments using distribution data. *Ecology and Evolution*, **7**, 11292–11303.

Dowle, M. and Srinivasan, A. (2020) data.table: Extension
 of 'data.frame'. *R Package Version 1.13.6*. URL: <https://cran.r-project.org/package=data.table>.

Funk, V. (2003) The Importance of Herbaria. *Plant Science Bulletin*, **49**, 94–95.

de Gasper, A. L., Stehmann, J. R., Roque, N., Bigio, N. C.,
 Sartori, Â. L. B. and Gritt, G. S. (2020) Brazilian herbaria:
 An overview. *Acta Botanica Brasilica*, **34**, 352–359.

Graham, C., Ferrier, S., Huettman, F., Moritz, C. and Peter-
 son, A. (2004) New developments in museum-based
 informatics and applications in biodiversity analy-
 sis. *Trends in Ecology Evolution*, **19**, 497–503. URL:
<https://linkinghub.elsevier.com/retrieve/pii/S0169534704002034>.

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**. URL: <https://onlinelibrary.wiley.com/doi/10.1002/aps3.11388>.

- 543 Lerner, B., Boose, E. and Perez, L. (2018) Using Introspection
544 to Collect Provenance in R. *Informatics*, **5**. URL: <http://www.mdpi.com/2227-9709/5/1/12>. 588
- 545 589 590
- 546 Lima, R. A. F., Souza, V. C., de Siqueira, M. F. and ter Steege,
547 H. (2020) Defining endemism levels for biodiversity con-
548 servation: Tree species in the Atlantic Forest hotspot. *Bi-*
549 *ological Conservation*, **252**, 108825. URL: <https://doi.org/10.1016/j.biocon.2020.108825>. 591
- 550 592 593 594 595 596 597 598 599
- 551 Liu, C., White, M. and Newell, G. (2018) Detecting outliers
552 in species distribution data. *Journal of Biogeography*, **45**,
553 164–176. 600
- 554 Owens, H. L., Merow, C., Maitner, B., Kass, J. M., Barve, V.
555 and Guralnick, R. P. (2021) occCite: Querying and Man-
556 aging Large Biodiversity Occurrence Datasets. *R pack-*
557 *age version 0.4.6*. URL: [https://cran.r-project.org/](https://cran.r-project.org/package=occCite)
558 [package=occCite](https://cran.r-project.org/package=occCite). 601 602
- 559 Pebesma, E. (2018) Simple Features for R: Standardized
560 Support for Spatial Vector Data. *The R Journal*, **10**,
561 439. URL: [https://journal.r-project.org/archive/](https://journal.r-project.org/archive/2018/RJ-2018-009/index.html)
562 [2018/RJ-2018-009/index.html](https://journal.r-project.org/archive/2018/RJ-2018-009/index.html). 603
- 563 Peterson, A. T., Asase, A., Canhos, D., de Souza, S. and Wic-
564 zorek, J. (2018) Data Leakage and Loss in Biodiversity
565 Informatics. *Biodiversity Data Journal*, **6**, e26826. URL:
566 <https://bdj.pensoft.net/articles.php?id=26826>. 604 605
- 567 R Core Team (2020) R: A language and environment for sta-
568 tistical computing. *R Foundation for Statistical Computing*,
569 *Vienna, Austria*. URL: <https://www.r-project.org>. 606
- 570 Rodrigues, A. V., Nakamura, G. and Duarte, L. (2020) natu-
571 ralList : a package to classify occurrence records in lev-
572 els of confidence in species identification. *bioRxiv*, 1–17.
573 URL: <https://doi.org/10.1101/2020.05.26.115220>. 607 608 609 610 611 612 613
- 574 Sánchez-Tapia, A., Mortara, S. R., Bezerra Rocha, D. S.,
575 Mendes Barros, F. S., Gall, G. and de Siqueira, M. F.
576 (2020) modler: a modular workflow to perform ecolog-
577 ical niche modeling in R. *bioRxiv*, 1–25. 606
- 578 Schuettpelz, E., Schneider, H., Smith, A. R., Hovenkamp, P.,
579 Prado, J., Rouhan, G., Salino, A., Sundue, M., Almeida,
580 T. E., Parris, B., Sessa, E. B., Field, A. R., de Gasper, A. L.,
581 Rothfels, C. J., Windham, M. D., Lehnert, M., Dauphin,
582 B., Ebihara, A., Lehtonen, S., Schwartsburd, P. B. et al.
583 (2016) A community-derived classification for extant ly-
584 cophytes and ferns. *Journal of Systematics and Evolution*,
585 **54**, 563–603. 606 607 608 609 610 611 612 613
- 586 Sweeney, P. W., Starly, B., Morris, P. J., Xu, Y., Jones, A.,
587 Radhakrishnan, S., Grassa, C. J. and Davis, C. C. (2018) 600
- Large-scale digitization of herbarium specimens: Devel-
opment and usage of an automated, high-throughput
conveyor system. *Taxon*, **67**, 165–178. 588 589 590
- Ulloa, C. U., Acevedo-Rodríguez, P., Beck, S., Belgrano,
M. J., Bernal, R., Berry, P. E., Brako, L., Celis, M.,
Davidse, G., Forzza, R. C., Gradstein, S. R., Hokche,
O., León, B., León-Yáñez, S., Magill, R. E., Neill, D. A.,
Nee, M., Raven, P. H., Stimmel, H., Strong, M. T.
et al. (2017) An integrated assessment of the vascular
plant species of the Americas. *Science*, **358**, 1614–
1617. URL: [http://www.sciencemag.org/lookup/doi/](http://www.sciencemag.org/lookup/doi/10.1126/science.aao0398)
10.1126/science.aao0398. 591 592 593 594 595 596 597 598 599
- Wickham, H. (2019) stringr: Simple, Consistent Wrappers
for Common String Operations. *R package version 1.4.0*.
URL: <https://cran.r-project.org/package=stringr>. 600 601 602
- Willemse, L. P., Van Welzen, P. C. and Mols, J. B. (2008)
Standardisation in data-entry across databases: Avoid-
ing Babylonian confusion. *Taxon*, **57**, 343–345. 603 604 605
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte
Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza,
M., Scharn, R., Svantesson, S., Wengström, N., Zizka,
V. and Antonelli, A. (2019) CoordinateCleaner: Stan-
dardized cleaning of occurrence records from biolog-
ical collection databases. *Methods in Ecology and Evolu-*
tion, **10**, 744–751. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13152>. 606 607 608 609 610 611 612 613