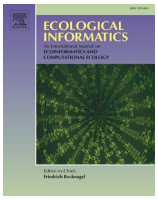
Ecological Informatics 93 (2026) 103543

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Bridging biodiversity gaps: Assessing R tools for harmonising vascular plant records

Diletta Santovito [a,*], Alessandro Chiarucci [a], Duccio Rocchini [a,b], Francesco Santi [a], Rocìo Beatriz Cortès Lobos [a], Riccardo Testolin [a]

[a] BIOME Lab, Department of Biological, Geological, and Environmental Sciences, Alma Mater Studiorum University of Bologna, Via Irnerio 42, 40126 Bologna, Italy
[b] Czech University of Life Sciences Prague, Faculty of Environmental Sciences, Department of Spatial Sciences, Kamýcka 129, Praha, Suchdol 16500, Czech Republic

ARTICLE INFO

Keywords:
Taxonomic harmonisation
Biodiversity databases
Biodiversity data cleaning
R
GBIF
BIEN

ABSTRACT

Biodiversity databases provide unprecedented opportunities for the use of species occurrence data for the development of large scale biodiversity analyses. However, these records often contain taxonomic uncertainties that can ultimately affect the outcomes of downstream analyses. Although several tools have been developed to address these issues, there is limited guidance on how to efficiently use and integrate them.

Here, we present a reproducible workflow for handling vascular plant occurrence data, and provide the first comparative analysis of R packages for the taxonomic harmonisation of vascular plant names. Our goal is to assess the differences in performance across the tested tools and to highlight best practices for leveraging large biodiversity databases.

We first downloaded occurrence data for vascular plants in Italy from the Botanical Information and Ecology Network (BIEN) and Global Biodiversity Information Facility (GBIF). We then compared seven R packages for taxonomic harmonisation, evaluating their ability to resolve names to accepted taxa and their overall performance.

Our results highlight heterogeneity in the number of names resolved by the different tools, with packages relying on plant-specific databases and implementing fuzzy matching outperforming those based on generalist databases and with no possibility of fuzzy matching. These findings underscore that the choice of both packages and taxonomic authorities can have a strong influence on data cleaning outcomes.

1. Introduction

Rapid global changes caused by the impact of human activities are leading to an unprecedented loss of biodiversity (Keesing et al., 2010; Pereira et al., 2012). According to the Intergovernmental Panel on Climate Change (IPCC, 2023) every region of the world is expected to see further rises in climatic hazards in the near future, raising numerous threats to ecosystems and people in the upcoming decades (Guo et al., 2023). It is therefore crucial to improve our understanding of the global patterns of biodiversity and how they change in time and space (Navarro et al., 2017).

In the past three decades, there has been a significant increase in the availability of large-scale biodiversity data, often referred to as biodiversity 'Big Data' (Musvuugwa et al., 2021), defined as 'the intensive data accumulation of digitized information on biodiversity, corresponding to a spatial and temporal description of species distribution', and as a 'techno-political tool to manage the distribution of biological species' (Devictor and Bensaude-Vincent, 2016).

Global databases have played a pivotal role in the aggregation and dissemination of biodiversity Big Data. Among them, the Global Biodiversity Information Facility (GBIF) is arguably the largest, now counting over three billion occurrence records distributed worldwide and spanning practically the entire spectrum of taxa (GBIF, 2025). Other open-source data aggregators focus on specific taxa, like the Botanical Information and Ecology Network (BIEN), containing only plant observations (BIEN, 2025).

The emergence of global-scale biodiversity databases has provided new opportunities for the use of occurrence data for the development of large scale analyses, such as species distribution models (Jetz et al., 2012), a powerful tool for guiding evidence-based conservation decisions (Guisan et al., 2013). However, leveraging this large amount of data is often a challenge, as each database has its limitations stemming

* Corresponding author.
E-mail address: diletta.santovito2@unibo.it (D. Santovito).

https://doi.org/10.1016/j.ecoinf.2025.103543
Received 1 September 2025; Received in revised form 27 November 2025; Accepted 28 November 2025
Available online 2 December 2025

from the heterogeneous nature of its sources (*e.g.,* historical museum specimens, researchers and citizen science data), which may also result in varying data quality (Zizka et al., 2019) and the presence of potential biases (Proença et al., 2017). Specifically, point-occurrence data often contain taxonomic, spatial, and temporal uncertainties (Meyer et al., 2016) which can affect the accuracy of downstream analyses (Gueta and Carmel, 2016). Thus, these data need to be harmonized before employing them to train species distribution models (Isaac et al., 2020).

As for taxonomic uncertainties, these are mainly due to the dynamic nature of taxonomic classification which, assisted by the increasing availability of data and the widespread use of genetic techniques, results in frequent nomenclature revisions (Grenié et al., 2023). Taxonomic names are also prone to human error, *e.g.*, during the digitization of handwritten labels or ledgers, which can lead to misspellings (Zermoglio et al., 2016).

When large amounts of data are involved, their manual handling is not an option for time-related reasons and the underlying risk of introducing additional errors. Moreover, manual approaches are inherently subjective and prevent reproducibility. To tackle these issues, several tools have been specifically built for the harmonisation of biodiversity data. Many of these tools have been developed for R, being the most widely used programming language by ecologists (Lai et al., 2019). A wide range of R packages have been developed to assist and facilitate taxonomic harmonisation following a common workflow (Box 1). These packages usually allow the user to match a list of species against a taxonomic database, *i.e.* a database containing a collection of information about taxa, to correct and standardise input names. Some of these databases have a great taxonomic breadth, such as the Catalogue of Life (COL), the GBIF Backbone Taxonomy, or the Integrated Taxonomic Information System (ITIS), while others are specific to certain taxa, *e.g.* the Leipzig Catalogue of Vascular Plants (LCVP) or the World Checklist of Vascular Plants (WCVP), specific to vascular plants. Grenié et al. (2023) reviewed most of R packages for taxonomic harmonisation, also providing some recommendations for users, database managers and package developers. Yet guidance on how to efficiently use and integrate these packages is still lacking, highlighting the need for a comparative analysis of their outputs.

To address this gap, we present a reproducible workflow for biodiversity data cleaning and conducted a comparative analysis of existing R packages for the taxonomic harmonisation of vascular plant species, using Italy as a test case. Italy is part of the Mediterranean biodiversity hotspot (Myers et al., 2000) and has one of the richest biodiversity heritages in Europe (ISPRA, 2009). This is mainly due to its range of biogeographic regions (Alpine, Continental, and Mediterranean), providing a highly heterogeneous climate, topography, and geology (ISPRA, 2009). Italy hosts more than 9000 plant taxa, of which more than 7500 are native and more than 1500 are endemic (Stinca et al., 2021). Additionally, the study area was selected to align with the strategic goals of the National Biodiversity Future Center, the first Italian research centre dedicated to biodiversity, within which this work is framed.

By evaluating the performance of taxonomic harmonisation packages on point-occurrence records from public biodiversity databases, we aim to provide practical insights into their relative strengths and limitations. Additionally, we propose a set of best practices for handling large biodiversity datasets, promoting standardized and reproducible approaches in biodiversity data cleaning.

2. Materials and methods

2.1. Data gathering and integration

We retrieved all the vascular plant occurrence data for Italy from the Botanical Information and Ecology Network (BIEN, https://bien.nceas.ucsb.edu/bien/) and the Global Biodiversity Information Facility (GBIF, 2025; https://www.gbif.org/), respectively using *BIEN* (Maitner et al., 2018; Maitner, 2023) and *rgbif* (Chamberlain et al., 2024) R packages.

We filtered georeferenced tracheophyte occurrences recorded after 1950 in Italy, for a total of 44,841 observations from BIEN and 1,243,640 from GBIF (Fig. 1). For the GBIF data, only occurrences with a basis of record of "OBSERVATION" and "HUMAN OBSERVATION" were kept, leading to the exclusion of living specimens, preserved specimens, fossil specimens, material citations and machine observations.

After downloading the data, we assigned each observation a unique identifier and merged the BIEN and GBIF datasets, only keeping the columns containing the scientific names, the date of collection, the coordinates and the unique identifiers, for a total of 1,288,481 occurrences.

Box 1
Summary of the main steps of the taxonomic harmonisation process.

1. Download point-occurrence data
Point-occurrence data can be retrieved from public databases, such as the Global Biodiversity Information Facility (GBIF) or the Botanical Information and Ecology Network (BIEN; specific to plants). Some of these databases can be directly queried in R using dedicated packages such as *rgbif* and *BIEN*.

2. Match the names against a taxonomic database
R packages for taxonomic harmonisation allow the user to match a list of scientific names against a taxonomic database. The matching process consists of two steps:

1. **Exact matching**: The species names are exactly matched with the corresponding names present in the taxonomic database, meaning that the output name must be identical to the input name.
2. **Fuzzy matching (optional)**: The remaining unmatched entries are matched to existing names based on their differences, for example by finding the name with the minimum number of single-character edits (insertions, deletions, or substitutions), since these discrepancies could result from typos or misspellings.

3. Retrieve the accepted names
Matched names are either synonyms or accepted names. In botany, a name is considered accepted if it is the legitimate and validly published name selected as the correct one for a taxon under the rules of the International Code of Nomenclature for algae, fungi, and plants (Turland et al., 2018). So the last step is to keep the accepted names associated with the matched names.

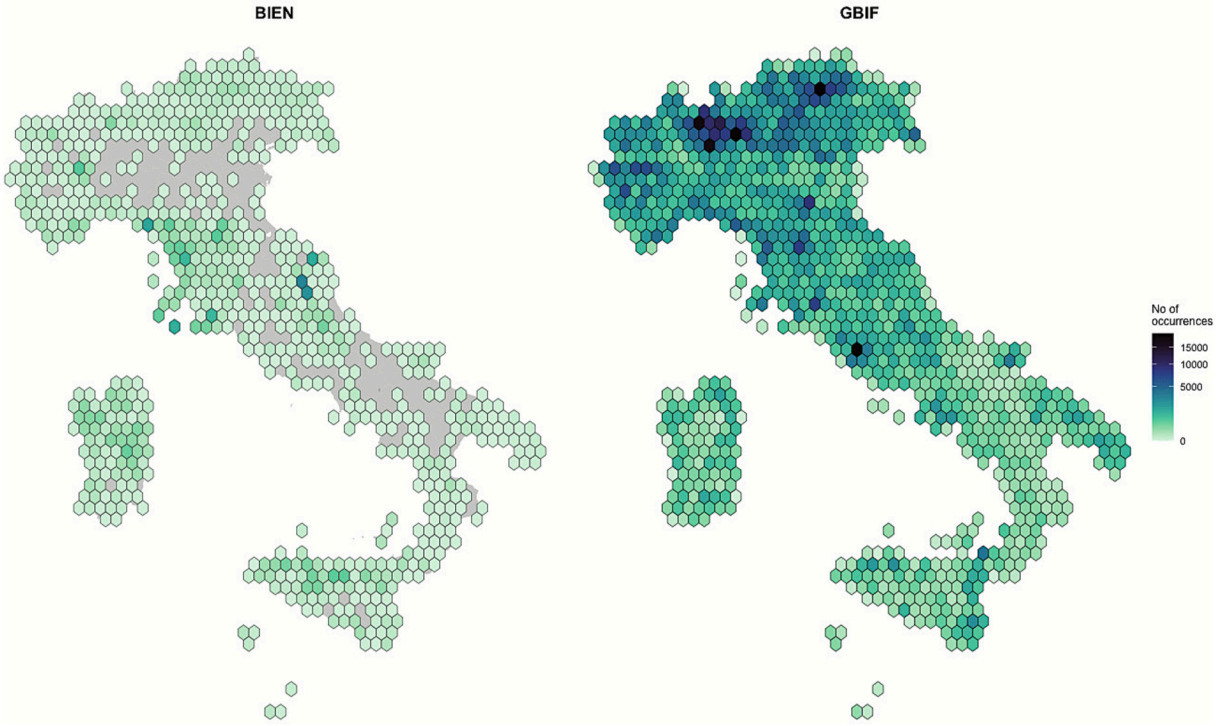


Fig. 1. Distribution of vascular plant point-occurrence data downloaded from BIEN (left) and GBIF (right) for the Italian territory. Values were square-root transformed to better highlight the differences in the spatial distribution of data.

2.2. Taxonomic harmonisation

Building on the review by Grenié et al. (2023), we selected R packages suitable for the harmonisation of vascular plant taxonomy. All selected packages were identified in that review, with the exception of *rWCVP* (Brown et al., 2023), which was developed afterwards. The final set of packages comprised: *lcvplants* (Freiberg et al., 2020), *rWCVP*, *taxadb* (Norman et al., 2020a, 2020b), *taxize* (Chamberlain and Szocs, 2013), *taxonomyCleanr* (Smith, 2025), *taxizedb* (Chamberlain et al., 2025), *TNRS* (Maitner and Boyle, 2024), and *WorldFlora* (Kindt, 2020). These packages were used to harmonise taxonomic names by correcting possible misspellings and updating them to the most recently accepted names. The execution time of the core functions of each package was measured using package *tictoc* (Izrailev, 2024).

2.2.1. lcvplants

Package *lcvplants* allows the user to standardise vascular plant names according to LCVP, which can be accessed through *LCVP* (Freiberg et al., 2020) data package. The package core function is *lcvp_search()* which first tries to find an exact match for the input names in the LCVP data. When it fails to do so, the package proceeds to perform a fuzzy match. To do so, we kept the default value of a maximum generalized Levenshtein distance of 0.2, meaning that given a name with length 10 only two changes (either an insertion, deletion, or substitution) are allowed.

2.2.2. rWCVP

Package rWCVP accesses the plant name and distribution data from WCVP *via* the *rWCVPdata* (Govaerts, 2024) data package. Core function *wcvp_match_names()* performs both exact and fuzzy matching. Results were filtered excluding matches with a match similarity lower than 0.8, *i.e.* 80 % of the characters.

2.2.3. taxadb

Package *taxadb* creates a local database of various taxonomic authorities and provides functions for the taxonomic harmonisation of scientific names against these databases. Among the available taxonomic authorities, we chose to use ITIS, COL and GBIF. Function *get_ids()* allows the user to retrieve the taxonomic identifiers associated with input names, while function *get_names()* converts the identifiers in accepted names. This package does not allow fuzzy matching.

2.2.4. taxonomyCleanr and taxize

Package *taxonomyCleanr* contains a set of functions for manually correcting misspelled taxa and resolving taxa to different authorities. It is based on *taxize,* another package developed for helping the users resolve taxonomic names, required for running *taxonomyCleanr*.

taxonomyCleanr does not include functions for automated fuzzy matching but only allows performing exact matching against supported taxonomic authorities *via* the function resolve_sci_taxa(). For this study, we selected the following authorities: ITIS and GBIF.

Package *taxize*, instead, was excluded from the analysis, because it does not include an automated way to retrieve the accepted names associated with the matched names.

2.2.5. taxizedb

Package *taxizedb* works similarly to *taxadb*, providing tools for working with taxonomic databases on the user's machine. Among the supported taxonomic authorities, we selected ITIS and GBIF. Species names can be converted in taxonomic identifiers *via* the function name2taxid() and the identifiers can then be conferred in accepted species names using the function taxid2name(). The package does not include functions for fuzzy matching.

2.2.6. TNRS

Package *TNRS* provides access to the Taxonomic Name Resolution Service API, an automated tool for the standardization of plant scientific names. Core function TNRS() resolves taxonomic names using various taxonomic sources. Among these, we selected WCVP and World Flora Online. The function also allows fuzzy matching. We selected only species with a match similarity greater than or equal to 0.8.

2.2.7. WorldFlora

Package *WorldFlora* provides methods for matching plant names against a static copy of the World Flora Online Taxonomic Backbone data. For this study we used version v.2024.06. Core function WFO.match() is used for matching plant names and performs both exact and fuzzy matching. For this purpose, we specified a maximum fuzzy distance of 2 characters.

2.2.8. Packages comparison

The species retrieved from occurrence data were provided as the input to the abovementioned package functions to compare the respective matches obtained using different taxonomic authorities. The overall output consistency was compared using non-metric multidimensional scaling (NMDS) based on the Jaccard similarity index to visually inspect differences between different packages and taxonomic backbones. To do so, we used package *vegan* (Oksanen et al., 2024) Intersections among outputs, *i.e.* the species names that were successfully resolved to an accepted name by two or more packages, were visualized using an UpSet plot *via* package *UpSetR* (Gehlenborg, 2019). The UpSet plot allows the visualisation of multiple sets of intersections, as an alternative to the Venn diagram when dealing with a large number of sets.

All the analyses were carried out in R version 4.4.2 (R Core Team, 2024). In addition to those already mentioned, we also used the following packages: *beepr* (Bååth, 2024), for sound notification after completing long processes; *ggpubr* (Kassambara, 2023), for arranging multiple ggplot2 objects on the same page; *ggrepel* (Slowikowski, 2024), to avoid overlapping text labels in plots; *lubridate* (Grolemund and Wickham, 2011), for date conversions; *mapdata* (Becker and Wilks, 2022), to download the base map of Italy; *sf* (Pebesma and Bivand, 2023), to convert the downloaded occurrences into spatial objects; *styler* (Müller and Walthert, 2024), to format the scripts according to the tidyverse style; *tidyverse* (Wickham et al., 2019), for data manipulation, visualisation, and analysis; *viridis* (Garnier et al., 2024), for colorblind-friendly color palettes for data visualisation.

3. Results

The final dataset contained 1,296,559 occurrences. From it, we extracted the unique species names, for a total of 6117 species names.

The packages with the highest number of resolved taxon names were *TNRS* (data provider = WCVP), *lcvplants* and *WorldFlora*, respectively matching to a proportion of 99.8 %, 99.5 % and 99.1 % accepted names, while the worst-performing packages were *taxonomyCleanr* (data provider = ITIS), *taxadb* (data provider = ITIS) and *taxizedb* (data provider = ITIS), respectively matching a proportion of 4.7, 48.7 % and 49.6 % accepted names (Table 1). There was also much heterogeneity in terms of timings. The fastest packages were *taxizedb, taxadb,* and *rWCVP,* taking less than one minute to match species names, while the slowest was WorldFlora, with an execution time of more than five hours.

Table 1
Number and percentage of species matched to an accepted name (6117 input species) for each combination of package taxonomic backbone, and of the execution timings.

Package	Accepted names	Accepted names %	Execution time (seconds)
TNRS (WCVP)	6106	99.82	486.676
lcvplants	6089	99.54	72.963
WorldFlora	6067	99.18	19,031.553
taxonomyCleanr (GBIF)	6049	98.89	3721.828
TNRS (WFO)	5941	97.12	466.413
taxadb (COL)	5931	96.96	24.086
taxadb (GBIF)	5824	95.21	34.946
rWCVP	5750	94.00	51.269
taxizedb (GBIF)	5393	88.16	0.197
taxizedb (ITIS)	3035	49.62	0.234
taxadb (ITIS)	2980	48.72	5.668
taxonomyCleanr (ITIS)	287	4.69	4931.376

The list of species that were not matched by the three packages that scored the highest number of resolved names is given in Appendix A. The table containing all the input species and the corresponding outputs per combination of package and data provider can be found in the supplemental files.

The positioning of the packages in the NMDS ordination space (Fig. 2) reflects the degree of similarity in their outputs. *taxonomyCleanr* (data source = ITIS), *taxizedb* (data source = GBIF, ITIS) and *taxadb* (data source = ITIS) were excluded from this visualisation because the number of resolved taxon names was much lower than the other packages that they obscured meaningful differences among the remaining methods, as shown in Appendix B. *TNRS* (data source = WCVP), *WorldFlora*, *lcvplants* and *taxonomyCleanr* (data source = GBIF) cluster tightly near the origin, suggesting that these methods produce consistent taxonomic resolutions. *taxadb* (data source = COL) is positioned slightly above this cluster, indicating moderate overlap in species name standardization, similarly to *TNRS* (data source = WFO) which is positioned slightly below instead. *rWCVP* and *taxadb* (data source = GBIF) are more dispersed from the main cluster, indicating more differences in species names resolution compared to the other methods.

The UpSet plot (Fig. 3) illustrates the degree of overlap in the outcomes across the packages. *taxonomyCleanr* (data source = ITIS) was excluded from this visualisation because of the excessively low number of matches. The UpSet plot comprising *taxonomyCleanr* (data source = ITIS) is in Appendix C. The largest intersection (2433 species names) involves all the considered packages, while the second one (2161 species names) includes all packages except for ITIS-based results, meaning that the number of shared matches among all packages except for ITIS-based packages is equal to 4594 (the sum of the first two intersections).

4. Discussion

Taxonomic harmonisation is an essential step in biodiversity research, as it helps reduce uncertainties and mistakes related to the inherent heterogeneous nature of biodiversity Big Data, which could potentially affect the results of downstream analyses. Despite its importance, there is no clear guidance on how to make the most out of

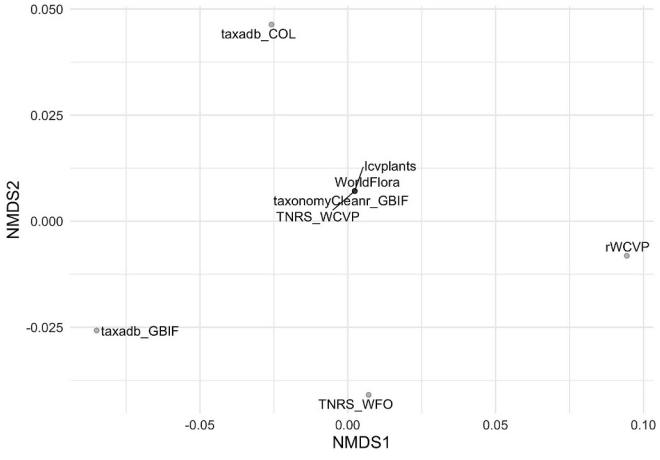


Fig. 2. Non-metric multidimensional scaling (NMDS) ordination based on Jaccard dissimilarities, highlighting the similarities and differences among the outputs of the tested R packages for the taxonomic harmonisation of vascular plant names. Each point represents one package-database combination, and the relative distances among points indicate the degree of dissimilarity in the resolved names. The axes represent the two main ordination axes summarising the multivariate dissimilarity structure in a reduced space.

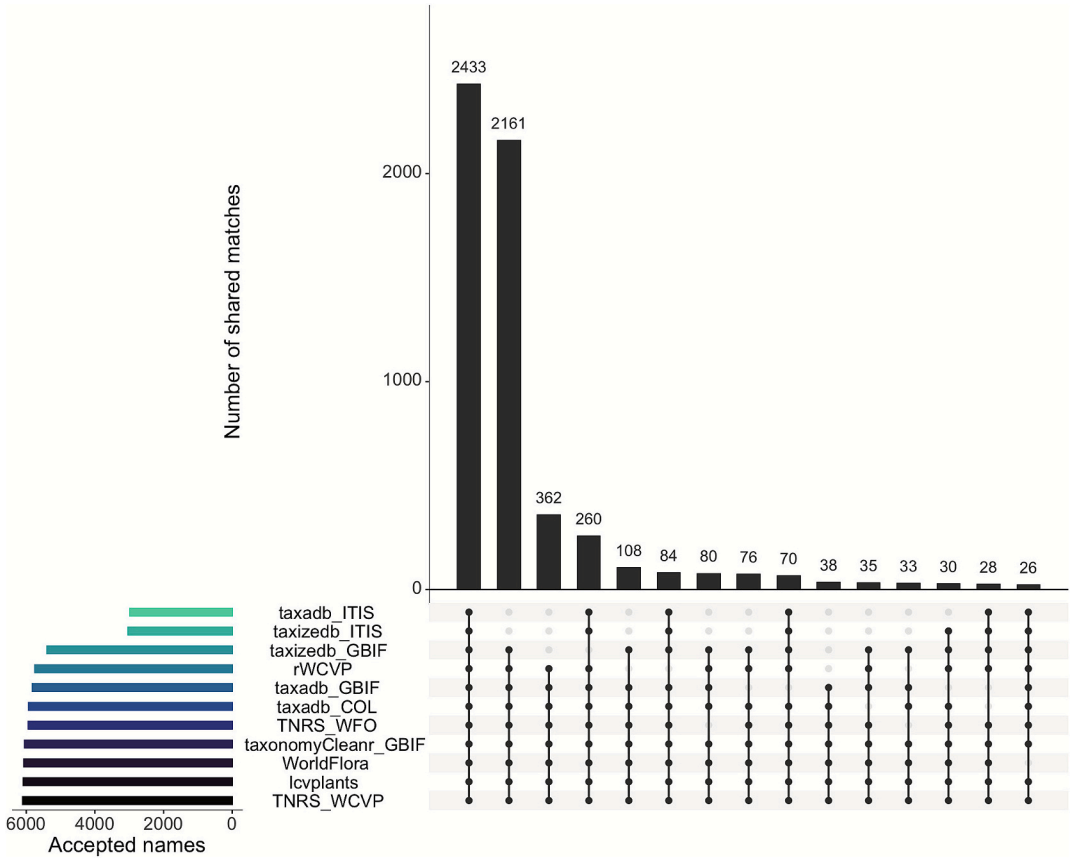


Fig. 3. UpSet plot illustrating the number of shared matched names among the tested methods (first 15 intersections shown). The bar chart on the bottom left represents the total number of species matched to an accepted name per package; the upper bar chart represents the number of shared species matched to an accepted name among the different combinations of packages and data providers, indicated by the connected black dots in the matrix below.

the wide range of tools built for the purpose. In this study, we developed a reproducible workflow for vascular plant name harmonisation, and conducted a comparative analysis of existing R packages.

We specifically focused on R packages due to their compatibility with reproducible workflows, an essential requirement in biodiversity informatics (Michener and Jones, 2012). Our results highlighted considerable heterogeneity in both the packages' functioning and the taxonomic databases they rely on, which both substantially affected the outputs.

Packages *TNRS* (data provider = WCVP), *lcvplants*, and *WorldFlora* were the most effective in terms of numbers, as they managed to resolve to an accepted name almost the entirety of the input names, while other packages, such as *taxizedb* (data provider = ITIS), *taxadb* (data provider = ITIS) and *TaxonomyCleanr* (data provider = ITIS) resolved less than 50 % of the names. The performance gap among the various packages appeared to stem from two main factors.

First, *TNRS*, *lcvplants*, and *WorldFlora* all relied on plant specific taxonomic databases, respectively WCVP, LCVP, and WFO, which are among the most comprehensive checklists for vascular plant names worldwide (Schellenberger Costa et al., 2023). Instead, *taxizedb*, *taxadb*, and *taxonomyCleanr* resolved the least amount of names when relying on ITIS. In fact, ITIS was originally developed as a taxonomic database covering multiple kingdoms, with a primary focus on North American taxa (ITIS, 2025), so it could be not well suited for the standardization of plant names outside of that area. ITIS is also part of GBIF and COL, two other databases with a broad taxonomic coverage. These were both used as taxonomic authorities in this study, *via* packages *taxonomyCleanr* (GBIF), *taxadb* (GBIF and COL), and *taxizedb* (GBIF), resolving much more names than when relying on ITIS, but still fewer names compared to packages relying on plant-specific taxonomic sources. However, in the case of *taxonomyCleanr* (data provider = ITIS) the result is likely due to technical inconsistencies in how the package interacts with the ITIS database. Consequently, this result should be interpreted as a technical artefact.

The second major factor influencing the number of matches appeared to be the possibility to use fuzzy matching. The three packages that resolved the greatest number of names all allowed fuzzy matching, unlike most of the other packages. Fuzzy matching can improve name resolution by accounting for typographical errors and misspellings. As pointed out by Grenié et al. (2023), however, fuzzy matching must be carefully used, since allowing too much distance could lead to matching two different taxa with similar names to the same accepted name.

Both the NMDS and the UpSet plot revealed substantial differences in taxonomic resolution outcomes among the evaluated R packages, underscoring how different taxonomic backbones and algorithmic implementations can lead to different outputs. The four best-performing packages clustered closely in the NMDS ordination space, suggesting a high degree of consistency in the names they managed to resolve. This pattern was further supported by their strong intersection in the UpSet plot, where these same packages appear together in the first 14 intersections, highlighting their overlap in the accepted names they managed to resolve. Excluding *taxonomyCleanr* (data provider = ITIS), less than a half of all accepted names were consistently resolved across all methods, showing the significant impact that the choice of both package and taxonomic backbone can have on biodiversity analyses.

Also the execution timings varied considerably across the packages, with the fastest tools generally being those that run locally and do not support fuzzy matching - such as *taxizedb* and *taxadb*, which completed in less than one minute. In contrast, *WorldFlora,* despite being a locally run package, exhibited one of the longest execution timing (> 5 h), likely due to its internal matching procedures. While local execution

offers speed, reproducibility, and independence from external servers, tools relying on APIs – like *TNRS* - benefit from up-to-date taxonomies, but can be limited by server availability and network speed (Norman et al., 2020a, 2020b).

There is almost no overlap between the species that the three best-performing packages did not manage to resolve. Among the species not resolved by the three packages, only input species "*Tormimalus fiorentina*" was unresolved by both *TNRS* (data source = WCVP) and *lcvplants*; instead, on WFO *Tormimalus florentina* is listed as a synonym of *Eriolobus florentinus* (WFO, 2025), and was indeed resolved by package *WorldFlora*. All the other unresolved species were not shared among the three packages, which is consistent with the findings of Schellenberger Costa et al. (2023), highlighting the differences among the four global checklists of plant names (WCVP, LCVP, WFO and WorldPlants). In their study, they also pointed out the fact that WFO had still not caught up with the other global checklists of plant names, which could explain why among the three best performing packages *WorldFlora* was the one with the greatest number of unresolved species. In particular, *WorldFlora* appeared to have issues with the species of the genera *Poa* and *Rosa*. *Poa* species were actually present in the version of the database we used, and listed as "Accepted" in the taxonomicStatus column of the WFO data used, but did not yield results after the matching process. Instead, species belonging to genus *Rosa* were all listed as "Unchecked" taxonomic status, so it is probably due to the database still being updated. In the case of *TNRS* (data source = WCVP) and *lcvplants*, most of the unresolved species are either because: (i) they are usually treated as hybrids, so they could be included in the database with an "x" symbol and/or could not be recognized as accepted names (*e.g.*, *Iris hybrida*, *Ophrys barlae*, *Populus canadensis*, *Salix fragilis, Anacamptis bornemannii*); (ii) their placement have been recently revised (*e.g.*, *Acrospelion distichophyllum* (Barberá et al., 2020); species belonging to genus *Lophiolepis* (Del Guacchio et al., 2022), *Soda inermis* (Mosyakin and Freitag, 2023), species belonging to genus *Thiliphthisa* (Del Guacchio and Caputo, 2020)). In these cases, the databases could have not caught up at the time of our analyses.

It is important to note that our approach may present some limitations. First of all, our assessment prioritised quantity, *i.e.* the number of input species matched, as we were not able to evaluate the quality of every single match because of the great number of names. Future work on the topic could include manual validation with experts to quantify misassignments attributable to fuzzy matching or taxonomic misalignments. Also, we did not account for historical taxonomic changes, where species names used in older records may have referred to different concepts or circumscribed taxa than their modern equivalents. Although our assessment focused exclusively on vascular plants, further analysis could be developed for other major taxonomic groups.

Before proceeding with the taxonomic harmonisation of names, it is advisable to remove occurrence records containing problematic coordinates. In fact, species occurrence records may contain erroneous or imprecise geographical coordinates and temporal information (Zizka et al., 2019). Geographical errors are particularly common and may include occurrences assigned to administrative centroids due to vague locality descriptions, swapped latitude and longitude values, incorrect zero coordinates from data entry mistakes, errors in the conversion to decimal degrees, and records from biodiversity institutions (*e.g.*, zoos, botanical gardens and museums) (Gueta and Carmel, 2016; Robertson et al., 2016; Zizka et al., 2019). Tools such as the *CoordinateCleaner* R package provide automated routines to flag and remove records located around administrative centroids, urban areas, or biodiversity institutions, that may otherwise bias spatial analyses. Although this study focused specifically on the taxonomic harmonisation step, spatial cleaning remains an essential component of biodiversity data cleaning, particularly when dealing with historical records or aggregated data from multiple sources, where locality information can be imprecise or outdated (Zizka et al., 2020).

Our comparative analysis of R packages for the taxonomic harmonisation also allowed us to identify a set of key features to maximise accuracy, efficiency, and flexibility. First, the package should rely on a plant-specific and actively maintained taxonomic backbone. Second, it should implement the possibility to perform fuzzy matching. Third, the tool should support local execution to enhance speed and reproducibility. In addition, high-quality support documentation is also crucial to ensure usability across a broad range of users. Finally, packages should ideally offer the flexibility to work with user-defined checklists, a possibility that was only given by *WorldFlora* among the tested packages, thanks to the function *new.backbone()*. Among the tested tools, *WorldFlora* was the only package that currently offers most of these functionalities, although it was also the slowest in terms of execution time and only allows local execution. Even though our analyses highlighted pros and potential pitfalls of the tested tools, we suggest users who are dealing with biodiversity data cleaning to test more than one tool, since results may vary based on the species they are studying, and the time and the area in which the analyses are carried out.

CRediT authorship contribution statement

Diletta Santovito: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Alessandro Chiarucci:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Duccio Rocchini:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Francesco Santi:** Writing – review & editing, Validation. **Rocìo Beatriz Cortès Lobos:** Writing – review & editing, Validation. **Riccardo Testolin:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Quentin Groom (Meise Botanic Garden, Nieuwelaan, Belgium) for valuable discussions and insights regarding the species that were not resolved by the best-performing taxonomic harmonisation packages.

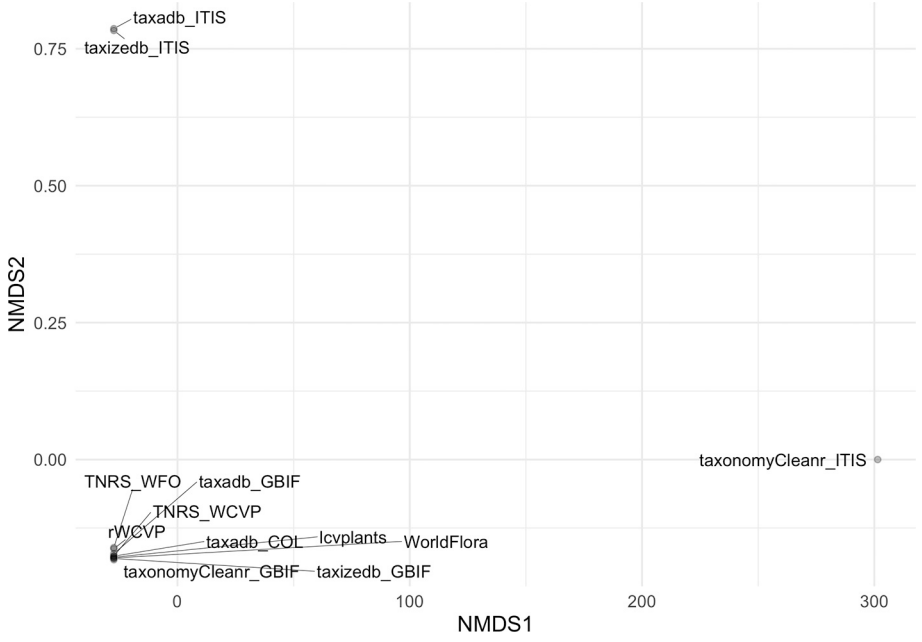
The project was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP J33C22001190001, Project title "National Biodiversity Future Center - NBFC".

Appendix A. List of the input species that were not resolved using the three packages that scored the highest amount of names matched to an accepted name

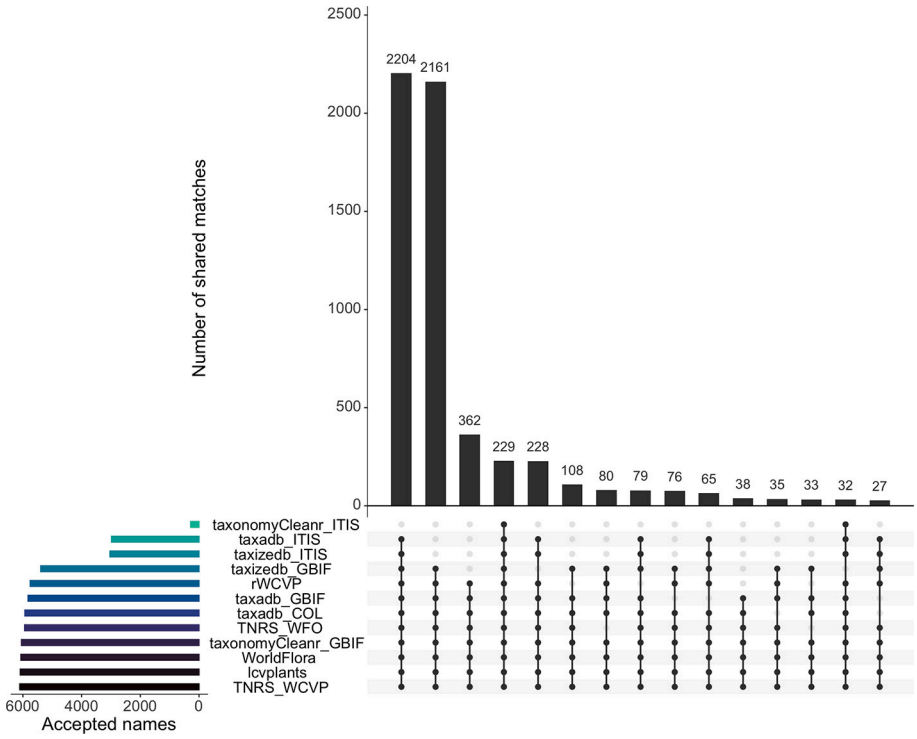
TNRS (data provider = WCVP), 11 unresolved species
Asplenium inexpectans
Facchinia villarsii
Geranium austroapenninum
Iris hybrida
Narcissus deficiens
Ophrys barlae
Phyteuma italicum
Populus canadensis
Salix fragilis
Taraxacum alpinum
Tormimalus fiorentina
lcvplants, 28 unresolved species
Acrospelion distichophyllum
Anacamptis bornemannii
Bellidiastrum michelii
Cardamine silana
Chrysojasminum fruticans
Kroenleinia grusonii
Lophiolepis eriophora
Lophiolepis ferox
Lophiolepis lobelii
Lophiolepis morisiana
Lophiolepis spatulata
Lophiolepis tenoreana
Lophiolepis vallis-demonii
Marcus-Kochia littorea
Marcus-Kochia ramosissima
Oloptum miliaceum
Oloptum thomasii
Pedicularis hoermanniana
Roemeria apula
Silene badaroi
Soda inermis
Soda oppositifolia
Solenopsis baccchettae
Straphisagria macrosperma
Stipa austroitalica
Thiliphthisa apuana
Thiliphtisa purpurea
Tormimalus fiorentina
WorldFlora, 50 unresolved species
Aira elegans
Alchemilla vulgaris
Allium cepa
Anthyllis vulneraria
Artemisia alba
Arum maculatum
Carex acuta
Centaurea alba
Ceropegia europaea
Cirsium heterophyllum x spinosissimum
Dactylorhiza fuchsii x incarnata
Facchinia rupestris
Festuca ovina
Festuca rubra
Geum montanum
Hieracium murorum
Hieracium pilosum
Iris japonica
Iris orientalis
Koenigia alpina
Linum alpinum

Linum grandiflorum
Mcneillia graminifolia
Mentha arvensis
Mentha longifolia
Ophrys fuciflora x insectifera
Orchis militaris x purpurea
Oryza sativa
Oxybasis rubra
Phyteuma nigrum x spicatum
Poa angustifolia
Poa compressa
Poa nemoralis
Poa palustris
Poa pratensis
Potentilla alba
Potentilla recta
Pyrus communis
Rabelera holostea
Ranunculus acris
Ranunculus auricomus
Rosa canina
Rosa glauca
Rosa micrantha
Rosa rubiginosa
Rosa tomentosa
Rosa villosa
Rubus hirtus
Thymus serpyllum
Viola williamsii

Appendix B. NMDS plot (distance = Jaccard) without the exclusion of any packages



Appendix C. UpSet plot including *taxonomyCleanr* (data provider = ITIS)



Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103543.

Data availability

Occurrence data for vascular plants in Italy were downloaded from the Botanical Information and Ecology Network (BIEN, https://bien.nceas.ucsb.edu/bien/) and the Global Biodiversity Information Facility (GBIF, https://www.gbif.org).

The BIEN data were retrieved using the *BIEN* R package (Maitner et al., 2018) on 22 January 2025 and are licensed under a CC-BY-NC-ND license. As BIEN does not provide DOIs for individual downloads, the dataset used in this study has been archived at doi: https://doi.org/10.5281/zenodo.16317606. The GBIF data are publicly available and can be accessed *via* doi: 10.15468/dl.efd2p7. The scripts needed to reproduce all the analyses carried out are provided in the supplementary files.

References

Bååth, R., 2024. beepr: Easily Play Notification Sounds on any Platform. R PACKAGE VERSion 2.0. https://CRAN.R-project.org/package=beepr.

Barberá, P., Soreng, R.J., Peterson, P.M., Romaschenko, K., Quintanar, A., Aedo, C., 2020. Molecular phylogenetic analysis resolves Trisetum (Poaceae: Pooideae: Koeleriinae) polyphyletic: evidence for a new genus, Sibirotrisetum and resurrection of Acrospelion. J. Syst. Evol. 58 (4), 517–526. https://doi.org/10.1111/jse.12523.

Becker, R.A., Wilks, A.R., 2022. mapdata: Extra Map Databases. R Package Version 2.3.1. https://CRAN.R-project.org/package=mapdata.

BIEN, 2025. http://bien.nceas.ucsb.edu/bien/ (accessed January 2025).

Brown, M.J.M., Walker, B.E., Black, N., Govaerts, R., Ondo, I., Turner, R., Nic, Lughadha E., 2023. rWCVP: a companion R package to the World Checklist of Vascular Plants. New Phytol. 240, 1355–1365. https://doi.org/10.1111/nph.18919.

Chamberlain, S., Szocs, E., 2013. taxize - taxonomic search and retrieval in R. F1000Research 2, 191. https://doi.org/10.12688/f1000research.2-191.v2.

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., Ram, K., 2024. rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.8.1. https://CRAN.R-project.org/package=rgbif.

Chamberlain, S., Arendsee, Z., Stirling, T., 2025. taxizedb: Tools for Working with 'Taxonomic' Databases. R package version 0.3.1. https://doi.org/10.5281/zenodo.1158055.

Del Guacchio, E., Caputo, P., 2020. Splitting Asperula (Rubiaceae): a proposal for consistency purposes within sections Cynanchicae, Thliphthisa and Hexaphylla. Plant Biosyst. 154 (5), 766–782. https://doi.org/10.1080/11263504.2020.1804008.

Del Guacchio, E., Bureš, P., Iamonico, D., Carucci, F., De Luca, D., Zedek, F., Caputo, P., 2022. Towards a monophyletic classification of Cardueae: restoration of the genus Lophiolepis (= Cirsium pp) and new circumscription of Epitrachys. Plant Biosyst. 156 (5), 1269–1290. https://doi.org/10.1080/11263504.2022.2131924.

Devictor, V., Bensaude-Vincent, B., 2016. From ecological records to big data: the invention of global biodiversity. Hist. Philos. Life Sci. 38, 1–23. https://doi.org/10.1007/s40656-016-0113-2.

Freiberg, M., Winter, M., Gentile, A., Zizka, A., Muellner-Riehl, A., Weigelt, A., Wirth, C., 2020. LCVP, the Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. Sci. Data 416. https://doi.org/10.1038/s41597-020-00702-z.

Garnier, S., Ross, N., Rudis, R., Camargo, A.P., Sciaini, M., Scherer, C., 2024. viridis(Lite) - Colorblind-Friendly Color Maps for *R. viridis* Package Version 0.6.5. https://doi.org/10.5281/zenodo.4679423.

GBIF, 2025. Global Biodiversity Information Facility. Occurrence download. https://www.gbif.org/occurrence/download/0001127-250121130708018.

Gehlenborg, N., 2019. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. R package version 1.4.0. https://CRAN.R-project.org/package=UpSetR.

Govaerts, R., 2024. WCVP: World Checklist of Vascular Plants. Facilitated by the Royal Botanic Gardens, Kew. https://doi.org/10.34885/nswv-8994.

Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G.M.L., Sagouis, A., Winter, M., 2023. Harmonizing taxon names in biodiversity data: a review of tools, databases

and best practices. Methods Ecol. Evol. 14 (1), 12–25. https://doi.org/10.1111/2041-210X.13802.
Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. J. Stat. Softw. 40 (3), 1–25. https://doi.org/10.18637/jss.v040.i03.
Gueta, T., Carmel, Y., 2016. Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. Ecol. Inform. 34, 139–145. https://doi.org/10.1016/j.ecoinf.2016.06.001.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, R., Tulloch, A. I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16 (12), 1424–1435. https://doi.org/10.1111/ele.12189.
Guo, Q., He, Z., Wang, Z., 2023. Long-term projection of future climate change over the twenty-first century in the Sahara region in Africa under four Shared Socio-Economic Pathways scenarios. Environ. Sci. Pollut. Res. 30, 2319–22329. https://doi.org/10.1007/s11356-022-23813-z.
IPCC, 2023. In: Lee, H., Romero, J. (Eds.), Climate Change 2023: Synthesis report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland, p. 184. https://doi.org/10.59327/IPCC/AR6-9789291691647.
Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B., 2020. Data integration for large-scale models of species distributions. Trends Ecol. Evol. 35 (1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006.
ISPRA, 2009. Environmental Data Yearbook 2009 - Key Topics. https://www.isprambiente.gov.it/en/publications/state-of-the-environment/key-topics-1.
ITIS, 2025. www.itis.gov. doi: https://doi.org/10.5066/F7KH0KBK (accessed January 2025).
Izrailev, S., 2024. tictoc: Functions for Timing R Scripts, as Well as Implementations of "Stack" and "StackList" Structures. R Package Version 1.2.1. https://CRAN.R-project.org/package=tictoc.
Jetz, W., McPherson, J.M., Guralnick, R.P., 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol. Evol. 27 (3), 151–159. https://doi.org/10.1016/j.tree.2011.09.007.
Kassambara, A., 2023. ggpubr: 'ggplot2' Based Publication Ready Plots. R Package Version 0.6.0. https://CRAN.R-project.org/package=ggpubr.
Keesing, F., Belden, L., Daszak, P., Dobson, A., Harvell, C.D., Holt, R.D., Hudson, P., Jolles, A., Jones, K.E., Mitchell, C.E., Myers, S.S., Bogich, T., Ostfeld, R.S., 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. Nature 468, 647–652. https://doi.org/10.1038/nature09575.
Kindt, R., 2020. WorldFlora: an R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. Appl. Plant Sci. 8 (9), e11388. https://doi.org/10.1002/aps3.11388.
Lai, J., Lortie, C.J., Muenchen, R.A., Yang, J., Ma, K., 2019. Evaluating the popularity of R in ecology. Ecosphere 10 (1), e02567. https://doi.org/10.1002/ecs2.2567.
Maitner, B., 2023. BIEN: Tools for Accessing the Botanical Information and Ecology Network Database. R Package Version 1.2.6. https://CRAN.R-project.org/package=BIEN.
Maitner, B., Boyle, B., 2024. TNRS: Taxonomic Name Resolution Service. R package version 0.3.6. https://CRAN.R-project.org/package=TNRS.
Maitner, B.S., Boyle, B., Casler, N., Condit, R., Donoghue, J., Duran, S.M., Guaderrama, D., Hinchliff, C.E., Jorgensen, P.M., Kraft, N.J.B., McGill, B., Merow, C., Morueta-Holme, N., Peet, R.K., Sandel, B., Schildhauer, M., Smith, S.A., Svenning, J.-C., Thiers, B., Enquist, B.J., 2018. The BIEN R package: a tool to access the Botanical Information and Ecology Network (BIEN) database. Methods Ecol. Evol. 9 (2), 373–379. https://doi.org/10.1111/2041-210X.12861.
Meyer, C., Weigelt, P., Kreft, H., 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol. Lett. 19 (8), 992–1006. https://doi.org/10.1111/ele.12624.
Michener, W.K., Jones, M.B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol. Evol. 27 (2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016.
Mosyakin, S.L., Freitag, H., 2023. (3000) proposal to conserve the name Soda against Sevada (Chenopodiaceae s. str./Amaranthaceae sl). Taxon 72 (6), 1371–1373. https://doi.org/10.1002/tax.13093.
Müller, K., Walthert, L., 2024. styler: Non-invasive Pretty Printing of R Code. R Package Version 1.10.3. https://CRAN.R-project.org/package=styler.
Musvuugwa, T., Dlomu, M.G., Adebowale, A., 2021. Big data in biodiversity science: a framework for engagement. Technologies 9 (3), 60. https://doi.org/10.3390/technologies9030060.
Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858. https://doi.org/10.1038/35002501.
Navarro, L.M., Fernandez, N., Guerra, C., Guralnick, R., Kissling, W.D., Londoño, M.C., Muller-Karger, F., Turak, E., Balvanera, P., Costello, M.J., Delavaud, A., El Serafy, G., Ferrier, S., Geijzendorffer, I., Geller, G.N., Jetz, W., Kim, E., Kim, H., Martin, C.S., McGeoch, Pereira, H.M., 2017. Monitoring biodiversity change through effective global coordination. Curr. Opin. Environ. Sustain. 29, 158–169. https://doi.org/10.1016/j.cosust.2018.02.005.
Norman, Kari E.A., Chamberlain, Scott, Boettiger, Carl, 2020a. Taxadb: A high-performance local taxonomic database interface. Methods Ecol. Evol. 11 (9), 1153–1159. https://doi.org/10.1111/2041-210X.13440.
Norman, K.E., Chamberlain, S., Boettiger, C., 2020b. taxadb: a high-performance local taxonomic database interface. Methods Ecol. Evol. 11 (9), 1153–1159. https://doi.org/10.1111/2041-210X.13440.
Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H., Weedon, J., 2024. vegan: Community Ecology Package. R Package Version 2.6-8. https://CRAN.R-project.org/package=vegan.
Pebesma, E., Bivand, R., 2023. Spatial Data Science: With Applications in R, 1st ed. Chapman and Hall/CRC, Boca Raton, p. 314. https://doi.org/10.1201/9780429459016.
Pereira, H.M., Navarro, L.M., Martins, I.S., 2012. Global biodiversity change: the bad, the good, and the unknown. Annu. Rev. Env. Resour. 37 (1), 25–50. https://doi.org/10.1146/annurev-environ-042911-093511.
Proença, V., Martin, L.J., Pereira, H.M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R.D., Pradinho Honrado, J., Jürgens, N., Opige, M., Schmeller, D.S., Tiago, P., van Swaay, C.A., 2017. Global biodiversity monitoring: from data sources to essential biodiversity variables. Biol. Conserv. 213, 256–263. https://doi.org/10.1016/j.biocon.2016.07.014.
R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Robertson, M.P., Visser, V., Hui, C., 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. Ecography 39, 394–401. https://doi.org/10.1111/ecog.02118.
Schellenberger Costa, D., Boehnisch, G., Freiberg, M., Govaerts, R., Grenié, M., Hassler, M., Kattge, J., Muellner-Riehl, A.N., Rojas Andrés, B.M., Winter, M., Watson, M., Zizka, A., Wirth, C., 2023. The big four of plant taxonomy - a comparison of global checklists of vascular plant names. New Phytol. 240 (4), 1687–1702. https://doi.org/10.1111/nph.18961.
Slowikowski, K., 2024. ggrepel: Automatically Position Non-overlapping Text Labels with 'ggplot2'. R Package Version 0.9.6. https://CRAN.R-project.org/package=ggrepel.
Smith, C., 2025. taxonomyCleanr: A Workflow and Set of Functions to Clean Taxonomy Data Using R. R Package Version 1.6.5. https://github.com/EDIorg/taxonomyCleanr.
Stinca, A., Musarella, C.M., Rosati, L., Laface, V.L.A., Licht, W., Farfarillo, E., Wagensommer, R.P., Galasso, G., Fascetti, S., Esposito, A., Fiaschi, T., Nicolella, G., Chianese, G., Ciaschetti, G., Salerno, G., Fortini, P., Di Pietro, R., Perrino, E.V., Angiolini, C., De Simone, L., Mei, G., 2021. Italian vascular flora: new findings, updates and exploration of floristic similarities between regions. Diversity 13 (11), 600. https://doi.org/10.3390/d13110600.
International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. In: Turland, N.J., Wiersema, J.H., Barrie, F.R., Greuter, W., Hawksworth, D.L., Herendeen, P.S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T.W., McNeill, J., Monro, A.M., Prado, J., Price, M.J., Smith, G.F. (Eds.), 2018. Regnum Vegetabile 159. Koeltz Botanical Books, Glashütten. https://doi.org/10.12705/Code.2018.
WFO, 2025. Tormimalus florentina (Zuccagni) Holub. Published on the Internet. http://www.worldfloraonline.org/taxon/wfo-0000991513. accessed 26 June 2025.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.
Zermoglio, P.F., Guralnick, R.P., Wieczorek, J.R., 2016. A standardized reference data set for vertebrate taxon name resolution. PloS One 11 (1), e0146894. https://doi.org/10.1371/journal.pone.0146894.
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., Antonelli, A., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10 (5), 744–751. https://doi.org/10.1111/2041-210X.13152.
Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N.M.C., Santos, T.A.B., Dos Santos-Costa, R.C., Serrano, F.C., Alves da Silva, A.P., de Souza Soares, A., Cavalcante de Souza, P.G., Calisto Tomaz, E., Vale, V.F., Antonelli, A., 2020. No one-size-fits-all solution to clean GBIF. PeerJ 8, e9916. https://doi.org/10.7717/peerj.9916.