Received: 7 September 2021 | Accepted: 21 December 2021

DOI: 10.1111/2041-210X.13802

Methods in Ecology and Evolution

REVIEW

Realising the Promise of Large Data and Complex Models

Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices

Matthias Grenié[1,2] | Emilio Berti[1,3] | Juan Carvajal-Quintero[1,2] | Gala Mona Louise Dädlow[1,2] | Alban Sagouis[1,4] | Marten Winter[1,2]

[1]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[2]Leipzig University, Leipzig, Germany

[3]Friedrich-Schiller University Jena, Jena, Germany

[4]Department of Computer Science, Martin Luther University, Halle, Germany

Correspondence
Matthias Grenié
Email: matthias.grenie@idiv.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG-FZT 118 and 202548816

Handling Editor: Laura Graham

Abstract

1. The process of standardizing taxon names, taxonomic name harmonization, is necessary to properly merge data indexed by taxon names. The large variety of taxonomic databases and related tools are often not well described. It is often unclear which databases are actively maintained or what is the original source of taxonomic information. In addition, software to access these databases is developed following non-compatible standards, which creates additional challenges for users. As a result, taxonomic harmonization has become a major obstacle in ecological studies that seek to combine multiple datasets.
2. Here, we review and categorize a set of major taxonomic databases publicly available as well as a large collection of R packages to access them and to harmonize lists of taxon names. We categorized available taxonomic databases according to their taxonomic breadth (e.g. taxon specific vs. multi-taxa) and spatial scope (e.g. regional vs. global), highlighting strengths and caveats of each type of database. We divided R packages according to their function, (e.g. syntax standardization tools, access to online databases, etc.) and highlighted overlaps among them. We present our findings (e.g. network of linkages, data and tool characteristics) in a ready-to-use Shiny web application (available at: https://mgrenie.shinyapps.io/taxharmonizexplorer/).
3. We also provide general guidelines and best practice principles for taxonomic name harmonization. As an illustrative example, we harmonized taxon names of one of the largest databases of community time series currently available. We showed how different workflows can be used for different goals, highlighting their strengths and weaknesses and providing practical solutions to avoid common pitfalls.
4. To our knowledge, our opinionated review represents the most exhaustive evaluation of links among and of taxonomic databases and related R tools. Finally, based on our new insights in the field, we make recommendations for users, database managers and package developers alike.

KEYWORDS
R packages, taxonomic databases, taxonomic harmonization, taxonomic name matching, taxonomic tools, taxonomy

1 | INTRODUCTION

In the era of big data, combining, harmonizing and analysing massive amounts of ecological data have played a central role in improving our understanding of biodiversity in a changing world (Hampton et al., 2013; La Salle et al., 2016; Michener & Jones, 2012; Wüest et al., 2020). While promising, this new era is also challenging. As exabytes of primary biodiversity data become publicly available, issues of quality control in data integration, interoperability and redundancy have become pressing concerns to address (Jin & Yang, 2020; Kissling et al., 2018; Lenters et al., 2021; Nelson & Ellis, 2019; Soberón & Peterson, 2004; Thomas, 2009; Wüest et al., 2020).

One of the biggest challenges in biodiversity data handling is maintaining a consistent taxonomy of species names associated with different biological attributes (Jin & Yang, 2020; Meyer et al., 2016; Tessarolo et al., 2017; Thomas, 2009). The dynamic nature of taxonomy, reinforced by the growing availability of information and the increasing use of genetic methods to identify species results in ever-changing taxon names considered accepted. Taxonomists start by sampling individuals in the field and when considered as not yet described, name them, based on best knowledge and defined procedures (Dayrat, 2005). These names become de facto accepted. However, some names can become obsolete, when, for example, researchers realize later on this species was named already before. Those names then are used as synonyms of another now accepted name (Lepage et al., 2014). In addition to the names per se, taxonomists refer to species through taxonomic concepts—that is, biological entities—(Lepage et al., 2014). Which taxonomic concepts researchers use, that is, are defined as legitimate and valid, can vary across research cultures (Lepage et al., 2014). For some taxonomic groups general consensus on one taxonomic concept is far from being reached (Chawuthai et al., 2016), generating confusion. This dynamic process results in difficulties for end users to point to single valid names referring unambiguously to single taxonomic concepts. The use of taxonomic databases helps resolve the different relationships that exist between names and taxonomic concepts (one-to-one, one-to-many, many-to-one or even many-to-many, see Lepage et al., 2014).

In an attempt to unify taxonomy across the tree of life, multiple initiatives have proposed curated lists of taxon names referenced against accepted taxon names. Taxonomic databases (Box 1) are usually based on extensive community and individual expert knowledge. Decisions which taxon names are accepted are usually based on robust scientific evidence. These decisions might also have to be based on less objective reasons, like reliability of original resources in comparison to conflicting studies or on individual preferences for grammar and spelling (e.g. Isoëtes vs. Isoetes; Isaac et al., 2004). However, despite significant efforts in creating a single authoritative list of the world's taxa (e.g. [37]), taxonomic unification has largely advanced through multiple independent efforts with different aims and scopes (e.g. per taxon group or region; Costello, 2020; Garnett et al., 2020). For example, some taxonomic databases, that is, databases that primarily offer reference taxonomic data, focus on specific taxonomic groups (e.g. Freiberg et al., 2020), others on environmental realms (e.g. [34]), providing a reference at either global or regional scale such as national databases (Figure 1). The last decade brought a lot of progress in taxonomy in general to overcome the 'taxonomic impediment' (Rouhan & Gaudeul, 2021), the lack of comprehensive information per taxonomic group. These efforts have generated a large number of taxa lists with taxonomic-curated information dispersed across very different repositories (König et al., 2019). For example, we are aware of four global taxonomic databases focusing on plants (Leipzig Catalogue of Vascular Plants [22]; World Flora Online [30]; Plants of the World Online [23]; World Plants, Hassler, 2021). While we know that different databases provide different scientific opinions on taxonomy (i.e. using different taxonomic concepts), meaning that they all contribute to the scientific debate and none of them is right or wrong, how should the non-taxonomy expert end user (e.g. macroecologists) know which resource is most suitable for her/his purposes? Researchers in need of validating taxon names are confronted with many different taxonomic databases that have often overlapping spatial or taxonomic coverage without a clear way to select which database to use.

Taxonomic information, through taxon names (Figure 2), can serve as a common basis to index and merge different biodiversity data (e.g. Dyer et al., 2017; occurrences: GBIF: The Global Biodiversity Information Facility, 2020; conservation status: IUCN, 2021; traits: Jones et al., 2009; Kattge et al., 2020; phylogenetic relationships: Smith & Brown, 2018; Upham et al., 2019; invasion status: van Kleunen et al., 2019). Aside from the challenges with maintaining updated and comprehensive taxonomic databases by themselves, combining and harmonizing additional biological data can be problematic since such datasets may have been created and updated at different times (sometimes spanning several decades), may use different taxonomic databases to standardize taxon names, and may not even be linked to any consistent taxonomic concept (Edwards et al., 2000; Farley et al., 2018; König et al., 2019). Ultimately, if taxonomic name harmonization is not properly executed, researchers are likely to introduce and propagate errors that can lead to misquantified biodiversity components or mismatched data (Bortolus, 2008). Larger amounts of data increase the issue, due to taxonomic inaccuracies introduced for increasing numbers of species and taxonomic breadth (Patterson et al., 2010).

Driven by the needs in data harmonization, multiple tools have emerged for this task. This has generated a diverse toolbox but no clear guidance on how these tools could be combined into a meaningful and efficient workflow. Improving our knowledge of the landscape of available taxonomic reference and tools is thus critical to developing robust and comprehensive workflows to achieve high levels of data quality and accurate downstream analyses.

Here, we fill this gap by reviewing publicly available taxonomic databases and R packages for taxonomic harmonization, describing common pitfalls to avoid when using them, and proposing hands-on approaches to achieve accurate and precise harmonized list of taxon names. To our knowledge, our study represents the most comprehensive review and assessment of tools and issues related to taxonomic name harmonization. We present and discuss main steps towards

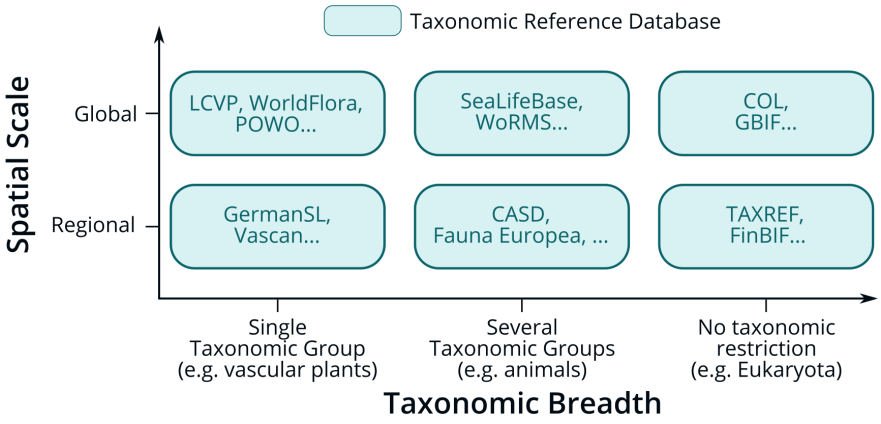

FIGURE 1 Typology of taxonomic databases according to their taxonomic breadth and their spatial scale. The *x*-axis represents increasing taxonomic breadth from a single taxonomic group to no clear taxonomic restriction (e.g. considering all biota or all Eukaryota). The *y*-axis represents spatial scale from regional to global. Each box represents a specific type of taxonomic database, with examples. LCVP, Leipzig Catalogue of Vascular Plants; WorldFlora, World Flora Online; POWO, Plants of the World Online; GermanSL, German Simple List; Vascan, Database of Vascular Plants of Canada; WoRMS, World Register of Marine Species; CASD, Chinese Animal Scientific Database; COL, Catalogue of Life; GBIF, Global Biodiversity Information Facility; TAXREF, French Taxonomic Referential; FinBIF, Finnish Biodiversity Information Facility

BOX 1 **The taxonomic terminology diversity**

Across the literature, the terms **taxonomic reference** (**list**; e.g. Freiberg et al., 2020), **taxonomic authority** (**list**/**file**; Vanden Berghe et al., 2015), **taxonomic databases** (Rees, 2014), **taxonomic backbone** (e.g. Schulman et al., 2021) or **taxonomic checklist** (Costello, 2020) are used interchangeably, often without clear definitions. The terminological diversity makes it difficult to understand differences between terms and potentially to find the correct resources. For example, the expression '**taxonomic authority**' can be confused with the authority when citing a species name, which is the citation of the author name associated with a taxon. Different expressions can sometimes reflect differences in sizes of provided databases, from a simple species **list** (e.g. to define the list of species names that occur in a given area), to a full nomenclatural **reference** (with a taxonomy), to systems that also provide synonymy resolution.

In this article, we use '**taxonomic databases**' as a generic expression of digital collections of taxonomic information on many individual species, with processes to mitigate potential conflicts between taxonomic designation.

robust and meaningful harmonization workflows. Specifically, we review taxonomic databases, R packages, and show how they depend on and interact with each other. We focus on R as it is the programming language of choice for ecologists (Lai et al., 2019). We present a Shiny R application that guides users through the labyrinth of tools and resources. We assess the efficiency of different possible taxonomic harmonization workflows through a concrete use case. We then formulate recommendations for end users, tool developers and taxonomic data managers.

2 | THE WILD WORLD OF TAXONOMIC RESOURCES

2.1 | A typology of taxonomic databases

We categorized taxonomic databases (see Box 1) along two axes: taxonomic breadth and covered spatial scale (Figure 1). Taxonomic breadth describes the amount of taxonomic groups covered by the database. We use the term 'taxonomic group' as a broad term to describe a group of taxa or taxonomic ranks at which people work (e.g. birds—class *Aves*, butterflies—order *Lepidoptera*). Databases have varying taxonomic and spatial breadths, from narrow taxonomic breadth but global scale (e.g. eBird [17]) to broad taxonomic breadth but regional/national scope (e.g. the Chinese Animal Species Database [4]). Some databases even aim to provide information without any taxonomic restriction at a global level, for example, Catalogue of Life [37].

Because navigating the landscape of taxonomic databases can be difficult for users, we provide a wide overview of available databases on as many taxonomic groups as possible at varying spatial scales and taxonomic breadths (Table 1). As one covering many databases, this list provides an entry point for users to get a sense of potential sources of taxonomy. The immense variety of taxonomic databases, especially at regional scales, prevents our list from being exhaustive but it includes most existing global databases.

2.2 | The wide landscape of R packages for taxonomy

With the increasing amount of data used in ecological studies, taxonomic harmonization cannot rely on manual curation. Computational tools are needed to help extract, evaluate, manipulate and visualize taxonomic information. Additionally, the use of computational tools

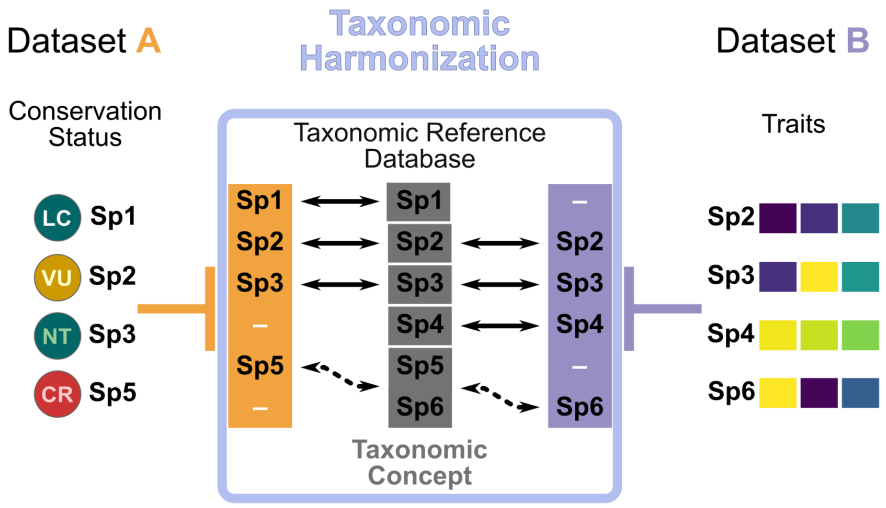

FIGURE 2 Taxonomy as a unifying key for ecological datasets. The two sides represent two exemplary datasets, with a containing conservation status of taxa (here species) and B their traits (colours show different traits). The datasets are indexed by taxon names 'Sp1' to 'Sp6'. The rounded rectangle in the middle depicts the taxonomic harmonization process: (a) the names are extracted from each dataset, respectively in the orange and purple rectangles; (b) both lists are then compared to a taxonomic database which harmonizes all names. Here the names 'Sp1' and 'Sp6' refer to the same taxon in the taxonomic database (as indicated by the dashed lines). Without taxonomic harmonization, the exact match of names would have resulted in the loss of Sp5 and Sp6 when merging both datasets. LC, NT, VU, and CR are abbreviations of Red List statuses, meaning least concern, not threatened, vulnerable, and critically endangered, respectively

increases the reproducibility of analyses compared to manual edits. In this section, we present the most extensive review, to our knowledge, of R packages that can be used to process taxonomic information (Table 2).

2.3 | Description of the landscape of tools

We identified some packages that provide standardized technical infrastructure for taxonomic experts to develop and work with taxonomic information within R. Infrastructure packages provide basic 'building blocks' for other packages to build onto. **taxa** [51], used by **metacoder** [104], provides R-native objects and methods to represent taxonomic data. **taxlist** [52] contains objects and functions to store taxa lists, synonyms, taxonomic hierarchy and functional traits in a standardized format; it is used by **vegdata** [102]. **taxview** [53] provides basic visualization of taxonomic hierarchies; it is used by no other packages. The fact that virtually no other packages rely on them means that several tools reinvent the wheel instead of relying on standardized functions. More widespread reliance on infrastructure packages and associated methods within the small community of R taxonomy package developers could foster best development practices, easier interoperability, as well as increased reproducibility, as it has been for example done already for spatial data through the **sp** and **sf** packages (Bivand et al., 2013; Pebesma, 2018; Pebesma & Bivand, 2005).

We identified 47 packages providing direct access to online taxonomic databases. These packages let the users search a given taxon name in one (or several) online taxonomic database(s) and get back a list of potential matching names, considering both accepted names and synonyms. Details about the packages, for example, which taxonomic databases they access are available in S2 and our specifically for this review developed shiny app **taxharmonizexplorer** (https://mgrenie.shinyapps.io/taxharmonizexplorer/). You can explore which package(s) access which database(s) as well as additional useful characteristics through **taxharmonizexplorer** described in the following section.

Accessing online databases does not come free of issues: (a) Online databases can be updated continuously, potentially leading to different versions used when harmonizing at different times or on- and offline, hindering reproducibility. (b) Database access is not always guaranteed because of technical issues with online resources (maintenance needed, server outage and Internet accessibility). (c) Some databases implement a form of request limitation, enforcing a maximum number of queries that can be made in a given period of time (e.g. one query every 3 s), with one query matching a single species only. (d) Online query execution speed can be limited compared to local queries (of the order of several seconds against tens of milliseconds, see [94, 95]) and potentially impossible if the Internet connection is unstable. (5) Databases also limit the complexity of queries with no standard format across databases, for example, the user can only get a list of accepted names from an input name and not ask more precise questions like 'What are all names with epithet *alba*?'.

To overcome these issues several packages provide or build local database copies. **lcvplants** [22] accesses the LCVP database fully offline through a local copy, it also offers functions to harmonize two lists of names. **ncbit** [60] provides a similar access but to the NCBI database [47]. **taxadb** [94, 95] creates a unified local database from different data sources as specified by the user. **taxalight** [96], which is maintained by the same developers, is faster and with fewer dependencies, it will supersede **taxadb** (C. Boettiger, pers. comm.). **taxizedb** [98] also downloads local copies of the database but, contrary to **taxadb** and **taxalight**, it provides the data without standardizing its format between sources. The user can then access

TABLE 1 A list of taxonomic databases. We included all databases accessed by the tools we referenced in the next section. Square brackets indicate supplementary references. Names in bold italics are taxonomic groups

Spatial Scale	Narrow Taxonomic Breadth (single taxonomic group)	Medium Taxonomic Breadth (several taxonomic groups)	Wide Taxonomic Breadth (no taxonomic restriction)
Regional	***Vascular Plants*** GermanSL (https://germansl.infinitenature.org/) [1], USDA (https://plants.usda.gov/home) [2], Vascan (https://data.canadensys.net/ipt/resource?r=vascan) [3]	***Animals*** CASD (http://zoology.especies.cn/) [4] ***All plants and fungis*** FB2020 (http://floradobrasil.jbrj.gov.br/) [5]	***No taxonomic restriction*** Dyntaxa (https://www.dyntaxa.se/) [6] EUBON (http://biodiversity.eubon.eu/web/guest/eu-bon-taxonomic-backbone) [7] FinBIF (https://laji.fi/en/) [8] NBN (https://nbn.org.uk/) [9] PESI (https://www.eu-nomen.eu/portal/index.php) [10] SP2000CN (http://sp2000.org.cn/) [11] TaiCOL (https://taibnet.sinica.edu.tw/eng) [12] TAXREF (https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref) [13] TWN (https://twnlist.aquadesk.nl/) [14]
Global	***Algae*** AlgaeBase (https://www.algaebase.org/) [15] ***Amphibians*** ASW (https://amphibiansoftheworld.amnh.org/) [16] ***Birds*** eBird/Clements (https://ebird.org/science/use-ebird-data/the-ebird-taxonomy) [17] ***Fungi*** Index Fungorum (http://www.indexfungorum.org/) [18] ***Fish*** FishBase (https://www.fishbase.in) [19] ***Mammals*** MMD (https://www.mammaldiversity.org/) [20] ***Plants*** IPNI[b] (https://www.ipni.org/) [21] LCVP (https://idiv-biodiversity.github.io/lcvplants/) [22] POWO (http://powo.science.kew.org/) [23] TPL[a,c] [24] TNRS (http://tnrs.iplantcollaborative.org/) [25] [26] Tropicos (https://tropicos.org/) [27] WCSP (https://wcsp.science.kew.org/) [28] WCVP (https://wcvp.science.kew.org/) [29] World Flora Online (http://worldfloraonline.org/) [30] ***Reptiles*** ReptileDB (https://www.reptile-database.org/) [31] ***Spiders*** World Spider Catalog (https://wsc.nmbe.ch/) [32]	***Marine organisms*** SeaLifeBase (https://sealifebase.ca/home/index.php) [33] WoRMS (https://www.marinespecies.org/) [34] ***Animals*** ZooBank[b] (http://zoobank.org/) [35]	***No taxonomic restriction*** BOLD (http://www.barcodinglife.org/) [36]COL (https://www.catalogueoflife.org/) [37] EOL (https://eol.org/) [38] GBIF (https://www.gbif.org/) [39] GNI[a] (https://index.globalnames.org/) [40] GNR[a] (https://resolver.globalnames.org/) [41] GNV (https://verifier.globalnames.org/) [42] ION (http://www.organismnames.com/) [43] ITIS (https://www.itis.gov/) [44] IUCN (https://www.iucnredlist.org/) [45] NatServe (https://explorer.natureserve.org/) [46] NCBI (https://www.ncbi.nlm.nih.gov/taxonomy) [47] Neotoma (http://neotomadb.org/) [48] OTL (https://opentreeoflife.github.io/) [49] PBDB (https://paleobiodb.org/#/) [50] Wikidata (https://www.wikidata.org/) Wikipedia (https://www.wikipedia.org/) Wikispecies (https://species.wikimedia.org/)

[a]Databases that can be considered as outdated.

[b]Rather a nomenclatural database (collection of names) than a taxonomic reference.

[c]The Plant List (https://www.theplantlist.org/), while still widely used and easy to access, has not been updated since the release of its version 1.1 in September 2013. It has been superseded notably by World Flora Online and other initiatives such as POWO and LCVP.

the original information through SQL queries tailored for each database. **`taxonlookup`** [99] provides a curated versioned taxonomy of land plants. **`taxastand`** [97] lets the user load local taxonomic data in Darwin Core format. **`vegdata`** [102] allows the download of the GermanSL database to access it offline. It also offers access to any (offline) TurboVeg database available on the user's computer within R. **`WorldFlora`** [103] lets the user access the World Flora Online database from R once it has been downloaded by the user.

Taxonomic harmonization is not limited to accessing databases and accessing lists of (un)accepted names. Several R packages offer functions to manipulate taxonomic data, parse taxonomic files or summarize taxonomic information. **`monographaR`** [105] uses standardized tables to produce a monograph on examined specimens in a paper, with associated maps and phenological diagrams. **`rgnparser`** [106] wraps within an R package a tool built by GlobalNames in the Go language that parses scientific names into components (i.e. genus, species, authority, year, etc.) efficiently. **`taxlist`** [52] and **`vegdata`** [102] provide help functions to harmonize one's own taxa list, including interaction with TurboVeg. **`taxonomyCleanr`** [78] processes and cleans taxonomic information, including a function to write taxonomy in Ecological Metadata Language (EML; Jones et al., 2006). **`taxotools`** [81, 82] contains functions to create your own taxonomic database and match it with other lists, it also parses data in Darwin Core format. **`yatah`** [111] parses taxonomic information from long strings with special characters as used in genomic data, outputs summary statistics about it and visualizes associated taxonomic hierarchy.

TABLE 2 Identified R packages useful for taxonomic name harmonization. Square brackets indicate supplementary references

Category name	Packages
Infrastructure	**taxa** [51], **taxlist** [52], **taxview**[53]
Database access (online)	**algaeClassify** [54], **AmphiNom** [55], **arakno** [56], **dyntaxa** [6], **finbif** [57], **kewr** [58], **natserv** [59], **ncbit** [60], **neotoma2** [61], **paleobioDB** [62], **plantlist** [63], **rcol** [64], **rebird** [65], **rentrez** [66], **rfishbase** [67], **rgbif** [68], **ritis** [69], **Rocc** [70], **rotl** [71], **rredlist** [72], **rreptiledb** [73], **rtaxref** [74], **SP2000** [75], **taxize** [76], [77], **taxonomyCleanr** [78], **Taxonstand** [79], [80], **taxotools** [81], [82], **taxreturn** [83], **TNRS** [84], [85], **twn** [86], **wikitaxa** [87], **worms** [88], **worrms** [89], **zbank** [90]
Database access (offline)	**AmphiNom** [55], **flattax** [91], **flora** [92], **lcvplants** [22], **mammals** [93], **ncbit** [60], **taxadb** [94], [95], **taxalight** [96], **taxastand** [97], **taxizedb** [98], **taxonlookup** [99], **taxonomizr** [100], **tpl** [101], **vegdata** [102], **WorldFlora** [103]
Data wrangling	**metacoder** [104], **monographaR** [105], **rgnparser** [106], **splister** [107], **taxastand** [97], **taxreturn** [83], **taxspell** [108], **traitdataform** [109], **vegdata** [102], **vegtable** [110], **yatah** [111]
Data visualization	**metacoder** [104], **taxview** [53]

We identified several packages that deal with taxonomic assignment from genomic data but considered them out of scope of this review (see S1 for the inclusion criteria).

2.3.1 | Tools: Lessons learned and future direction

To avoid reinventing the wheel, whenever possible, package developers should build their tools on top of existing packages and functions; however, we found little evidence for package or function reuse across packages (see lack of network links in **taxharmonizexplorer**). As an exception, **taxize** [76, 77] relies on individual packages that provide functions to access specific online databases (e.g. it relies on **rfishbase** [67] to access FishBase). The lack of dependencies between packages is inefficient from a developer standpoint and unclear for end users, due to packages performing virtually identical tasks but in a slightly different way, with different syntaxes, and different ways of handling errors. For example, **plantlist** [63], **taxadb** [94,95], **taxalight** [96], **taxize** [76,77], **taxizedb** [98], **Taxonstand** [79,80] and **tpl** [101] all access The Plant List data. While evaluating relevant tools, we identified several packages in early development. **splister** [107] and **taxastand** [97] both allow the user to match its own custom reference database, which can be useful for areas or taxa where no commonly accepted taxonomy exists. **taxreturn** [83] fetches data from BOLD and NCBI taxonomies for metabarcoding. **taxspell** [108] checks the spelling of taxon names through dictionaries that reference the most common spelling mistakes.

Our review was facilitated by the fact that the packages are deposited in standardized central repositories such as CRAN or Bioconductor. Many packages were also accessible in their last development state on open development platforms such as GitHub. Thanks to this accessibility, we identified the tools in development mentioned in the paragraph above, showing the trends in tools for taxonomy.

Of the 60 packages we included, 20 were made available through rOpenSci, many of which are central in global taxonomic harmonization such as **taxize** [76, 77]. rOpenSci is a not-for-profit organization that aims to '*[...] help develop R packages for the sciences via community driven learning, review and maintenance of contributed software in the R ecosystem*' (Boettiger et al., 2015). The fact that rOpenSci supported the development and the publicity of many tools important for taxonomy underlines how rOpenSci filled quasi an 'ecological' package niche that was not filled by traditional scientific developers. Resolving taxonomic name conflicts requires good taxonomic knowledge, which is rare outside of taxonomists. While the manipulation of online databases requires a good knowledge of web technologies, uncommon among scientists. The intersection of both is thus even rarer. Furthermore, there are few incentives to build and maintain scientific software (Jay et al., 2020; Mislan et al., 2016). The combined expertise found among rOpenSci members greatly helped advance the development and maintenance of tools to interact with taxonomic data.

Several tools we reviewed accessed data that can be considered outdated. For example, several packages access The Plant List [24], which used to be the main global taxonomic database for plants, but has not been updated since 2013 and is considered outdated by its authority (see https://www.theplantlist.org/). It refers now to the World Flora Online database as the updated successor [30]. Despite this, because of its easy access, standardized format and continuous availability it is still used by packages created long after 2013. The Plant List has gained ~1,000 citations, since 2020, (according to Google Scholar) of which very likely many used the outdated list, leading to results based on outdated knowledge. Similarly, **taxize** [76, 77] accesses both Global Names Index and Global Names Resolver, which are massive collections of other taxonomic databases (Mozzherin et al., 2021). Global Names Index has not been updated since 2018 and it has been superseded by Global Names Resolver in 2018 (D. Mozzherin, pers. comm.). Global Names Resolver has in turn been superseded by Global Names Verifier (Mozzherin, 2021), with even faster software and continuously updated data. While maintaining access to older databases is paramount to ensure the reproducibility of taxonomic name harmonization, users should check the date of last update of the resource they are accessing. The tools should explicitly warn their users when they are using outdated taxonomic databases and point them to alternative, more up-to-date, sources.

2.4 | A tool to guide users in the network of resources

To help the users navigate the complex network of tools and databases, we developed a shiny application that lets users explore the

relationships between resources and their main characteristics (date of last update, taxonomic breadth, URL, etc.). We called it **taxharmonizexplorer** and it is available as a perennial archive on Zenodo (Grenié et al., 2021) but also accessible online at: https://mgrenie.shinyapps.io/taxharmonizexplorer.

The application presents on the right side a network that links taxonomic databases and packages (Figure 3). Global databases with a wide taxonomic breath often aggregate taxonomies trying to provide a unified taxonomic backbone for all covered organisms, such as Catalogue of Life (COL) or Encyclopedia of Life (EOL) [37, 38]. The databases are connected when they rely on one another, while packages are connected when they depend on each other. Finally, packages are connected to databases when they provide access to the databases. The top left panel displays information about the node selected on the network and includes a link to the package or database website. The bottom left of the app shows a table where the user can select and search for nodes through their name, type and taxonomic group.

The dataset that backs the network is continuously improving as we are identifying the links that connect the different databases and add new R packages. The dataset is open for contributions for packages and databases that we may have missed (through GitHub or email to the corresponding author).

3 | STEPPING OUT OF THE TAXONOMIC HARMONIZATION LABYRINTH: RECOMMENDATIONS AND A COMPARISON OF EXAMPLE WORKFLOWS

In this section, we provide general guidelines and best practices to harmonize taxonomy in large biodiversity datasets to avoid common pitfalls. As an illustrative example, we harmonize taxon names from BioTIME (v. 02_04_2018, BioTIME Consortium, 2018; Dornelas et al., 2018), the largest global compilation of time-series assemblages, which includes 44,440 taxa spanning multiple taxonomic groups at broad spatial and temporal scales. BioTIME is often used (~145 citations) and is particularly interesting as it gathers information from different data sources (361 studies), which potentially leads to taxonomic inconsistencies between them. For the sake of simplicity we only focus here on birds, fishes and vascular plants in BioTIME. We detailed the process and tools used for our taxonomic harmonization (packages, including versions, specific functions and parameter values used). To achieve full reproducibility we encourage others to detail their workflow in a similar fashion, as taxonomic harmonization workflows can be highly sensitive to the exact version of the tools or data used.

We applied four different workflows (WF, Figure 4), to harmonize the taxonomy of BioTIME. WF1 and WF2 use taxon-specific databases whenever available. WF1 matches all species names against all chosen taxon-specific databases and conflicts are resolved afterwards, whereas in WF2 taxa are first assigned to higher taxonomic groups (birds, fish or vascular plants) and only then matched against relevant taxon-specific databases. WF1 and WF2 can be summarized as follows: Step 1, taxon names are preprocessed to unify writing style. Step 1.5 (only in WF2) taxa are assigned to high taxonomic groups using a multi-taxa global database. Step 2, taxon names are matched against taxon-specific databases. The other two workflows, WF3 and WF4, only use GBIF to harmonize all names. In WF3 names are preprocessed

BOX 2 The double-edged sword of 'fuzzy matching'

'Fuzzy matching' is a method to match taxon names that differ by some characters.

How it works

Similarity measures are used to quantify the discrepancy between two names (Meyer et al., 2016). For example, orthographic distance metrics measure similarity as the reciprocal of the number of characters to be modified to obtain one string from another. The obtained score indicates how close two names are to each other. The highest score name is then matched to the name of interest. One common metric is measuring single-character deletions, substitutions or insertions with the Levenshtein Distance (e.g. [95]). An alternative is the phonetic modified Damerau–Levenshtein distance weighting transpositions lower than individual character substitutions (Taxamatch; Rees, 2014).

When to use it

Fuzzy matching is useful when orthographic and spelling errors are suspected in the list of taxon names, meaning that exact matching cannot resolve them. These typos can have multiple causes, for example, transcription mistakes, wrong Latin name, differences in spelling style among taxonomic authorities, changes in the spelling style of accepted names, etc.

Risks

When two different taxa display similar names (low orthographic distance), they can be fuzzy matched to the same accepted name. If used blindly to match taxon names at broad spatial and temporal scale and taxonomic coverage, there is a relatively high risk of fuzzy matching a wrong name in a different part of the tree of life. The Interim Register of Marine and Nonmarine Genera (Rees, 2021) provides a database of possible name colliders at genus level.

Resort to fuzzy matching should only come at the end of the harmonization process to cast a bigger net of candidate names. Use of fuzzy matching should always be explicitly stated by users; tools that implement fuzzy matching by default should highlight this feature and give the option to toggle it off. Tools should also mention to what extent are results based on fuzzy matching. When resorting to fuzzy matching, sensitivity analyses should be performed using fuzzy matching scores, for example, by random sampling taxon names using matching scores as probability weights.

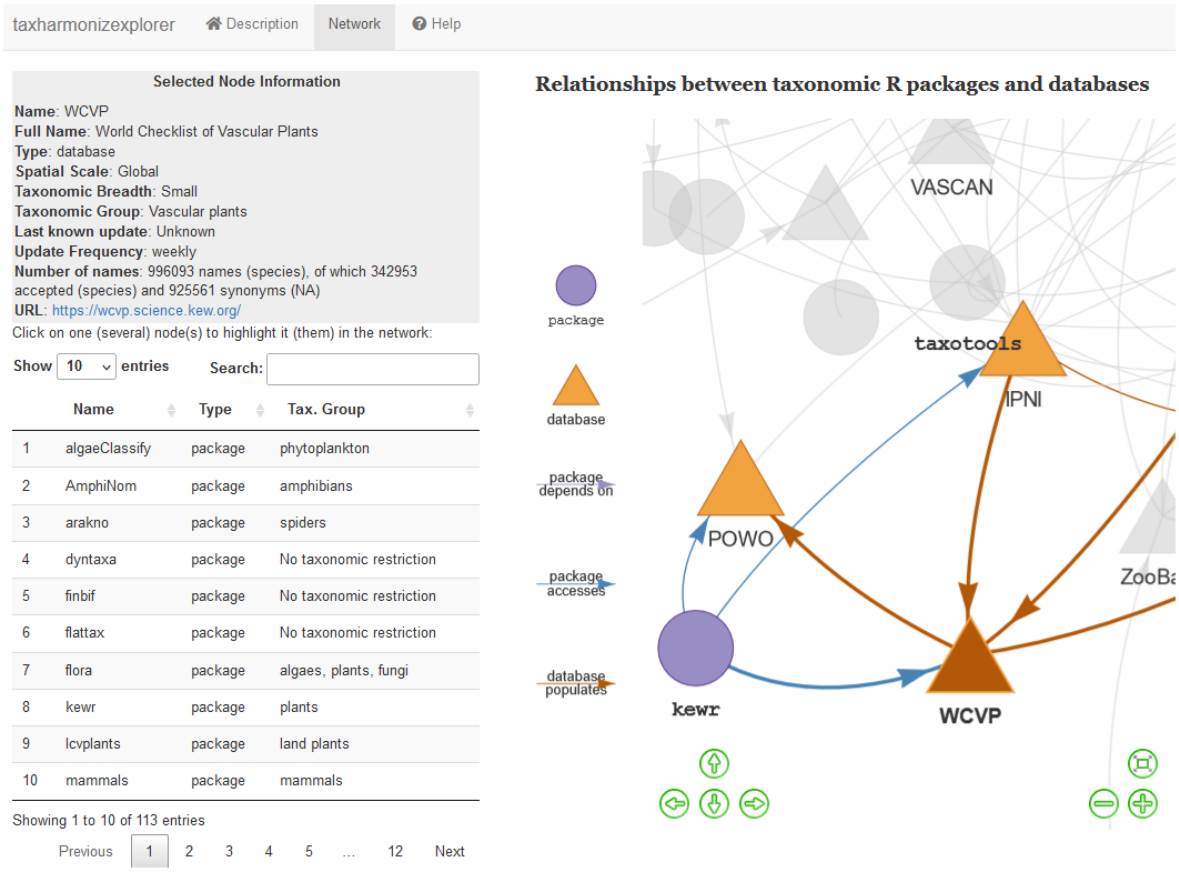

FIGURE 3 Screenshot showing the network view of **taxharmonizexplorer**. The left section shows a table of each of the nodes in the network to let the user select manually nodes of interest, the top part presents a summary of the information on the selected node in the network. The right section displays the relationships between packages (which depends on which other), between databases (how one populates another one) and between packages and databases (which packages access which databases)

(Step 1 as in WF1 and WF2), while in WF4 taxon names are passed directly from BioTIME to GBIF. We included these two workflows because they are intuitive and easy to implement and, as such, appeal particularly to non-taxonomists. We compared the performance of the different workflows by the number of identified names in the different taxonomic groups (birds, fishes and vascular plants).

4 | STEP 1: PREPROCESS NAMES (A.K.A. CLEAN/UNIFY WRITING STYLE)

Taxon names writing style can vary between sources, complicating harmonization (D. Patterson et al., 2016; Patterson et al., 2010) and becoming a source for errors. These differences arise because of the disparate use of upper and lower case, abbreviations, annotations, depictions of hybrids, authorships, etc. Removing these syntactic issues and standardizing taxon names are thus the starting point of taxonomic harmonization. To match all possible variations of a scientific name, these need to be divided into their stable (e.g. genus, species epithet and authorships) and prone-to-change elements (e.g. annotations) and then combined into only stable elements (Mozzherin et al., 2017). The result is a syntactically normalized list of names. We recommend keeping authorship, whenever possible, along the taxon names because it decreases errors. Using taxa authorship information also disambiguates between accepted and synonyms names (e.g. the IRMNG referencing binomial homonyms, Rees, 2021).

To standardize the writing style of taxon names across BioTIME, we used the function `gn_parse_tidy()` from package `rgnparser` v.0.2.0 [106]. After parsing taxon names, we only kept the two first words of each parsed name, which ideally represent the scientific binomial name of species (*Genus species*). We did not keep authorship as most names in BioTIME did not have it. We applied this step for all workflows except WF4. We found that of the 44,326 names reported in the original file, 4,734 taxa (11%) had spelling style differences, that is, species with the same binomial name after parsing. Of the remaining 39,592 unique taxon names, 6,692 were composed of only one word. We removed these taxa as our aim was to match only binomial names. Importantly, the remaining 32,900 names also contained common names and undetermined taxa with taxonomic abbreviation and keywords, for example, '*Family fam*'. As our aim was to programmatically harmonize

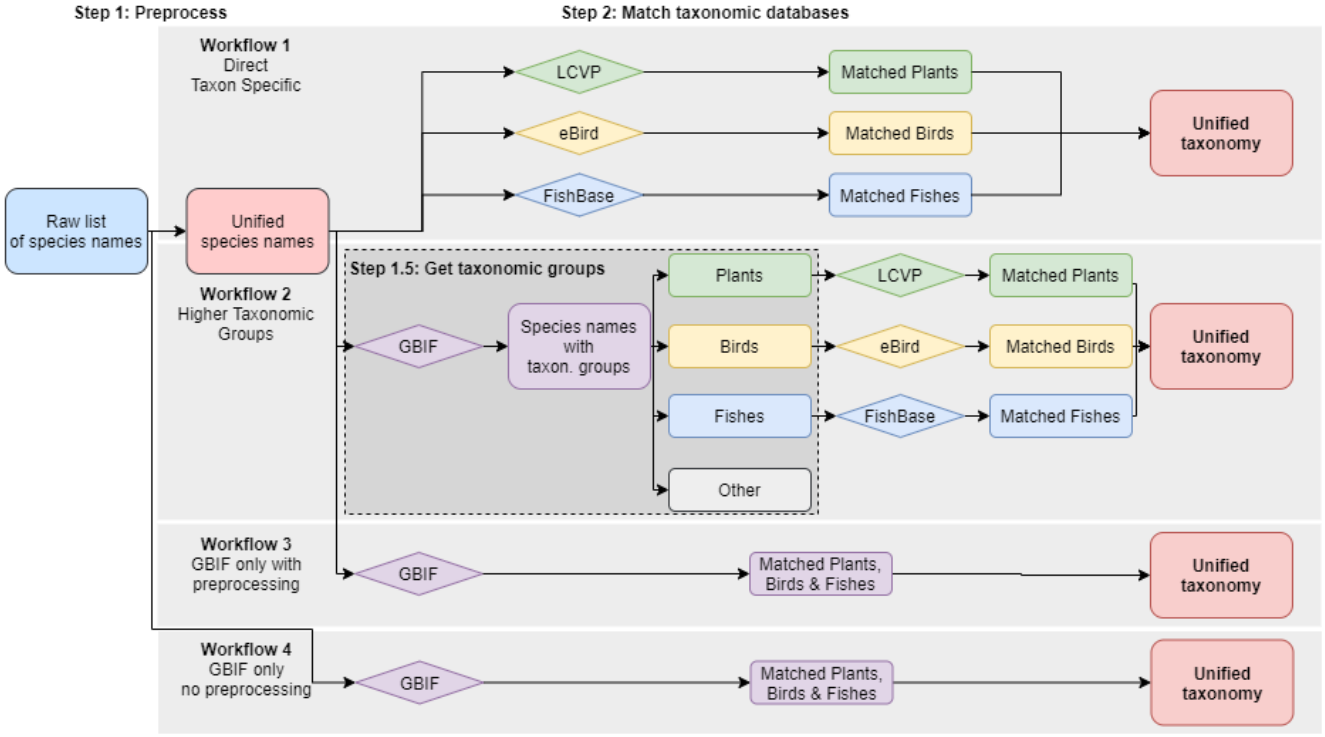

FIGURE 4 Diagram of different taxonomic harmonization workflows. The workflows differ in the number of steps they consider and the databases they leverage on. Rounded rectangles are lists of taxon names while diamonds represent taxonomic databases against which the names are matched. The different colours used at step 2 represent different taxonomic groups

taxonomy using available R packages, we kept such binomial entries as they were returned from `rgnparser` [106]; such inaccuracies will be solved in the next steps. GBIF offers an alternative name parser, which can be used through `rgbif` with the `parsenames()` function [68].

4.1 | Step 1.5: (if needed) Divide taxa in higher taxonomic groups

In WF2, taxon names are passed only to the relevant taxon-specific databases, for example, plants are matched only against a plant-specific database. Multi-taxa global databases (e.g. GBIF [39]) can provide classification to divide taxa into taxonomic groups. The potential errors should be fairly limited for higher taxonomic groups as multi-taxa databases generally offer reliable higher taxonomy (regna, phylum, class, etc.), even though some binomial names could match across different phyla (e.g. the *Aotus* genus is present in both plants and monkeys). These cases are referenced in the Interim Register of Marine and Non-Marine Genera (IRMNG, Rees, 2021).

BioTIME originally assigns taxonomic groups, but these are at the study level rather than for each species. For example, the species *Abalistes stellatus* was correctly assigned to the fish group except in one study, where it was assigned to the benthos group (to which most of the species in this study belong). To achieve maximal taxonomic accuracy, we reclassified species names into higher taxonomic groups using GBIF. We queried all names against GBIF and, based on higher clades (mostly taxonomic classes, e.g. *Sarcopterygii*, and unranked clades, e.g. *Tracheophyta*), we grouped names into three groups that could be referred to by taxon-specific databases: birds, fishes and vascular plants.

5 | STEP 2: MATCH TAXONOMIC DATABASES

The selection of databases and packages for harmonization depends on the taxonomic breadth and the spatial coverage of the species list under study (Figure 1). In general, we recommend using the most updated and taxa-specific databases. For example, if this contains species names for one taxonomic group (e.g. fishes) from a specific region (e.g. France), the most appropriate approach should be to use a taxon-specific global database (e.g. FishBase [19]) or a regional database (e.g. TAXREF [13]). For instance, if the aim is to merge the list of species names with other global datasets, then FishBase would be preferred, whereas if the goal is to provide a comprehensive list of species in France, then TAXREF can be used instead. This approach can present some caveats in specific cases. For example, if the regional studied dataset comprises non-native or aquatic species that may not be present in the regional or terrestrial focused database respectively, but would likely be present in a global database. Another example would be using fuzzy matching (Box 2) on a database of

TABLE 3 Number of species matched using each workflow. Numbers of species matched were calculated after performing Step 2 but before performing Step 3

Taxonomic group	WF1 (direct taxon specific)	WF2 (pre-assign taxonomic groups)	WF3 (GBIF with preprocessing)	WF4 (GBIF without preprocessing)
Birds	878	877	1,092	1,093
Fishes	5,123	5,122	5,491	5,496
Vascular Plants	4,435	4,412	4,647	4,649
Other	–	–	19,458	19,466

large taxonomic scope which could end up matching names in the wrong part of the tree of life (e.g. *Fucus* to *Ficus*).

The type of search, exact matching versus fuzzy matching (see details in Box 2), performed during taxon name matching can strongly affect the results. While fuzzy matching can correct misspellings, it increases the chances of mismatching errors. A way to safeguard against potential mismatches is to perform a first harmonization without fuzzy matching and then a second process (Step 3 below) including fuzzy matching algorithms only if many species names are left without matches. The use of higher taxonomic ranks can also help control that fuzzy matched names correspond to the appropriate part of the tree of life.

Finally, we strongly recommend tracking package versions and version or date of access of the taxonomic database(s) used. Tracking versions increases replicability, as different versions of packages and databases can give different results. For example, **`taxadb`** [94, 95] uses yearly snapshots of taxonomic databases, provided by the developers, to create a local database. On the other hand, **`taxize`** [76, 77] uses the last available version accessing databases online APIs.

As BioTIME has global scope, we used only global databases. The choice of taxonomic references and R packages to use was informed by our Shiny app, providing a direct example of its utility. The databases and R packages used were: eBird v.2021 and **`rebird`** v.1.2.0 for birds, FishBase v.21.04 and **`rfishbase`** v.3.1.8 for fishes, **`lcvplants`** v.1.1.1 and **LCVP** v. 1.0.4 for plants, and GBIF (accessed August 2021) and **`rgbif`** v.3.6.0 for assigning taxonomic groups in WF2 and for WF3 and WF4. We only used exact matching. Of the 32,900 parsed names, WF1 matched, as unique names, 878 birds, 5123 fishes and 4435 plants (Table 3). WF2 matched slightly less ($n = 25$) species names, caused by misclassification of higher taxonomic groups, mostly plants ($n = 23$), by GBIF (Step 1.5). WF3 and WF4 matched the highest number of species, with 795 and 803 more species than WF1 respectively. The higher number of species matched was, however, due for a large proportion to species names that were considered synonyms in WF1 and WF2 and that were thus assigned to the same accepted name by taxon-specific databases. For instance, 734 synonyms were identified in WF2, while there were only 484 in WF3. Because of this, WF3 and WF4 should be generally avoided when suitable taxon-specific databases are available.

In summary, the workflows using taxon-specific databases performed relatively similar in the number of matched names, with WF1 matching slightly more species than WF2, but requiring three times the queries needed for WF2. WF3 and WF4 were faster, easier and matched the most species names, but this was at the expense of not resolving many synonyms. Which of these workflows is best depends ultimately on the goal of the taxonomic harmonization process and users must choose what suits most the task at hand. Yet, using taxon-specific databases (WF2) to match species names already divided into high taxonomic groups seems an optimal trade-off between computational speed, programmatic complexity, accuracy and robustness of the harmonization process.

6 | STEP 3: (DO AT YOUR OWN RISK) RESOLVE UNMATCHED NAMES WITH FUZZY MATCHING

If not satisfied with the number of matches achieved through Steps 1–2, further steps can be implemented to maximize the number of matched names, looking for misspellings not corrected in Steps 1–2. These spelling errors correspond to errors associated with the wrong spelling of Latin names (e.g. the use *Breviraja caerulia* instead *Breviraja caerulea*), either due to typos or caused by using different databases (Costello et al., 2013; Patterson et al., 2016; Patterson et al., 2010). Some misspellings may have been corrected during Step 2 if species names were matched using fuzzy matching.

To correct spelling errors, algorithms are available to calculate the probability of correspondence between an input taxon name and long lists of names. Although these fuzzy searches have some risks (Box 2), functions like **`gnr_resolve()`** from package **`taxize`** have arguments that reduce the probability of mismatching. Its argument `with_context` restricts the search to a narrower taxonomical context, reducing the probability of matching homonyms from different taxonomic groups (Costello et al., 2013; Shipunov, 2011). The IRMNG database, that references colliding genera names across the tree of life, can also be used to check potential typos (Rees, 2021). As fuzzy algorithms programmatically match names based on their orthographic similarity, often without considering additional taxonomic information, extra care should be taken if step 3 is implemented, including sensitivity analyses and manual checking of matched names.

We applied this step only to WF2. We looked for misspellings across the 777 names belonging to birds, fishes and plants (from Step 1.5) that were not matched in WF2. We used the function

BOX 3 **Recommendations and best practices for robust taxonomic harmonization**

Target group	Recommendations
Users	1. Learn common principles of taxonomy to be able to develop a meaningful workflow and to understand potential outputs of the used tools 2. Use single-taxon-group databases to get the most reliable resources of taxonomic authorities 3. Use the most recently updated databases to get the most up-to-date taxonomic knowledge 4. Parse taxonomic names with specific tools to standardize their writing style (e.g. **`rgnparser`**) 5. If some data are already matched against one taxonomic database, use this database as a basis to harmonize the rest of the data to avoid mixing different taxonomic concepts and potential spelling styles 6. Flag potentially inaccurate matches (fuzzy matching, orthographic corrections) for sensitivity analyses 7. Describe your taxonomic harmonization workflow in detail, for both credit and reproducibility (e.g. which databases and packages were used?; mention the used software and database versions; and which functions and steps were taken and why?)
Package developers	1. Use updated and at best regularly maintained taxonomic databases 2. Use infrastructure packages to enforce standard methods 3. Check if other packages already provide the functionality to avoid duplication of tools, for example, start checking with **`taxharmonizexplorer`** https://mgrenie.shinyapps.io/taxharmonizexplorer/ 4. Put your package in a standardized repository (CRAN, Bioconductor) or at least in a long-term archive (Zenodo, OpenScienceFramework) 5. Contribute to other tools that provide similar functionality rather than create your own 6. Use multi-language tags (keywords), and at best short abstracts in several UN languages to make them better discoverable 7. If your tool accesses a database, always report the date of access and version of the database; if you know the database has been superseded, issue a warning to the users 8. Publish widely (targeting all end user research communities) release notes about a new tool and new major updates
Database managers	1. Provide detailed information on how the database was compiled: cite original publications 2. Use harmonized explicit grammar and spelling styles rules of the taxon names and communicate them clearly 3. Develop new databases and tools as much as possible consistent with what is already out there: do not force users to adopt a new workflow 4. Detail publicly the links between your database and other existing databases (which backbone is it using, etc.) 5. Give clear version numbers and dates to the different versions of your database and communicate it clearly to your users (what is the update frequency and how to identify it?) 6. Give clear citation guidelines of the database as a structured file such as a BibTeX file 7. Publish widely (targeting all end user research communities) release notes about a new database and major updates

`gnr_resolve()` from `taxize` v.0.9.99 and selected only the best matching names. We thus corrected spelling errors for 293 names and matched an additional 218 unique species applying again Step 2: 22 of 267 bird names, 130 of 253 fish names and 66 of 257 plant names. Despite the improvement in the number of matches, these may be wrong due to fuzzy matching and orthographic corrections. Therefore, we recommend flagging matches obtained during this step and analysing their influence on downstream analyses to account for such potential issues (Box 2), for example, by randomizing the accepted fuzzy matched names based on their score.

7 | CONCLUSION

The correct treatment of taxon names is a prerequisite for robust biodiversity research. We proposed a typology of widely used taxonomic databases and extensively reviewed R packages that work with taxonomic data. Throughout our review we identified several areas to be improved aiming for more integrated and user-friendly resources and processes to harmonize taxon names (Box 3). Many issues we came across could have been prevented by a more open and inclusive communication across research communities (e.g. ecologists, data scientists and taxonomists). For instance, rigorous and widely spread communication on important new or updated taxonomic resources or relevant tools would help prevent using outdated data or developing redundant tools either as end user or developer. We suggest publishing short release notes of taxonomic databases and tools (and major updates of them) also in target journals of the respective user communities (often possible additionally to data papers).

On a technical side, we specifically see the design and documentation of taxonomic databases and tools as a major field to improve. We

urge any researcher and potential tool developer starting with taxonomic name harmonization to do a thorough search for the most suitable (i.e. most reliable, most up-to-date) databases and existing related tools. Users should also document fully their harmonization workflow (software versions, functions, parameters and database versions) for the sake of reproducibility. Vice versa, database managers and tool developers need to make their resources discoverable for all researchers globally and describe them with all necessary meta-data (Box 3). From our review, it is clear that joint efforts between taxonomists and ecologists are strongly needed to understand how these two related fields can inform each other better, improving taxonomic harmonization on one side and making use of and improving existing tools and functions on the other. Teaching and workshops focused on taxonomic name harmonization could foster knowledge and best practices while helping connect both disciplines.

What can the broad research community do to support these services for many of us? We can start by acknowledging more this type of community service, for example, in similar ways as for reviewing papers. Developing and especially maintaining databases and tools, used by many, should be more visible and valuable than just counting citations. Scientific evaluation should fully comprise these aspects. And developers and data managers should mention these services prominently in their CVs. Funding agencies should also fund these types of projects and specifically their long-term maintenance or should support, at least, relevant existing structures, which could serve as home for these resources.

Ultimately we are convinced that joint synthesis efforts across research communities towards a comprehensive resource overviewing taxonomic databases and useful tools, including meta-data and dependencies, will help any user to discover and work with the most suitable and robust information. This resource could be hosted, for example, on platforms already offering global cross-taxa information such as COL [37]. The research community will always need taxonomic experts and initiatives working on these individual resources, but we, as users, also need more guidance on where to find them and how to use them best. Our review and the shiny app can only be a start, even hopefully a very useful one.

ACKNOWLEDGEMENTS

The authors acknowledge the support of iDiv funded by the German Research Foundation (DFG–FZT 118, 202548816). The authors thank all the participants of the iDiv taxonomic harmonization workshop that led to the ideas developed in this paper; Sulochana Swathi Kannan for her work on the characteristics of databases; Martin Freiberg, David Schellenberger-Costa, Alexander Zizka and Markus Döring for fruitful discussions about taxonomic harmonization in practice; Erik F. Y. Hom for a friendly review; Brian Maitner as well as one anonymous reviewer for their inputs on our work. We thank all the database managers who answered our data collection emails: Darrel Frost, Michael D. Guiry, Jennifer Hammock, Chantal Huijbers, Congtian Lin, Johan Liljeblad, Paul Kirk and Chris Raper. We are grateful to all the many colleagues, often not acknowledged enough, who built, curated and maintained the mentioned data and tools and continue to do so.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

AUTHORS' CONTRIBUTIONS

M.W. initiated the project; All authors conceived the ideas of the manuscript; M.G. led the writing of the manuscript with substantial contributions from J.C.-Q. and M.W.; A.S. and M.G. led the development of the companion shiny app; G.M.L.D. acquired data on databases; J.C.-Q. and E.B. developed the example workflows. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13802.

DATA AVAILABILITY STATEMENT

Code and data available on GitHub (https://github.com/Rekyt/taxo_harmonization) with a perennial archive on Zenodo 10.5281/zenodo.5121244 (Grenié et al. 2021). The repository contains the table of included packages and network links. It contains the code to run the shiny app **`taxharmonizexplorer`**. The online shiny app is available at https://mgrenie.shinyapps.io/taxharmonizexplorer/

ORCID

Matthias Grenié https://orcid.org/0000-0002-4659-7522
Emilio Berti https://orcid.org/0000-0001-9286-011X
Juan Carvajal-Quintero https://orcid.org/0000-0001-6758-8118
Alban Sagouis https://orcid.org/0000-0002-3827-1063
Marten Winter https://orcid.org/0000-0002-9593-7300

REFERENCES

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). Springer. Retrieved from https://asdar-book.org/

Boettiger, C., Chamberlain, S., Hart, E., & Ram, K. (2015). Building software, building community: Lessons from the rOpenSci project. *Journal of Open Research Software*, *3*(1), e8. https://doi.org/10.5334/jors.bu

Bortolus, A. (2008). Error cascades in the biological sciences: The unwanted consequences of using bad taxonomy in ecology. *AMBIO: A Journal of the Human Environment*, *37*(2), 114–118. https://doi.org/10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2

Chawuthai, R., Takeda, H., Wuwongse, V., & Jinbo, U. (2016). Presenting and preserving the change in taxonomic knowledge for linked data. *Semantic Web*, *7*(6), 589–616.

Consortium, B. (2018). BioTIME. *Zenodo*. https://doi.org/10.5281/zenodo.3265871

Costello, M. J. (2020). Taxonomy as the key to life. *Megataxa*, *1*(2), 105–113. https://doi.org/10.11646/megataxa.1.2.1

Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., Walter, T. C., Vanhoorne, B., Decock, W., & Appeltans, W. (2013). Global Coordination and Standardisation in Marine Biodiversity

through the World Register of Marine Species (WoRMS) and Related Databases. *PLOS ONE*, *8*(1), e51629. https://doi.org/10.1371/journal.pone.0051629

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, *85*(3), 407–417. https://doi.org/10.1111/j.1095-8312.2005.00503.x

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Dyer, E. E., Redding, D. W., & Blackburn, T. M. (2017). The global avian invasions atlas, a database of alien bird distributions worldwide. *Scientific Data*, *4*(1), 170041. https://doi.org/10.1038/sdata.2017.41

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, *289*(5488), 2312–2314. https://doi.org/10.1126/science.289.5488.2312

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*, *68*(8), 563–576. https://doi.org/10.1093/biosci/biy068

Freiberg, M., Winter, M., Gentile, A., Zizka, A., Muellner-Riehl, A. N., Weigelt, A., & Wirth, C. (2020). LCVP, The Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. *Scientific Data*, *7*(1), 416. https://doi.org/10.1038/s41597-020-00702-z

Garnett, S. T., Christidis, L., Conix, S., Costello, M. J., Zachos, F. E., Bánki, O. S., Bao, Y., Barik, S. K., Buckeridge, J. S., Hobern, D., Lien, A., Montgomery, N., Nikolaeva, S., Pyle, R. L., Thomson, S. A., van Dijk, P. P., Whalen, A., Zhang, Z.-. Q., & Thiele, K. R. (2020). Principles for creating a single authoritative list of the world's species. *PLOS Biology*, *18*(7), e3000736. https://doi.org/10.1371/journal.pbio.3000736

GBIF: The Global Biodiversity Information Facility. (2020, June 24). What is GBIF? *GBIF*. Retrieved from https://www.gbif.org/what-is-gbif

Grenié, M., Berti, E., & Sagouis, A. (2021). Rekyt/taxo_harmonization: Revised version. *Zenodo*. https://doi.org/10.5281/zenodo.5121244

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/10.1890/120103

Hassler, M. (2021). *World plants. Synonymic checklist and distribution of the world flora. Version 12.4*. Retrieved from https://www.worldplants.de

Isaac, N. J. B., Mallet, J., & Mace, G. M. (2004). Taxonomic inflation: Its influence on macroecology and conservation. *Trends in Ecology & Evolution*, *19*(9), 464–469. https://doi.org/10.1016/j.tree.2004.06.004

IUCN. (2021). *The IUCN Red List of Threatened Species. Version 2021-1*. Retrieved from https://www.iucnredlist.org

Jay, C., Haines, R., & Katz, D. S. (2020). Software must be recognised as an important output of scholarly research. *ArXiv:2011.07571 [Cs]*. Retrieved from http://arxiv.org/abs/2011.07571

Jin, J., & Yang, J. (2020). BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, *21*, e00852. https://doi.org/10.1016/j.gecco.2019.e00852

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., … Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*(9), 2648–2648. https://doi.org/10.1890/08-1494.1

Jones, M. B., Schildhauer, M. P., Reichman, O. j., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Carolina Alcázar, C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant trait database – Enhanced coverage and open access. *Global Change Biology*, *26*(1), 119–188. https://doi.org/10.1111/gcb.14904

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., Amengual, J., Basset, A., van Bodegom, P. M., Cornelissen, J. H. C., Denny, E. G., Deudero, S., Egloff, W., Elmendorf, S. C., Alonso García, E., Jones, K. D., Jones, O. R., Lavorel, S., Lear, D., … Guralnick, R. P. (2018). Towards global data products of Essential Biodiversity Variables on species traits. *Nature Ecology & Evolution*, *2*(10), 1531–1540. https://doi.org/10.1038/s41559-018-0667-3

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration – The significance of data resolution and domain. *PLOS Biology*, *17*(3), e3000183. https://doi.org/10.1371/journal.pbio.3000183

La Salle, J., Williams, K. J., & Moritz, C. (2016). Biodiversity analysis in the digital era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150337. https://doi.org/10.1098/rstb.2015.0337

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., & Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1), e02567.

Lenters, T. P., Henderson, A., Dracxler, C. M., Elias, G. A., Kamga, S. M., Couvreur, T. L. P., & Kissling, W. D. (2021). Integration and harmonization of trait data from plant individuals across heterogeneous sources. *Ecological Informatics*, *62*, 101206. https://doi.org/10.1016/j.ecoinf.2020.101206

Lepage, D., Vaidya, G., & Guralnick, R. (2014). Avibase – A database system for managing and organizing taxonomic concepts. *ZooKeys*, *420*, 117–135. https://doi.org/10.3897/zookeys.420.7089

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *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 in Ecology & Evolution*, *27*(2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016

Mislan, K. A. S., Heer, J. M., & White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology & Evolution*, *31*(1), 4–7. https://doi.org/10.1016/j.tree.2015.11.006

Mozzherin, D. (2021). gnames/gnverifier: V0.3.3. *Zenodo*. https://doi.org/10.5281/zenodo.5111543

Mozzherin, D., Myltsev, A. A., & Patterson, D. J. (2017). 'gnparser': A powerful parser for scientific names based on Parsing Expression Grammar. *BMC Bioinformatics*, *18*(1), 279. https://doi.org/10.1186/s12859-017-1663-3

Mozzherin, D., Shorthouse, D., ashipunova, & pdevries. (2021). GlobalNamesArchitecture/gni: V0.9.40 Global Names Index (no fuzzy matching), *Zenodo*. https://doi.org/10.5281/zenodo.5121908

Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1763), 20170391. https://doi.org/10.1098/rstb.2017.0391

Patterson, D., Mozzherin, D., Shorthouse, D., & Thessen, A. (2016). Challenges with using names to link digital biodiversity information. *Biodiversity Data Journal*, *4*, e8080. https://doi.org/10.3897/BDJ.4.e8080

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in Ecology & Evolution*, *25*(12), 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Pebesma, E. J. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. J., & Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13.

Rees, T. (2014). Taxamatch, an algorithm for near ('Fuzzy') matching of scientific names in taxonomic databases. *PLOS ONE*, *9*(9), e107510. https://doi.org/10.1371/journal.pone.0107510

Rees, T. (2021). *The interim register of marine and nonmarine genera*. Retrieved from https://www.irmng.org at VLIZ https://www.irmng.org

Rouhan, G., & Gaudeul, M. (2021). Plant taxonomy: A historical perspective, current challenges, and perspectives. In P. Besse (Ed.), *Molecular plant taxonomy: Methods and protocols* (pp. 1–38). Springer US. https://doi.org/10.1007/978-1-0716-0997-2_1

Schulman, L., Lahti, K., Piirainen, E., Heikkinen, M., Raitio, O., & Juslén, A. (2021). The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. *Scientific Data*, *8*(1), 137. https://doi.org/10.1038/s41597-021-00919-6

Shipunov, A. (2011). The problem of hemihomonyms and the on-line hemihomonyms database (HHDB). *Bionomina*, *4*(1), 65–72. https://doi.org/10.11646/bionomina.4.1.3

Smith, S. A., & Brown, J. W. (2018). Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, *105*(3), 302–314. https://doi.org/10.1002/ajb2.1019

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, *359*(1444), 689–698. https://doi.org/10.1098/rstb.2003.1439

Tessarolo, G., Ladle, R., Rangel, T., & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, *7*(17), 6863–6870. https://doi.org/10.1002/ece3.3259

Thomas, C. (2009). Biodiversity databases spread, prompting unification call. *Science*, *324*(5935), 1632–1633. https://doi.org/10.1126/science.324_1632

Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*, *17*(12), e3000494. https://doi.org/10.1371/journal.pbio.3000494

van Kleunen, M., Pyšek, P., Dawson, W., Essl, F., Kreft, H., Pergl, J., Weigelt, P., Stein, A., Dullinger, S., König, C., Lenzner, B., Maurel, N., Moser, D., Seebens, H., Kartesz, J., Nishino, M., Aleksanyan, A., Ansong, M., Antonova, L. A., … Winter, M. (2019). The Global Naturalized Alien Flora (GloNAF) database. *Ecology*, *100*(1), e02542. https://doi.org/10.1002/ecy.2542

Vanden Berghe, E., Coro, G., Bailly, N., Fiorellato, F., Aldemita, C., Ellenbroek, A., & Pagano, P. (2015). Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics*, *28*, 29–41. https://doi.org/10.1016/j.ecoinf.2015.05.004

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography*, *47*(1), 1–12. https://doi.org/10.1111/jbi.13633

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M., Sagouis, A., & Winter, M. (2023). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution*, *14*, *12–25*. https://doi.org/10.1111/2041-210X.13802