Matching species names across biodiversity databases: sources, tools, pitfalls and best practices for taxonomic harmonization

Matthias Grenié‡, Emilio Berti‡, Juan David Carvajal-Quintero‡, Marten Winter‡, Alban Sagouis‡

‡ German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

Corresponding author: Emilio Berti (emilio.berti@idiv.de)

Abstract

Quantity and quality of ecological data have rapidly increased in the last decades, bringing ecology into the realm of big data. Frequently, multiple databases with different origin and data characteristics are combined together to address new research questions. Taxonomic name harmonization, i.e. the process of standardize taxa names according to common sources (i.e. taxonomic backbones, TB), is necessary to properly combine multiple databases through species names. In order to be able to develop proper data matching workflows, TBs and tools using them need to be clearly and comprehensively described. But this is rarely the case. Common problems users have to deal with are: not well described taxonomic concepts behind biological databases, lack of information if TBs are actively updated and details from where the primary source of taxonomic information comes from, e.g. with secondary TBs taking information from primary TBs. In addition, software to access these TBs are not always advertised for, partly redundant, or developed following non-compatible standards, creating additional challenges for users. As a result, taxonomic name harmonization has become a major difficulty in ecological studies. Researchers face a jungle of primary and secondary TBs with a diversity of tools to access them and no clear workflow on how to practically proceed. As a consequence, it is hard for users to know which TB, tool and workflow will fit the task at hand and lead to the most robust results when combining different biological datasets.

Here, we present an overview of major TBs as well as an extensive review of R packages to access TBs, and to harmonize taxa names. We developed a Shiny app summarizing meta-data and linkages among TBs and R packages (Figs 1, 2), which users can explore to learn about general features of TBs and tools and how they are linked among each other. This is particularly helpful to help users decide on the TBs and tools that best fit the tasks and data at hand and to develop more informed workflow for taxonomic name harmonization. Finally, from our review and using the Shiny app, we were able to provide general best practice principles to harmonize taxonomic names and avoid common pitfalls.

To our knowledge, this study represents the most exhaustive review of TBs and R tools for taxonomic name harmonization. Our intuitive Shiny app can help taking practical decision when harmonizing taxonomic names across multiple datasets. Finally, our proposed workflows, based on conservative guideline principles, provide a hands-on approach for taxonomic harmonization that still focus on the quality of the end results, e.g. making sure that incompatible taxonomic hypotheses are not combined together, while maximizing the number of species correctly matched.

Keywords

taxonomy, standardization, backbone, taxonomic reference, R packages, workflow, guidelines

Presenting author

Emilio Berti

Hosting institution

German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany

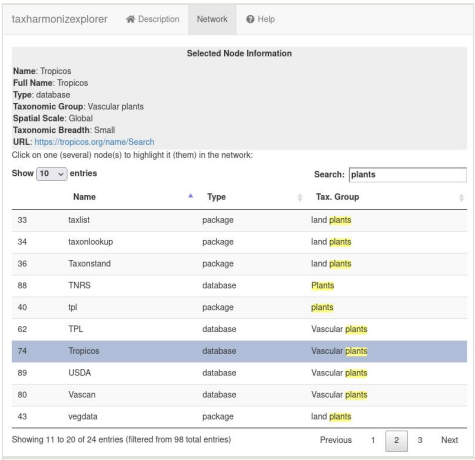

Figure 1.

First screenshot of the interactive Shiny application to explore taxonomic databases and R packages to access them. On the bottom, a table of the availalbe databases and packages is displayed with information about their taxonomic coverage. The search bar can be used to subset the taxonomic group of interest (plants in this case). On the top, information about the chosen database or package is displayed.

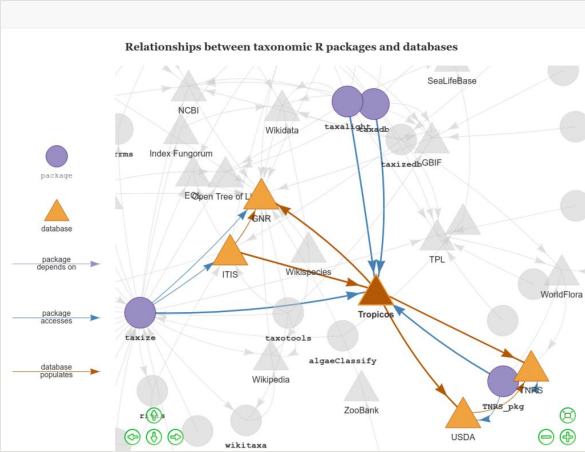

Figure 2.

Second screenshot of the interactive Shiny application to explore taxonomic databases and R packages to access them, showing the network of connections between databases and R packages. Packages accessing the taxonomic database (Tropico in this case) are displayed in blue; arrows from packages to other databases indicate that these packages can access also other taxonomic databases. Databases are displaiyed in yellow, with arrows indicating if information from a database is used to populate another database.