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

The 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 origins and data characteristics are combined to address new research questions. Taxonomic name harmonization, i.e., the process of standardizing taxon names according to common sources such as taxonomic databases (TD), is necessary to properly combine multiple databases using species names. In order to be able to develop proper data matching workflows, TDs 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: poorly described taxonomic concepts behind biological databases, lack of information when TDs are actively updated, and details regarding where the primary source of taxonomic information comes from (e.g., secondary TDs taking information from primary TDs). In addition, software to access these TDs is not always advertised, partly redundant, or developed with incompatible 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 TDs 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 TD, 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 TDs as well as an extensive review of R packages to access TDs, and to harmonize taxa names. We developed an R Shiny web application summarizing meta-data and linkages among TDs and R packages (Figs 1, 2), which users can explore to learn about general features of TDs and tools and how they are linked among one another. This is particularly helpful to assist users when deciding on the TDs and tools that best fit the tasks and data at hand and to develop more informed workflows 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 TDs and R tools for taxonomic name harmonization. Our intuitive Shiny app can help make practical decisions when harmonizing taxonomic names across multiple datasets. Finally, our proposed workflows, based on conservative guideline principles (e.g., making sure incompatible taxonomic hypotheses are not combined together), provide a hands-on approach for taxonomic harmonization, which focuses on the quality of the end results while maximizing the number of species correctly matched.

Keywords

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

Presenting author

Emilio Berti

Presented at

TDWG 2021

Hosting institution

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

Conflicts of interest

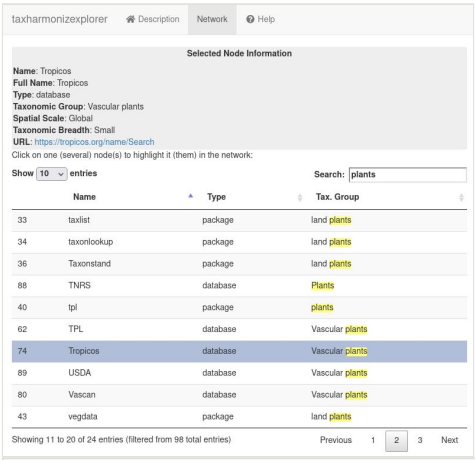

Figure 1.

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

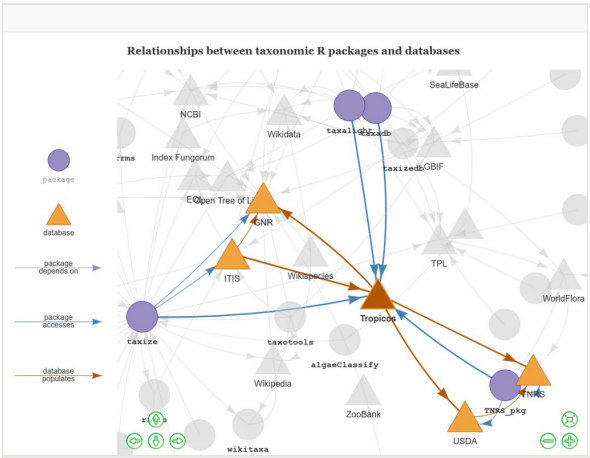

Figure 2.

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