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

Matthias Grenié1, Emilio Berti1, Juan Carvajal-Quintero1, Marten Winter1, Alban Sagouis1

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

Keywords

Taxonomy, harmonization, standardization, backbone, reference, R, workflow, guidelines.

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 of different origins with different 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 currently 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 information 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 partly redundant or developed following non-compatible standards, creating additional challenges for users. As as results, taxonomic name harmonization has become a major obstacle in ecological studies. Researchers face a jungle of primary and secondary TBs with a diversity of tools to access them and no accepted workflow on how to practically proceed. As a consequence, it is hard for users to know which TB, tool and workflow will fit the the task at hand and lead to the most robust results when combining different biological data sets.

Here, we present an overview of major TBs as well as a large quantity of R packages and other tools to access TBs, and to correctly harmonize taxa names. We developed a shiny app summarizing meta-data and linkages among TBs and R packages (available at https://mgrenie.shinyapps.io/taxharmonizexplorer/), 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 taxa 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.