**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 are combined together to adress new questions and test theories. Taxonomic harmonization, i.e. the process of standardize species names according to common sources, is necessary to properly combine multiple databases through species names. For researchers to harmonize species names, taxonomic databases and tools need to be communicated clearly and comprehensively. However, this is not currently the case, but 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 are partly redundant or developed following non-compatible standards, creating additional challenges for users. As as results, taxonomic harmonization of species names has become a major obstacle in ecological studies that seek to combine multiple databases based on species names.

Here, we overview the major taxonomic databases as well as the main R packages to access them and to harmonize species names of multiple datasets. Importantly, our findings are summarized in a GUI (available at: ...), which users can explore to decide which database and tools to use for their goal. From this review, we are able to provide guideline principles that can be used to avoid common pitfalls when harmonizing species names (Fig. 1). To provide explicit reasons and examples of such principles, we support them with a case study.

Accessible taxonomic databases can be categorized according to their taxonomic breadth and spatial scope into: regional (e.g. country level) vs global databases and taxon-specific vs multiple taxonomic groups (Fig. 2). At one extreme, regional single-taxon references (e.g. GermanSL) provide standardized taxonomic information for a specific taxon (e.g. plants) in a spatially/politically well-defined area (e.g. Germany). At the other end, global multiple-taxa references (e.g. COL) contain information for several taxonomic groups (e.g. Eucaryota) globally. One of the main differences among the two extremes is that single-taxon references are often compiled only by a small group of experts, whereas multi-taxa references combine expertise from several groups into one taxonomic backbone. The latter is a more challenging task that may generate issues when combining taxonomic hypotheses that are not inter-compatible, for instance due to unresolved synonyms and homonyms in different taxa. Because of this, we recommend (Fig. 1) to use, whenever possible, taxon-specific databases to minimize conflicts between different taxonomic hypotheses.

R packages for taxonomic harmonization can be divided into: syntax standardization tools (e.g. rgnparser), tools to access online databases (e.g. rgbif), and packages to access a local version of the database (e.g. taxadb). Strikingly, packages to achieve similar goals often operate following different standards, protocols, and syntax. This lack of standardization among tools make the task of harmonizing species names even more challenging, adding another potential source of confusion and error. This can be partially relieved if developers would follow common practices on how to access and deal with taxonomic information, e.g. by prioritizing collaborative open-source communities. Not surprisingly, the only R packages, to our knowledge, that are inter-operable are the ones developed by the rOpenSci community (e.g. ). Following such examples, we recommed (Fig. 1) to develop new tools that can operate, whenever possible, in synchrony with others tools already available.

With this new understanding of taxonomic databases and R packages, and following our guideline principles, we propose a general workflow for taxonomic harmonization. We present two flavors of the workflow (Fig. 3), one maximizing correct matches of species names (MaxAcc), the other performing only marginally worse in accuracy, but faster (MaxSpeed). To illustrate these two workflows, we applied them to harmonize species names of BioTIME database. In both MaxAcc and MaxSpeed, we pre-processed species names using rgnparser to standardize syntax, followed by parsing species names to taxon-specific databases (eBird for birds, lcvplants for plants, and fishbase for fishes). For species not belonging to one of these three groups, we parsed them to GBIF. The choice of the R packages for each step was suggested by exploring our GUI. In MaxAcc, all species names were parsed to all taxonomic databases, with conflicts, i.e. species names matched in multiple sources, considered not harmonized and excluded a-posteriori. In MaxSpeed, we used GBIF to split species into four major taxonomic groups (birds, plants, fishes, and others) and parsed each species name to the specific taxon-specific database or GBIF.

Both workflow flavors identified most (93%) of the species names in BioTIME, with only a marginal increase in accuracy (16 species out of 30,613) in MaxAcc. Note that, as MaxAcc requires four times the queries of MaxSpeed, which may thus be preferred when speed is limiting. Overall, MaxAcc and MaxSpeed give almost identical results and one of the two can be chosen depending on the specific task at hand. We also tried a third workflow where we use only the multi-taxa global reference GBIF to harmonize taxonomy, as this is likely one of the most intuitive way to harmonize species names for newcomers, especially for macroecologists already familiar with GBIF and rgbif. We found that, despite being intuitive and easy to apply, using only GBIF cannot resolve circa 620 synonyms that were identified as the same species in taxon-specific databases. Potentially, this can bias downstream analyses, e.g. by inflating species richness and turnover metrics.

To our knowledge, ours represents the most exhaustive review of taxonomic databases and R tools to harmonize species names. Our intuitive GUI can help taking practical decision when harmonizing species 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 different, incompatible taxonomic hypotheses are not combined together.