Taxonbridge: an algorithm for the creation and analysis of custom taxonomies

Werner P. Veldsman[1,✉], Giulia Campli[1], Sagane Dind[1], Valentine Rech de Laval[1], Harriet B. Drage[2], Robert M. Waterhouse[1] and Marc Robinson-Rechavi[1]

[1] University of Lausanne and Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland.
[2] Institute of Earth Sciences, University of Lausanne, Geopolis, Lausanne 1015, Switzerland.
✉ WernerPieter.Veldsman@unil.ch

INTRODUCTION

Taxonomies such as the NCBI and GBIF Backbone establish conventions by which researchers can reference and systematically compare their studies. But like all databases, taxonomies can be difficult to collate due to their heterogeneous nature (Figure 1). Researcher therefore tend to reference a single taxonomy. However, in some areas of biological studies such as Evo-Devo reliance is often placed on data from diverse fields including palaeontology and genomics, which necessitates the collation of different standard taxonomies.

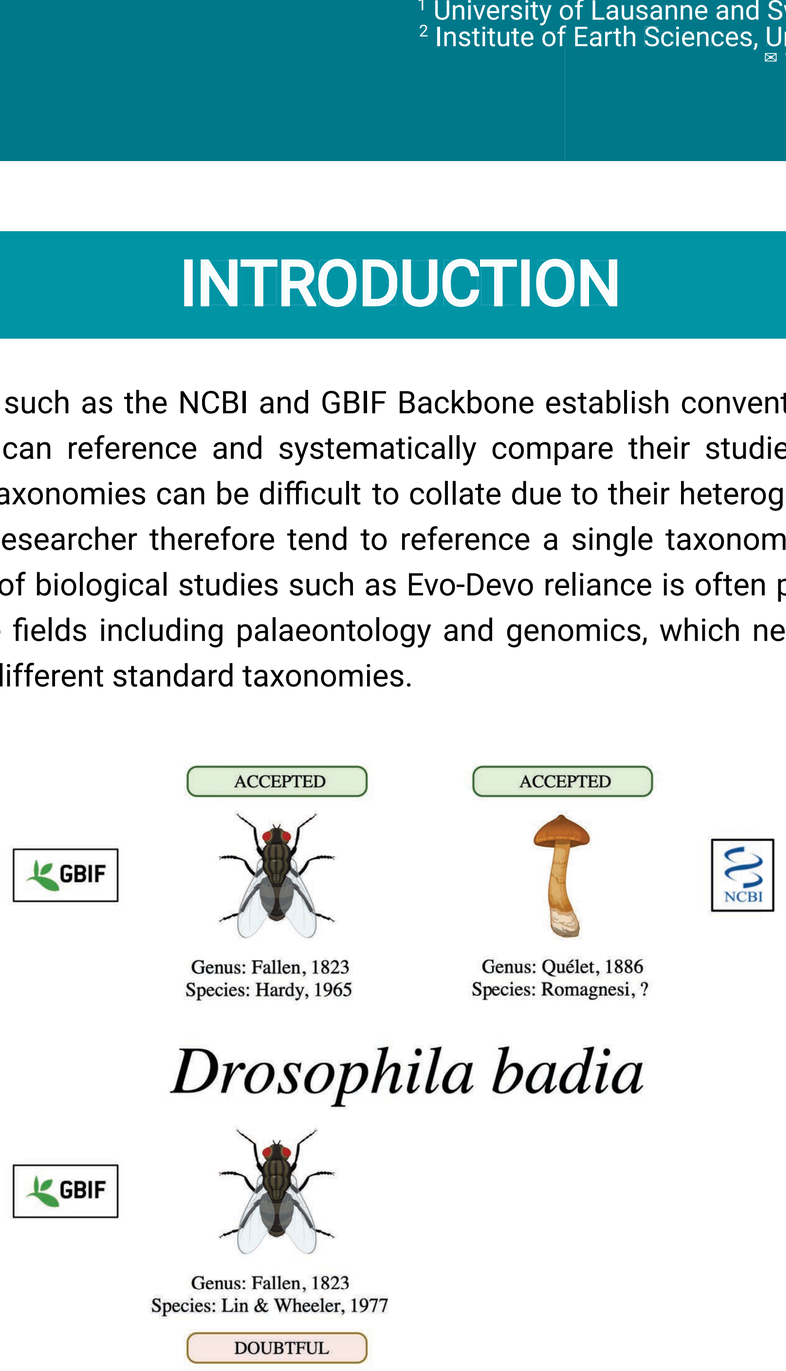

Figure 1: Ambiguity, duplication and inconsistency in taxonomic data. In this example the binomial name *Drosophila badia* is the accepted scientific name for a vinegar fly in the GBIF Backbone Taxonomy and a mushroom in the NCBI Taxonomy.

METHODS

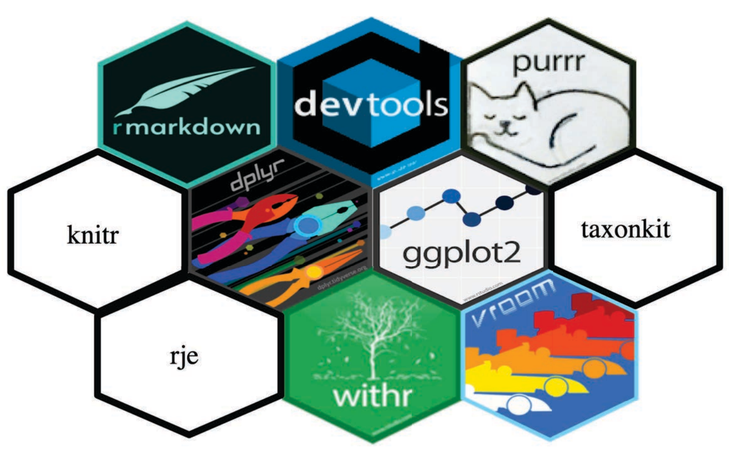

Figure 2: Taxonbridge implementation. Taxonbridge is implemented in a functional paradigm as an R package with S3 class constructs. The package uses nine other R packages as well as the Golang library Taxonkit.

CONCLUSION

Taxonbridge provides generic tools to create and analyse deduplicated, disambiguated and consistent custom taxonomies that include data on both extant species with sequence data and extinct species without sequence data.

Veldsman *et al.* (2022) *BioRxiv*, **490269**.

SCAN ME

RESULTS

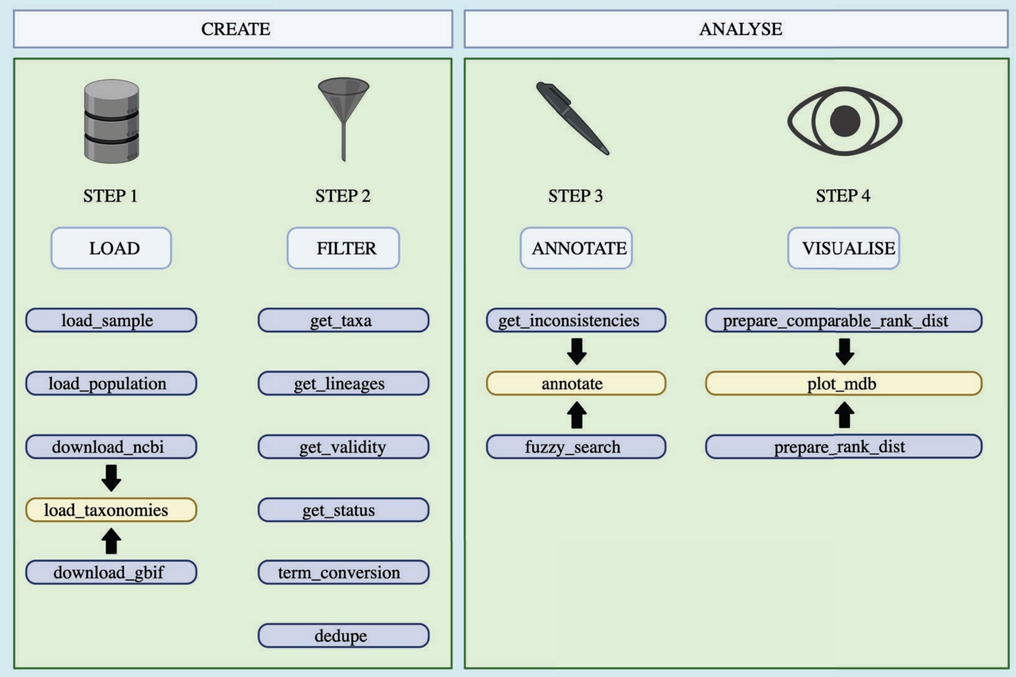

Figure 3: Workflow to create and analyse a custom taxonomy in four steps. The Taxonbridge workflow consists of seventeen methods grouped into four categories ("steps"). The four categories are designed to be used consecutively and are respectively useful for parsing, filtering, annotation, and visualisation of taxonomic data.

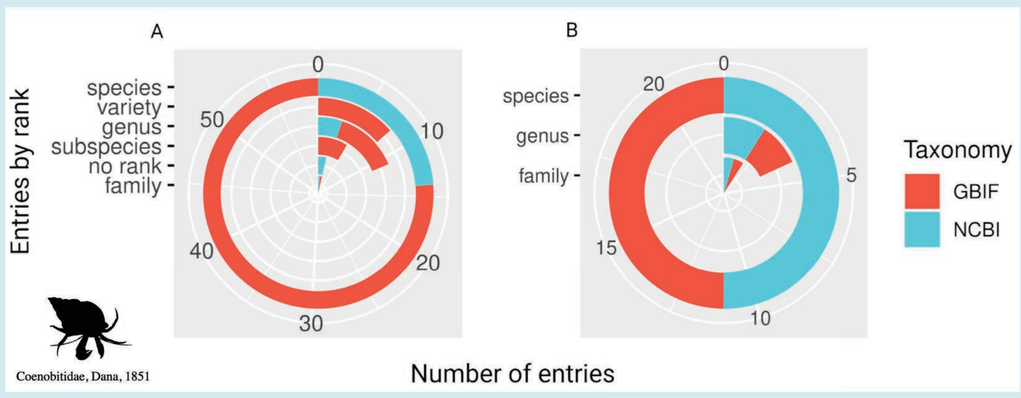

Figure 4: Visualising a custom taxonomy. These two graphs Illustrate the process of preparing a merged taxonomy for the detection of inconsistencies and ambiguity. (**A**) Raw taxonomic entries for the hermit land crab family (Coenobitidae) after merging the NCBI and GBIF Backbone taxonomies. (**B**) Entries for the same family after ensuring that both the GBIF and NCBI data contain complete lineage information, and that the GBIF data only includes entries that have a status of "accepted". This process excludes all synonyms and extinct species without sequence data in the GBIF data, and should therefore not be followed if the purpose is to retain the latter data.

		GBIF	NCBI	TAXONBRIDGE
A	Number of taxonomic entries	6,957,235	2,418,466	8,009,282
	Entries with scientific names or phrases	6,186,255	2,418,466	8,009,282
	All against all matches			696,094
B	Number of taxonomic entries after deduplication			7,391,740
	Entries with "accepted" status			3,082,679
	Entries with complete lineage data			549,635

Figure 5: Taxonbridge summary statistics. (**A**) Statistics on the results of merging the NCBI and GBIF taxonomies by full joining scientific name columns. (**B**) Statistics on filtering stepwise for duplicate entries, taxonomic statuses and complete lineage data.

ACKNOWLEDGEMENTS

Taxonbridge was developed with funding from a Swiss National Science Foundation (SNSF) Sinergia award [grant number 198691].