taxize: taxonomic search and retrieval in R

Eduard Szöcs [1] and Scott A. Chamberlain [2]

[1] University of Koblenz-Landau, [2] rOpenSci

Summary

Taxize is a R package that provides an interface to various taxonomic data sources around the web[1]. Data cleaning steps, like fixing taxonomic names, aggregating data to a specific taxonomic level, resolving ambiguous taxa or matching tables with different taxonomic resolution are crucial steps before a statistical analysis[2]. The functionality of taxize simplifies these steps and eases handling of taxonomic data in R.

Data Sources

Taxize currently provides simple and programmatic access to taxonomic data from 15 data sources around the web.

ITIS, iPlant Collaborative, IPNI, Global Invasive Species Database, Catalogue of Life, Plantminer, Tropicos, GBIF, NCBI, eol Encyclopedia of Life, gni, uBio, IUCN RED LIST, Canadensys, ThePlantList

Features

Resolve taxonomic names

We often have a list of species names and we want to know
a) if we have the most up-to-date names,
b) if our names are spelled correctly,
c) and the scientific name for a common name.

Taxize provides an interface to the EOL Global Names Resolver and Taxonomic Name Resolution Service, e.g.

```
gnr_resolve('Baetis roodani')
##   submitted_name   matched_name
## 1 Baetis roodani Baetis rhodani
```

Retrieve higher taxonomic names

Another common task is to retrieve the complete taxonomic hierarchy for a taxon. Different sources with different coverages can be used.

```
classification('Baetis rhodani', db = 'col')
##              name         rank
## 1       Animalia      Kingdom
## 2     Arthropoda       Phylum
## 3        Insecta        Class
## 4  Ephemeroptera        Order
## 5      Baetoidea Superfamily
## 6       Baetidae       Family
## 7         Baetis        Genus
## 8 Baetis rhodani      Species
```

Retrieve children taxa

One can also search in the opposite direction, i.e. search species within a genus:

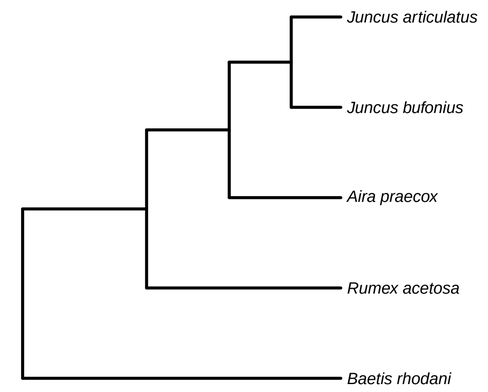
```
downstream('Baetis', db = 'col', downto = 'Species')
##      childtaxa_name childtaxa_rank
## 1   Baetis acceptus        Species
## 2  Baetis aculeatus        Species
## 3 Baetis acuminatus        Species
## 4     Baetis adonis        Species
## 5     Baetis aeneus        Species
## 6      Baetis alius        Species
## 7    Baetis alpinus        Species
```

Features (cont.)

Hierarchy trees

The taxonomic relationships between species can be displayed in hierarchy trees. These could be used for example as surrogates when phylogenetic data is scarce[3].

```
species <- c("Juncus bufonius", "Juncus articulatus",
    "Aira praecox", "Rumex acetosa", "Baetis rhodani")
hier <- classification(species, db = 'ncbi')
plot(class2tree(hier))
```



Aggregate taxa

Using the taxonomic information taxa can be easily aggregated to different levels, e.g. to study effects on different taxonomic levels. Taxize provides the `tax_agg()` function for this purpose:

```
tax_agg(dune, rank = 'family', db = 'ncbi')
```

Other use cases

restax, IUCN, Invasive, ...

```
codecodecode
```

Under the hood

taxize grabs data from the internet, formats and returns it to the user. This would not have been possible without the work of others:

Calling Servers

`httr` and `RCurl`

Parsing

`XML` and `RJSONIO`

Data manipulation

`stringr`, `plyr`, `reshape2` and `vegan`

And, of course, base-R ;)

Get involved!

taxize is currently developed collaboratively on GitHub. Feature requests, bug reports and contributions are strongly encouraged!

https://github.com/ropensci/taxize

References

[1] Scott A. Chamberlain and Eduard Szöcs. taxize: taxonomic search and retrieval in r [v2; ref status: indexed, http://f1000r.es/24v]. *F1000Research*, 2(191), 2013.

[2] Brad Boyle, Nicole Hopkins, Zhenyuan Lu, et al. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, 14(1):16, 2013.

[3] Guillaume Guénard, Peter Carsten von der Ohe, Dick de Zwart, Pierre Legendre, and Sovan Lek. Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecological Applications*, 21(8):3178–3190, 2011.

Eduard Szöcs, Mail: szoecs@uni-landau.de