taxize: taxonomic search and retrieval in R

Eduard Szöcs[1] and Scott A. Chamberlain[2]
[1]University of Koblenz-Landau, [2]Simon Fraser University

Summary

Taxize is a R package that provides an interface to various taxonomic data sources around the web. Data cleaning steps, like fixing taxonomic names, aggregating data to a specific taxonomic level or matching tables with different taxonomic resolution are crucial steps before a statistical analysis. 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 14 data sources around the web.

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

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:

```
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:

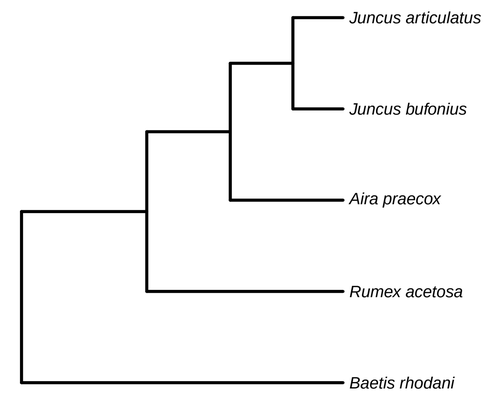
```
downstream('Baetis', db = 'col', downto = 'Species')
##      childtaxa_name childtaxa_rank
## 1   Baetis acceptus        Species
## 2  Baetis aculeatus        Species
## 3 Baetis acuminatus        Species
## 4     Baetis adonis        Species
## 5     Baetis aeneus        Species
```

Features (cont.)

Building taxonomic trees

Using this taxonomic information we can build taxonomic trees. The can be used a surrogates when phylogenetic data is scarce.

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



Aggregate data to a specific taxonomic rank

Using the taxonomic information taxa can be easily aggregated to different levels, e.g. to study effects on different taxonomic levels. This is provided via the `tax_agg()` function:

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

Match tables with different taxonomic resolution

Using the taxonomic information taxa can be easily aggregated to different levels, e.g. to study effects on different taxonomic levels. This is provided via the `tax_agg()` function:

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

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 in github. Feature requests, bug reports and contributions are strongly encouraged!

https://github.com/ropensci/taxize

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