taxize - taxonomic search and retrieval in R

Scott Chamberlain[1] and Eduard Szöcs[2]
1 Biology Department, Simon Fraser University, Canada
2 Institute for Environmental Sciences, University Koblenz-Landau, Fortstr. 7, 76829 Landau, Germany
∗ **E-mail: myrmecocystus@gmail.com**

Abstract

All species are hierarchically related to one another, and we use taxonomic names to label the nodes in this hierarchy. Taxonomic data is becoming easily available on the web, but scientists need a way to access taxonomic data on the web in a programmatic fashion that's easy and reproducible. We have developed taxize, an open-source software package (freely available from `http://cran.r-project.org/web/packages/taxize/index.html`) for the R language. taxize provides simple, programmatic access to taxonomic data for 13 data sources around the web. We discuss the need for a taxonomic toolbelt in R, and outline a suite of use cases for which taxize is ideally suited (including a full workflow as an appendix). The taxize package will facilitate open and reproducible science by allowing taxonomic data collection to be done in the open-source R platform.

Author Summary

Introduction

Evolution by natural selection has led to a hierarchical relationship among all living organisms. Thus, species are categorized using a taxonomic hierarchy, starting with the binomial species name (e.g, *Homo sapiens*), moving up to genus (*Homo*), then family (*Hominidae*), and on up to Domain (*Eukarya*). Biologists, whether studying organisms at the cell, organismal, or community level, can put their study taxa into taxonomic context, allowing them to know close and distant relatives, find relevant literature, and more. Discovering the correct taxonomic names is, unfortunately, not straightforward. Taxonomic names often change due to name changes at the generic or specific levels, lumping or splitting lower taxa (genera, species) among higher taxa (families), and name spelling changes. In addition, there is no one authoritative taxonomic names source. Instead, there are essentially competing sources (e.g., uBio, Tropicos, ITIS) that may have different accepted names for the same taxon. The goal of taxize, an R package in development, is to make all use cases having to do with retrieving and resolving taxonomic names easy and replicable.

Taxonomic data is getting easier to obtain through web interfaces (e.g., [1]). However, there are a number of good reasons to obtain taxonomic information programatically rather than through a web interface. First, if you have more than a few names to lookup on a website, it can take quite a long time to enter each name, get data, and repeat for each species. Second, programatically getting taxonomic names solves the first problem by looping over a list of names. In addition, doing taxonomic searching, etc. is reproducible. With increasing reports of irreproducibility in science [2, 3], it is extremely important to make science workflows repeatable. Science workflows can now easily incorporate text, code, and images in a single executable document [4].

The R language is the dominant language used by biologists (reference), and now has over 5,000 packages on the R package repository (CRAN) and more than 2,500 packages on other repositories to extend R. R is great for manipulating, visualizing and fitting statistical models to data. However, the key missing piece in R is the ability to get data from the internet within R. Getting data from the web will be increasingly common as more and more data gets moved to the cloud. Increasingly, data is available from

the web via API's, or application programming interfaces. These are bits of code that allow computers to talk to one another using code that is not human readable, but is machine readable. Web APIs often define a number of methods that allow users to search for a species name, or retrieve the synonyms for a species name, for example. A further strength of APIs is that they are language agnostic, meaning that data can be consumed in almost any computing context, allowing users to interact with the web API without having to know the details of the code. Whereas, if data are stored in an Excel file, for example, the file can only be opened in a few programs.

In taxize, we have written a suite of R functions that interact with many taxonomic data sources via their web APIs (Table **??**). The interface to each function is usually a simple list of species names, just as a user would do with a web API. Therefore, we hope moving from a web to R interface for taxonomic names will be relatively seamless (if one is already nominally familiar with R).

Here, we justify the need for taxize, discuss our data sources, and run through a suite of use cases to demonstrate the variety of ways that users can interact with taxize.

Results

Subsection 1

Subsection 2

Discussion

Materials and Methods

Acknowledgments

References

1. Encylopedia of Life (2013). Available: `http://eol.org/`. Accessed May 27 2013.

2. Stodden VC (2010) Reproducible research: Addressing the need for data and code sharing in computational science. Computing in Science & Engineering 12: 8–12.

3. Zimmer C (2012) A sharp rise in retractions prompts calls for reform. New York Times .

4. Xie Y (2013) Dynamic Documents with R and knitr. Chapman and Hall/CRC. URL `http://yihui.name/knitr/`.

Figure Legends

Tables