Rdataretriever: R Interface to the Data Retriever

Henry Senyondo¹, Daniel J. McGlinn², Pranita Sharma³, David J. Harris¹, Hao Ye⁴, Shawn D. Taylor¹, Jeroen Ooms⁶, Francisco Rodríguez-Sánchez⁷, Karthik Ram⁶, Apoorva Pandey⁸, Harshit Bansal⁹, Max Pohlman¹⁰, and Ethan P. White¹¹,¹²

¹ Department of Wildlife Ecology and Conservation, University of Florida  ² Department of Biology, College of Charleston  ³ North Carolina State University, Department of Computer Science  ⁴ Health Science Center Libraries, University of Florida  ⁵ USDA-ARS Jornada Experimental Range  ⁶ Berkeley Institute for Data Science, University of California, Berkeley  ⁷ Department of Agricultural Economics, Sociology, and Education, Penn State University  ⁸ Department of Electronics and Communication, Indian Institute of Technology, Roorkee  ⁹ Ajay Kumar Garg Engineering College, Ghaziabad  ¹⁰ Departamento de Biología Vegetal y Ecología, Universidad de Sevilla. ¹¹ Informatics Institute, University of Florida ¹² Biodiversity Institute, University of Florida

Summary

The rdataretriever provides an R interface to the Python-based Data Retriever software. The Data Retriever automates the multiple steps of data analysis including downloading, cleaning, standardizing, and importing datasets into a variety of relational databases and flat file formats. It also supports provenance tracking for these steps of the analysis workflow by allowing datasets to be committed at the time of installation and allowing them to be reinstalled with the same data and processing steps in the future. Finally, it supports the installation of spatial datasets into relational databases with spatial support. The rdataretriever provides an R interface to this functionality and also supports importing of datasets directly into R for immediate analysis. The system also supports the use of custom data processing routines to support complex datasets that require custom data manipulation steps. The Data Retriever and rdataretriever are focused on scientific data applications including a number of widely used, but difficult to work with datasets in the environmental sciences. This actively facilitates research on important ecological and environmental questions that would otherwise be limited.
Implementation

The main Data Retriever software is written in Python (Morris & White, 2013). The rdatalretriever allows R users to access this data processing workflow through a combination of the reticulate package (Allaire et al., 2017) and custom features developed for working in R (R Core Team, 2020). Because many R users, including the domain researchers most strongly supported by this package, are not familiar with Python and its package management systems, a strong emphasis has been placed on simplifying the installation process for this package so that it can be done entirely from R. Installation requires no direct use of Python or the command line. Detailed documentation has been developed to support users in both installation and use of the software. A Docker-based testing system and associated test suite has also been implemented to ensure that the interoperability of the R package and Python package are maintained, which is challenging due to frequent changes in reticulate and complexities in supporting cross-language functionality across multiple operating systems (Windows, Mac OS, Linux) and R programming environments (terminal-based R and RStudio).

For tabular datasets requiring relatively simple workflows the software uses the JSON based Frictionless Data tabular data metadata package standard (Frictionlessdata, 2017). For more complex data processing workflows, custom Python code is used to process the data into cleaned and standardized formats. Spatial data support is available for PostgreSQL using PostGIS. The information required for handling these datasets is based on a customized version of the Frictionless Data Geo Data schema (Frictionlessdata, 2017) that also supports raster datasets.

Acknowledgements

Development of this software was funded by the Gordon and Betty Moore Foundation’s Data-Driven Discovery Initiative through Grant GBMF4563 to Ethan White and the National Science Foundation CAREER Award 0953694 to Ethan White.

References


