The Journal of Open Source Software

Rdataretriever: R Interface to the Data Retriever

Henry Senyondo[1], **Daniel J. McGlinn**[2], **Pranita Sharma**[3], **David J. Harris**[1], **Hao Ye**[4], **Shawn D. Taylor**[1, 5], **Jeroen Ooms**[6], **Francisco Rodríguez-Sánchez**[7], **Karthik Ram**[6], **Apoorva Pandey**[8], **Harshit Bansal**[9], **Max Pohlman**[10], **and Ethan P. White**[1, 11, 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

DOI: 10.21105/joss.02800

Software

- Review
- Repository
- Archive

Editor: Frederick Boehm

Reviewers:

- @rmhogervorst
- @jsgalan

Submitted: 05 October 2020
Published: 10 December 2020

rdataretriever: An R package for downloading, cleaning, and installing publicly available datasets

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 ecology and the environmental sciences.

Statement of Need

Finding, cleaning, standardizing, and importing data into efficient data structures for modeling and visualization represents a major component of most research workflows. This is a time-consuming process for researchers even when working with relatively simple datasets. For more complex datasets, these steps can be so complex as to prevent domain experts from engaging with the dataset at all. Systems that operate like package managers for scientific data can overcome these barriers, allowing researchers to move quickly to the final steps in the data analysis workflow (visualization and modeling) and allowing domain experts to leverage the most complex data appropriate to their research questions. The rdataretriever allows R users to automatically conduct these early steps of the analysis workflow for over 200 datasets including a number of the most widely used and difficult to work with datasets in the environmental sciences including the North American Breeding Bird Survey and the Forest Inventory and Analysis datasets. 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), (Senyondo et al., 2017). The rdataretriever 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

Allaire, J., Ushey, K., Tang, Y., & Eddelbuettel, D. (2017). *Reticulate: R interface to python*. https://github.com/rstudio/reticulate

Frictionlessdata. (2017). *Specs: Specifications for frictionless data*. https://github.com/frictionlessdata/specs

Morris, B. D., & White, E. P. (2013). The EcoData retriever: Improving access to existing ecological data. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0065848

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Senyondo, H., Morris, B. D., Goel, A., Zhang, A., Narasimha, A., Negi, S., Harris, D. J., Digges, D. G., Kumar, K., Jain, A., Pal, K., Amipara, K., & White, E. P. (2017). Retriever: Data retrieval tool. *Journal of Open Source Software*, *2*(19), 451. https://doi.org/10.21105/joss.00451