rcldf: R library for reading CLDF files

15 September 2025

Summary

Cross-Linguistic Data Formats (CLDF) is a standardized data format becoming increasingly common for storing and distributing a wide range of comparative linguistic, cultural, ethnographic, geographic, and religious data. The `rcldf` package provides a lightweight *R* toolkit for loading and reading CLDF files from both local and remote sources. The package facilitates analysis with *R* by providing a number of convenience methods for converting CLDF data, and connecting to standard "reference catalogues". The aim of `rcldf` is to provide researchers with a robust toolkit for seamlessly integrating CLDF datasets into their workflows, enhancing the efficiency of linguistic and cultural research.

Statement of need

Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018) is a standardized data format designed to handle cross-linguistic and cross-cultural datasets. CLDF provides a consistent specification and package format (https://cldf.clld.org/) for common types of linguistic and cultural data from word lists, to grammatical features, and cultural traits. The aim of CLDF is to provide a simple, reliable data format to facilitate the storage, sharing, and re-use for these data.

There are currently more than 250 CLDF datasets available containing data from the world's languages and cultures including everything from catalogues of linguistic metadata, to word lists of lexical data, grammatical features, phonetic information, geographic information, and religious and cultural databases (Table 1).

Dataset	CLDF
Metadata	
Glottolog (Hammarström et al. 2020)	1
EndangeredLanguages.com	2
Lexicon	
Lexibank (Johann-Mattis List et al. 2022)	3

Dataset	CLDF
TransNewGuinea.org (Greenhill 2015)	4
Indo-European Cognate Relationships (Anderson et al. 2025))	5
Grammatical	
Grambank (Skirgård et al. 2023)	6
AUTOTYP (Bickel et al. 2023)	7
The World Atlas of Language Structures (Dryer and Haspelmath 2013)	8
The Electronic World Atlas of Varieties of English (Kortmann, Lunkenheimer, and Ehret 2020)	9
Phonetic	
Phoible (Moran and McCloy 2019)	10
Illustrations of the International Phonetic Assoc. (Baird, Evans, and Greenhill 2021)	11
Geographic	
Glottography (Ranacher et al. 2025)	12
Cultural	
D-PLACE: The Database of Places, Language, Culture, & Environment (Kirby et al. 2016)	13
Religious Data	
Pulotu: Database of Austronesian Religions (Watts et al. 2015)	14

Table 1: Examples of CLDF Datasets showing the dataset, the type of data it contains, the source, and a link to the dataset.

CLDF describes a lightweight data-package format containing one or more data tables containing tabular data in "CSV on the Web" format (CSVW) following the World Wide Web Consortium (W3C) recommendations for Tabular Data and Metadata. These tables are described and connected by a metadata file in Javascript Object Notation (JSON) format.

While there is existing functionality in R (R Core Team 2025) to read CSVW and JSON files, the `rcldf` package extends the (Gower 2022) package in a number of key ways. First, `rcldf` is metadata aware, and uses the metadata JSON file that is part of the CLDF specification to identify tables, what those tables contain, and how they are connected to each other by foreign keys. All of this information is available in a single `S3` object which incorporates each table, metadata, and source information into one namespace. Second, `rcldf` supports loading CLDF files from not just local sources but websites and remote archives as well. Third, there are functions for automatically loading the CLDF reference catalogs that describe the languages (Glottolog Hammarström et al. 2020), lexical concepts (Concepticon Johann Mattis List et al. 2025) and phonetic transcriptions Anderson et al. (2018). Finally, `rcldf` contains tools to convert the 'long' CLDF tables into 'wide' formats while resolving the foreign keys into expanded columns into one data frame for easier analysis.

Acknowledgements

I thank Robert Forkel, Sam Passmore, and Hedvig Skirgård for feedback on rcldf.

References

Anderson, Cormac, Matthew Scarborough, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, et al. 2025. "The Indo-European Cognate Relationships Dataset." *Scientific Data* 12 (1). https://doi.org/10.1038/s41597-025-05445-3.

Anderson, Cormac, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. "A Cross-Linguistic Database of Phonetic Transcription Systems." *Yearbook of the Poznan Linguistic Meeting* 4 (1): 21–53. https://doi.org/10.2478/yplm-2018-0002.

Baird, Louise, Nicholas Evans, and Simon J. Greenhill. 2021. "Blowing in the wind: Using "North Wind and the Sun" texts to sample phoneme inventories." *Journal of the International Phonetic Association* 52 (3): 453–94. https://doi.org/10.1017/s002510032000033x.

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. 2023. "The AUTOTYP Database." Zenodo. https://doi.org/10.5281/ZENODO.7976754.

Dryer, Matthew S, and Martin Haspelmath, eds. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." *Scientific Data* 5 (1): 180205. https://doi.org/10.1038/sdata.2018.205.

Gower, Robin. 2022. *Csvwr: Read and Write CSV on the Web (CSVW) Tables and Metadata*. https://doi.org/10.32614/CRAN.package.csvwr.

Greenhill, Simon J. 2015. "TransNewGuinea.org: An online database of New Guinea languages." *PLoS One* 10 (10): 1–17. https://doi.org/10.1371/journal.pone.0141563.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.15525265.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, et al. 2016. "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity." *Plos One* 11 (7): e0158391. https://doi.org/10.1371/journal.pone.0158391.

Kortmann, Bernd, Kerstin Lunkenheimer, and Katharina Ehret, eds. 2020.

eWAVE. https://ewave-atlas.org/.

List, Johann Mattis, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, eds. 2025. *CLLD Concepticon 3.4.0.* Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org/.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, Christoph Rzymski, and Robert Forkel. 2024. "CLTS. Cross-Linguistic Transcription Systems." Zenodo. https://doi.org/10.5281/ZENODO.10997741.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. "Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features." *Scientific Data* 9 (1): 316. https://doi.org/10.1038/s41597-022-01432-0.

Moran, Steven, and Daniel McCloy, eds. 2019. *PHOIBLE 2.0.* Jena: Max Planck Institute for the Science of Human History. https://phoible.org/.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ranacher, Peter, Robert Forkel, Nour Efrat-Kowalsky, Matthias Urban, Antonia Hehli, Micha Franz, Gregory Biland, et al. 2025. "A Global and Interoperable Dataset of Linguistic Distributions Derived from the Atlas of the World's Languages." *Scientific Data* 12 (1). https://doi.org/10.1038/s41597-025-05828-6.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, et al. 2023. "Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss." *Science Advances* 9 (16): eadg6175. https://doi.org/10.1126/sciadv.adg6175.

Watts, Joseph, Oliver Sheehan, Simon J. Greenhill, Stephanie Gomes-Ng, Quentin D. Atkinson, Joseph Bulbulia, and Russell D. Gray. 2015. "Pulotu: Database of Austronesian Supernatural Beliefs and Practices." *PLoS One* 10 (9): e0136783. https://doi.org/10.1371/journal.pone.0136783.