The Journal of Open Source Software

rcldf: R library for reading CLDF files

Simon J. Greenhill [1,2]

1 School of Biological Sciences, University of Auckland, New Zealand. **2** Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Germany.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 16 September 2025
Published: unpublished

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
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 & Haspelmath, 2013)	8
The Electronic World Atlas of Varieties of English (Kortmann et al., 2020)	9
Phonetic	
Phoible (Moran & McCloy, 2019)	10
Illustrations of the International Phonetic Assoc. (Baird et al., 2021)	11

Dataset	CLDF
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, C., Scarborough, M., Jocz, L., Kümmel, M. J., Jügel, T., Irslinger, B., Pooth, R., Liljegren, H., Strand, R. F., Haig, G., Geupel, U., Macak, M., Kim, R. I., Anonby, E., Pronk, T., Belyaev, O., Dewey-Findell, T. K., Boutilier, M., Freiberg, C., … Heggarty, P. (2025). The indo-european cognate relationships dataset. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-05445-3

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., & List, J.-M. (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, L., Evans, N., & Greenhill, S. J. (2021). Blowing in the wind: Using "North Wind and the Sun" texts to sample phoneme inventories. *Journal of the International Phonetic Association*, *52*(3), 453–494. https://doi.org/10.1017/s002510032000033x

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B. (2023). *The AUTOTYP database*. Zenodo. https://doi.org/10.5281/ZENODO.7976754

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., & Gray, R. D. (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, R. (2022). *Csvwr: Read and write CSV on the web (CSVW) tables and metadata*. https://doi.org/10.32614/CRAN.package.csvwr

Greenhill, S. 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, H., Forkel, R., Haspelmath, M., & Bank, S. (2020). *Glottolog 5.2*. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.15525265

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., Leehr, D., Low, B. S., McCarter, J., Divale, W., & Gavin, M. C. (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, B., Lunkenheimer, K., & Ehret, K. (Eds.). (2020). *eWAVE*. https://ewave-atlas.org/

List, Johann Mattis, Tjuka, A., Blum, F., Kučerová, A., Ugarte, C. B., Rzymski, C., Greenhill, S., & Forkel, R. (Eds.). (2025). *CLLD concepticon 3.4.0*. Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org/

List, Johann-Mattis, Anderson, C., Tresoldi, T., Rzymski, C., & Forkel, R. (2024). *CLTS. Cross-linguistic transcription systems*. Zenodo. https://doi.org/10.5281/ZENODO.10997741

List, Johann-Mattis, Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., & Gray, R. D. (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, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History. https://phoible.org/

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ranacher, P., Forkel, R., Efrat-Kowalsky, N., Urban, M., Hehli, A., Franz, M., Biland, G., Kreienbühl, A., Hermida Rodríguez, A., Azevedo, M., Romar, M., Klaussova, A., Takahashi, T., Neureiter, N., Gijn, R. van, Roose, M., Vesakoski, O., Weibel, R., Kaiping, G., & Norder, S. (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, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., … Gray, R. D. (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, J., Sheehan, O., Greenhill, S. J., Gomes-Ng, S., Atkinson, Q. D., Bulbulia, J., & Gray, R. D. (2015). Pulotu: Database of Austronesian Supernatural Beliefs and Practices. *PLoS One*, *10*(9), e0136783. https://doi.org/10.1371/journal.pone.0136783