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: 10.21105/joss.10811

Software

- Review
- Repository
- Archive

Editor: Richard Littauer

Reviewers:

- @thomaskrause
- @annagraff

Submitted: 16 September 2025
Published: 22 September 2026

Summary

Cross-Linguistic Data Formats (CLDF) is an increasingly common standardized data format 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. There are currently hundreds of CLDF datasets available, representing many of the major publicly available linguistic and cultural datasets. These datasets contain a wide variety of different types of data from the world's languages and cultures including catalogues of linguistic metadata, lexical word lists, grammatical features, phonetic information, and geographical locations and areas, as well as religious and cultural traits (Table 1).

CLDF provides a consistent specification and package format (https://cldf.clld.org/). This format is a lightweight data-package comprised of 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 that details the datatypes in the CSVW files and how they are related to each other, and to global 'reference catalogs' of linguistic and cultural information (Forkel et al., 2018).

CLDF is rapidly becoming a standard framework for storing cultural data as it provides a simple, reliable data format that facilitates the storage, sharing, and re-use for these data. The format has been used both as the native format for newly released datasets (Blum et al., 2025; e.g. Johann-Mattis List et al., 2022; Skirgård et al., 2023), as well as to republish legacy datasets after 'retrostandardizing' them for modern use (Forkel et al., 2024).

The package `rcldf` was developed to make it easy for users to access CLDF data by providing a robust metadata-aware data loader that correctly identifies column types, missing data, and other key information. To make analysis easier, `rcldf` also provides a suite of methods for manipulating CLDF data, converting to 'tidy' data formats (Wickham, 2014), and plot selected data onto maps. Finally, the package enables users to access reference catalog data to connect and aggregate data across any dataset.

A full `vignette` is provided with the R package showing an example analysis and how the package can be used.

Dataset	CLDF
Metadata	
Glottolog (Hammarström et al., 2020)	1
EndangeredLanguages.com (Holton, 2026)	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
Electronic World Atlas of Varieties of English (Kortmann et al., 2020)	9
Phonetic	
Phoible (Moran & McCloy, 2019)	10
Illustrations of the I.P.A. (Baird et al., 2021)	11
Geographic	
Glottography (Ranacher et al., 2025)	12
Cultural	
D-PLACE: 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.

State of the field

Recently a number of *R* packages have emerged to support analysis of cross-linguistic and cross-cultural data, indicating a growing need for R infrastructure for comparative analysis. For example, `lingtypology` (Moroz, 2017) provides an interface to Glottolog and a handful of other CLDF datasets, `glottospace` (Norder et al., 2022) provides geographical mapping tools, and `glottoTrees` (Round, 2026) provides the taxonomic trees from (Hammarström et al., 2020).

However, these packages have two important limitations. First, they typically bundle a fixed set of pre-processed datasets, with ad-hoc parsing code tailored to each release. This limitation means they do not generalise to the broader ecosystem of CLDF datasets. More critically, they require ongoing manual maintenance every time an upstream dataset is updated or restructured or else risk distributing stale out-of-date data.

Second, because these packages largely bypass the CSVW/JSON metadata in a CLDF package, they do not reliably parse the column types, missing-data sentinels, foreign keys, and reference catalogues that the CLDF specification uses (Forkel et al., 2018). As a result, users of these packages are restricted to whichever subset and version of data the package author chose to process, and downstream analyses inherit any silent parsing errors introduced along the way.

The rcldf package provides the missing link that solves these issues: A general-purpose package for loading and parsing any CLDF dataset. Dataset-specific metadata is incorporated and handled correctly, and the dataset is exposed in a form ready for easy analysis in R. This package should help shift the maintenance burden from per-dataset wrappers to a single shared layer, and open the full CLDF ecosystem to *R* users.

Software design

The decision to build a bespoke package for CLDF datasets rather than reuse existing packages was driven by a number of key friction points in the existing functionality in R (R Core Team, 2025). `rcldf` builds on existing R packages (e.g. Gower, 2022; Hester et al., 2023; Johnson, 2025; Ooms, 2014; Wickham et al., 2024) and extends them to reduce the risk of analysis bugs and enhance the findability, re-use, and aggregation of CLDF datasets.

1. Reducing the risk of potential bugs. While the Comma Separated Value (CSV) files that underlie CLDF are highly flexible and critically allows them to store a range of datatypes, this flexibility introduces issues with naïve CSV parsers (e.g. Ziemann et al., 2016). Many of these issues are solved by CSVW which provides a metadata file to describe the column format of each file and we use this information in the metadata to make parsing more robust.

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. The CLDF specification describes the expected values in a CSV column and this information is passed to the `readr` (Wickham et al., 2024) package to load the data correctly and avoid these common issues with datatype parsing.

Second, the CLDF specification indicates which sentinel values indicate missing data (e.g. the strings ' ', or'-','?', or 'NA') and `rcldf` automatically converts these to a standard R `NA` value. This handling reduces the risk of R treating missing data as real data, or mis-identifying strings like 'NA' as 'North America'.

Third, `rcldf` transparently handles a number of CLDF specific implementation details that can trip up unaware users: for example, to save space CLDF can compress particular media files (e.g. Nexus formatted phylogenetic data (Maddison et al., 1997) or large BibTex files (as found in e.g. Hammarström et al., 2020)). The `rcldf` package transparently handles these implementation details to make it easier for users to work with these data.

Finally, all the information for a dataset is wrapped in a single `S3` object which incorporates each table, metadata, and source information into one namespace. This namespacing makes it easier for users to keep the relevant metadata for each dataset separate from each other, reducing the risk of bugs from data from difference datasets blending into each other.

2. Enabling the findability, re-use, and aggregation of dataset. One of the key aims of CLDF is to make data Findable, Accessible, Interoperable, and Reusable i.e. 'FAIR' (Wilkinson et al., 2016). The package `rcldf` fosters this aim in three key ways.

First, `rcldf` supports loading CLDF files from not just local sources but websites and remote archives as well, especially from common repositories of scientific data like Github.com or Zenodo.com. As an added aid for enabling dataset findability, `rcldf` implements a `datasets` function which retrieves the full list of known CLDF datasets, and gives the user the required URL identifier to download and open any of these datasets immediately.

Second, `rcldf` contains 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). These reference catalogs are the key mechanisms which enable datasets to be connected to each other by, for example, aggregating the datasets that contain the Glottolog 'Glottocode' language identifier for a particular language e.g. Aotearoa New Zealand Māori maor1246, or finding datasets with a particular lexical concept of interest.

Third, CLDF datasets are normally stored in a 'long' format where multiple variables are included in the same column matching common database normalisation practices (Forkel et al., 2018). However, often in statistical analysis users prefer a 'wide' format (Wickham, 2014) where each variable is contained in its own column, and each observation is a row. To facilitate

this, `rcldf` contains tools to converts the 'long' or 'tidy' CLDF tables into 'wide' formats while resolving the foreign keys into expanded columns into one data frame.

Research impact statement

Recent years have seen a major quantitative turn towards large-scale quantitative analysis in linguistics (Kortmann, 2021) and humanities (McGillivray et al., 2022). Cross-Linguistic Data Formats has rapidly become a very common standard for producing, releasing, and distributing a rich array of cultural data. CLDF has been suggested as a best-practice approach for distributing data alongside research publications (e.g. Tresoldi et al., 2022). Publicly available datasets in this format have rapidly increased with more than 300 datasets online. However, the archival database nature of CLDF is often at conflict with how users want to interact with and analyse data, and the flexibility of CSV formats means that there are many potential bugs waiting to happen with naïve parsing and loading of data. In addition, CLDF datasets are scattered across various websites, dataset repositories, and in journal supplementary materials, requiring users to know how to find, download, unarchive, and load these datasets.

The `rcldf` package resolves these issues by, first, providing metadata-aware loading capabilities that easily load datasets from wherever they may be found, and providing a catalogue of common datasets to make this even easier. This functionality will enable users to readily access and re-use any of these datasets for their own work within the `R` language. Second, the package provides automatic functionality to identify column types, standardize missing data sentinels, and resolve compressed archives like NEXUS and BibTeX. These capabilities will ensure reproducible workflows free from parsing bugs common in ad-hoc CSV imports. Finally, `rcldf` provides a number of tools to explore datasets, from describing the metadata schema, to plotting of data onto maps, to allow users to rapidly understand the dataset they are exploring.

To make the package usable and community ready, the package is fully documented along with a 'vignette' tutorial describing how to analyse and merge datasets into a common analysis. Almost all of the code is rigorously tested with test-case coverage above 95%. The package aims to provide an almost invisible infrastructure layer, lowering the barrier for researchers to seamlessly integrate the rich array of legacy and modern datasets into R, thereby facilitating novel, reliable large-scale comparative analysis across languages, cultures, and environments. This toolkit is ready for immediate adoption by the linguistic data science community.

AI usage disclosure

ChatGPT, Claude, and Google Gemini were used to review and critique the code for potential bugs, provide documentation templates for some functions, and suggest improvements to the manuscript.

Acknowledgements

I thank Robert Forkel, Anna Graff, 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

Blum, F., Barrientos, C., Englisch, J., Forkel, R., Greenhill, S. J., Rzymski, C., & List, J.-M. (2025). Lexibank 2: Pre-computed features for large-scale lexical data. *Open Research Europe*, *5*, 126. https://doi.org/10.12688/openreseurope.20216.2

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.11040

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

Forkel, R., Swanson, D., & Moran, S. (2024). Converting legacy data to CLDF: A FAIR exit strategy for linguistic web apps. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 3978–3982). ELRA; ICCL. https://doi.org/10.63317/5oyw2vutpg6g

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

Hester, J., Wickham, H., & Bryan, J. (2023). *Vroom: Read and write rectangular text data quickly*. https://doi.org/10.32614/CRAN.package.vroom

Holton, G. (Ed.). (2026). *Catalogue of Endangered Languages*. University of Hawai'i at Mānoa. http://www.endangeredlanguages.com/

Johnson, K. F. (2025). *bib2df: Parse a BibTeX file to a data frame*. https://doi.org/10.32614/cran.package.bib2df

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. (2021). Reflecting on the quantitative turn in linguistics. *Linguistics*, *59*(5), 1207–1226. https://doi.org/10.1515/ling-2019-0046

Kortmann, B., Lunkenheimer, K., & Ehret, K. (Eds.). (2020). *eWAVE*. https://doi.org/10.5281/zenodo.17433568

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://doi.org/10.5281/zenodo.14923561

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

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An extensible file format for systematic information. *Systematic Biology*, *46*(4), 590–621. https://doi.org/10.2307/2413497

McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. *Publications*, *10*(4), 39. https://doi.org/10.3390/publications10040039

Moran, S., & McCloy, D. (Eds.). (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.2677911

Moroz, G. (2017). *Lingtypology: Easy mapping for linguistic typology*. https://doi.org/10.32614/CRAN.package.lingtypology

Norder, S., Becker, L., Skirgård, H., Arias, L., Witzlack-Makarevich, A., & van Gijn, R. (2022). Glottospace: R package for the geospatial analysis of linguistic and cultural data. *Journal of Open Source Software*, *7*(77), 4303. glottospace: R package for language mapping and geospatial analysis of linguistic and cultural data

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between JSON data and r objects. *arXiv:1403.2805 [Stat.CO]*. https://doi.org/10.32614/CRAN.package.jsonlite

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.32614/r.manuals

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

Round, E. R. (2026). *glottoTrees: Phylogenetic trees in linguistics*. https://github.com/erichround/glottoTrees

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

Tresoldi, T., Rzymski, C., Forkel, R., Greenhill, S. J., List, J.-M., & Gray, R. D. (2022). Managing Historical Linguistic Data for Computational Phylogenetics and Computer-Assisted Language Comparison. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 345–354). MIT Press. https://doi.org/10.7551/mitpress/12200.003.0033

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

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10). https://doi.org/10.18637/jss.v059.i10

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. https://doi.org/10.32614/CRAN.package.readr

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1044-7