cldf for dummies

Hedvig Skirgård

2023-09-01

CLDF for dummies

This document outlines some of the very basics of the Cross-Linguistic Data Format (CLDF). CLDF is a way of organising language data, in particular datasets with many different languages in it. The basic organisation is a set of csv-sheets (languages.csv, forms.csv etc). These documents are linked to each other in a specific way which makes it possible to combine them into a interlinked database. At the same time, because they are just plain csv-sheets so they can easily be read in by most data analysis software programs like python, R, julia etc or just regular spreadsheet programs like LibreOffice or Microsoft Excel. It is not necessary to use FileMakerPro, Microsoft Access or similar programs.

It's plain, flat and simple and you can get the hang of it in a jiffy! In this document, you will learn the very basics on how it works and how to read them in.

The data format was first published in 2018 [1] and has since then expanded to include a large amount of different datasets.

How to know if you're dealing with a CLDF-dataset

You are dealing with a CLDF-dataset if there is a folder called "cldf" with files like "languages.csv", "values.csv" and "StructureDataset-metadata.json" in it. The last file will be different depending on the type of dataset.

Here are some examples of datasets that are available in CLDF-format that you may have encountered:

- WALS (World Atlas of Language Structures)
- PHOIBLE (Phonetics Information Base and Lexicon)
- D-PLACE (Database of Places, Language, Culture and Environment)
- Glottolog
- Lexibank
- Grambank

Types of CLDF-datasets

There are five types of CLDF-datasets. They are also known as "modules".

- Wordlist (lexicon, has Forms and often Cognates)
- Structure dataset (grammar or other types of information with one value for a Parameter and a Feature, has Values)
- Dictionary (particular kind of lexicon, has Entries and Senses)
- Parallel text (collections of paragraphs of the same text in different languages, has Forms, Segments and FunctionalEquivalents)
- generic (no specific type)

Contents

Each CLDF-dataset consists of

- a set of tables
- a bibTeX-file
- a json-file

The tables are usually in csv-format and contain the data itself. The json file has information *about* the dataset, for example what type it is, what the contents are etc. The bibTeX-file contains bibliographic references for the data. Each data-point is tied to a reference by the key in the bibTeX entry.

Tables inside datasets

There are some tables that occur in most CLDF-datasets, and some that occurr only in certain types. For example, there is no table with word forms for Structure datasets - that's for wordlists and Dictionaries.

The tables have specific names in the CLDF-world. The names are different from their filenames. You can see which name is tied to which csv-file in the json. "LanguageTable" is usually found in the file languages.csv, "CodeTable" in codes.csv, "ValueTable" in values.csv, "CognateTable" in cognates.csv etc.

Tables in most CLDF-dataset Here are CLDF-tables that occur in most CLDF-datasets

- LanguageTable - list of all of the languages in the dataset. May also include things classified by Glottolog as dialects or proto-languages. Includes meta-information like longitude, language family etc.
- ParameterTable - contains a definition of the variables. For lexicon, these are the concepts, for grammar these are the features.

Wordlist also contain

- FormTable - the forms for each concept for each language
- CognateTable (not obligatory) - the cognate classification per form per concept per language

Structure datasets also contain

- ValueTable - the value for each Parameter and language
- CodeTable - The list of possible values for each parameter. For example, GB020 in Grambank is a binary feature and can take 0, 1 and ? whereas EA016 in the Ethnographic Atlas (D-PLACE) can take 1, 2 or 9. The options are exclusive of each other for each data-point.

Example: Wordlist

Below is a tiny Wordlist CLDF-dataset. This dataset contains 3 words in 2 languages. The first two tables, LanguageTable and ParameterTable contains information about the languages and parameters - in this case concepts. The FormTable contains the actual forms. For one of the concepts, one of the languages has two words and both are listed.

LanguageTable

One row = one language (or sometimes dialect or proto-language, i.e above language in a tree). The ID column uniquely identifies each language in the dataset. In other tables, the column that links to the ID column here is called "Language_ID".

ID	Name	Glottocode
15	Bintulu	bint1246
18	CHamorro	cham1312

ParameterTable

One row = one parameter. The ID column uniquely identifies each parameter in the dataset. In other tables, the column that links to the ID column here is called "Parameter_ID".

ID	Name	Concepticon_ID
144_toburn	to burn	2102
2_left	left	244

FormTable

One row = one form. The ID column uniquely identifies each form in the dataset. In other tables, the column that links to the ID column here is called "Form_ID". Here we also see Parameter_ID, which links to the column ID in the ParameterTable and Language_ID which linkes to the column ID in the LanguageTable.

ID	Parameter_ID	Language_ID	Form	Source
15-144_toburn-1	144_toburn	15	pegew	Blust-15-2005
15-144_toburn-2	144_toburn	15	tinew	Blust-15-2005
18-2_left	2_left	18	akague	38174

Source

bibTeX file called "sources.bib" One entry = one source.

```
@misc{Blust-15-2005,
    author = {Blust},
    date = {2005},
    howpublished = {personal communication}
}

@book{38174,
    author = {Topping, Donald M. and Ogo, Pedro M. and Dungca, Bernadita C.},
    address = {Honolulu},
    publisher = {The University Press of Hawaii},
    title = {Chamorro-English dictionary},
    year = {1975}
}
```

example: Wordlist - linking together

Each of the tables has a column called "ID". This column allows us to link the tables together. The column "Language_ID" in the FormTable maps onto the column "ID" in the LanguageTable, and so on.

Langugage_ID -> ID column in LanguageTable Parameter_ID -> ID column in ParameterTable Form_ID -> ID column in FormTable.

There is no column "Form_ID" inside the FormTable, it's just called ID there. Same with Parameter_ID and the ParameterTable and so on.

WARNING Some LanguageTables contain a column called "Language_ID" which is not the same as the ID column. For dialects, this column contains the Glottocode of the language that they are a dialect of. For example, Eastern Low Navarrese is a dialect of Basque. The dialect glottocode is east1470. The glottocode of the language Basque is basq1248. If a LanguageTable has the column Language_ID, it would contain basq1248 for the dialect.

Example: Structure

TBA

Advanced

If you want to learn more, go to: https://github.com/cldf/cldf/tree/master.

References

[1] Forkel, R., List, J. M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M. Hammarström, H., Haspelmath, M., Kaiping, G.A. and Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific data, 5(1), 1-10.