



HAL
open science

Using cross-linguistic data formats (CLDF) databases for historical linguistics and comparative anthropological research

Matthias Urban

► To cite this version:

Matthias Urban. Using cross-linguistic data formats (CLDF) databases for historical linguistics and comparative anthropological research. 1ère Rencontre annuelle Langues et langage à la croisée des disciplines, Sep 2024, Paris, France. <hal-04914156>

HAL Id: hal-04914156

<https://hal.science/hal-04914156v1>

Submitted on 27 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ETALAB - Open licence

Matthias Urban (Univ. of Tübingen, Germany & DDL, CNRS)

Using cross-linguistic data formats (CLDF) databases for historical linguistics and comparative anthropological research: the example of CINWA, a database of names of cultivated plants in South American Indigenous communities

In this presentation, I will illustrate how cross-linguistic data formats (CLDF, Forkel et al. 2018) can be used to construct intuitively accessible online databases of different types of cross-linguistic data, and store them according to FAIR principles. After briefly introducing some of CLDF's design principles, I will focus on presenting the CINWA database for names of cultivated plants in South American Indigenous communities (Aguilar Panchi et al. 2023, Urban et al. 2023) as an example of a CLDF database. CINWA contains names for cultivated plants in Indigenous languages of South America, which we use in ongoing work on linguistic aspects of the transition from hunting and gathering to sophisticated forms of agriculture that occurred at different times in different parts of pre-Columbian South America. I will discuss some of the decisions we made in designing the database and extracting data, highlighting in particular the value of (i) consistency in the generation of characters with diacritics in the orthographies of Indigenous languages and (ii) full documentation of the original glosses in the lexicographic sources that we consulted, and then demonstrate the functionalities of CINWA available to users through the online interface available at <http://www.cinwa.org/>

Panchi, A., Michelle, E., Lee, S. & Brodetsky, E. (2022). CINWA – Database of Cultivated plants and their names in the indigenous languages of South America. www.cinwa.org

Forkel, R., List, J-M., Greenhill, S. J., Rzymiski, Ch., 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: 180205. <https://doi.org/10.1038/sdata.2018.205>

Urban, M., Michelle, E., Panchi, A., Lee, S. & Brodetsky, E. (2023). CINWA (database of terminology for cultivated plants in indigenous languages of northwestern South America): introducing a resource for research in ethnobiology, anthropology, historical linguistics, and interdisciplinary research on the neolithic transition in South America. *Language Resources and Evaluation* 57: 1787-1817. <https://doi.org/10.1007/s10579-022-09628-x>