Converting legacy data to CLDF: A FAIR exit strategy for linguistic web apps

Robert Forkel[1], **Daniel Swanson**[2], **Steven Moran**[3,4]
Max Planck Institute for Evolutionary Anthropology[1], Indiana University[2], University of Neuchâtel[3], University of Miami[4]
Leipzig, Germany[1], Bloomington, IN, USA[2], Neuchâtel, Switzerland[3], Coral Gables, FL, USA[4]
robert_forkel@eva.mpg.de, daniel@linguistlist.org, steven.moran@unine.ch

Abstract

In the mid 2000s, there were several large-scale US National Science Foundation (NSF) grants awarded to projects aiming at developing digital infrastructure and standards for different forms of linguistics data. For example, MultiTree encoded language family trees as phylogenies in XML and LL-MAP converted detailed geographic maps of endangered languages into KML. As early stand-alone website applications, these projects allowed researchers interested in comparative linguistics to explore language genealogies and areality, respectively. However as time passed, the technologies that supported these web apps became deprecated, unsupported, and inaccessible. Here we take a future-oriented approach to digital obsolescence and illustrate how to convert legacy linguistic resources into FAIR data via the Cross-Linguistic Data Formats (CLDF). CLDF is built on the W3C recommendations *Model for Tabular Data and Metadata on the Web* and *Metadata Vocabulary for Tabular Data* developed by the CSVW (CSV on the Web) working group. Thus, each dataset is modeled as a set of tabular data files described by metadata in JSON. These standards and the tools built to validate and manipulate them provide an accessible and extensible format for converting legacy linguistic web apps into FAIR datasets.

Keywords: linguistic data, FAIR data, CLDF, converting legacy data

1. Introduction

National science foundations and other funding agencies of many countries support fundamental research in science and the humanities. With the rise of internet technologies in the 2000s, many grant funded projects leveraged browser-based applications to provide easy web access to their research data and findings. A product of their time, it was mainstream to disseminate information through cross-platform applications, such as Adobe Flash, which like many other technologies of the time, is now a long defunct software platform. Cross-platform, cross-browser, they were the go-to technology for the development of applications for a wide-ranging audience, e.g., in science, sales, and gaming. However, the lure of using these technologies, especially during the "early" web, often meant that the underlying raw data was inaccessible and instead the user interacted with the data or findings through some software front-end application. Thus, the idea of open access to research data was in principle achieved, but too often in hindsight, it was through web apps instead of standardized formats for accessing raw data. Here, we use as two case studies two large-scale NSF funded projects that focused on making available the linguistics literature on language family relationships, and detailed geographic information about languages and where they are spoken. We illustrate how legacy web apps can be transformed into FAIR data formats (Wilkinson et al., 2016). By using W3C standard data formats, we show how one can model a dataset as tabular data, with associated and standardized metadata, such that users can access the original raw data and its bibliographic provenance.

The workflow we employed for these two datasets provides a blueprint for converting legacy web apps that were supported through tax payer money, e.g., grants through various funding agencies, to FAIR data formats that leverage existing standards in linguistic data formats. Thus, not only do our case studies illustrate how to extract and convert legacy web formats into W3C standard compliant formats for long-term data accessibility, but also how existing linguistic datasets, whether FAIR or not, may be able to be converted in ways that allow them to leverage data and metadata in the CLDF universe. This allows language scientists, for example, to extract information about a language's genealogy, geography, phonology, orthography, grammar, and its speakers, e.g., demography, socio-cultural factors, to combine disparate datasets in new ways.

In Sections 2 & 3 we provide a brief overview of MultiTree and LL-MAP. In Section 4, we describe the processes and the lessons learned in converting these legacy datasets into a FAIR format.[1] Lastly, in Section 5 we summarize our work.

[1]Our reproducible workflows are hosted on GitHub: https://github.com/linguistlist.

LREC-COLING 2024, pages 3978–3982
20-25 May, 2024.

2. MultiTree

MultiTree (MT) was designed to be a library of scholarly hypotheses about the genealogical relationships among languages. It was funded through the NSF, and in 2007, it was innovative because its developers used advancements in web technologies to produce a "hyperbolic" display of language phylogenies (language family trees) in the web browser. MT was not only accessible across platforms, but it was also interactive, so that users could explore, and also compare, different language family trees (including different proposals of the same language family by different scholars). Users could also access bibliographic reference data about proposed language genealogies and they could make comments, on the website, about them. Figure 1 is a screenshot from the now defunct website.

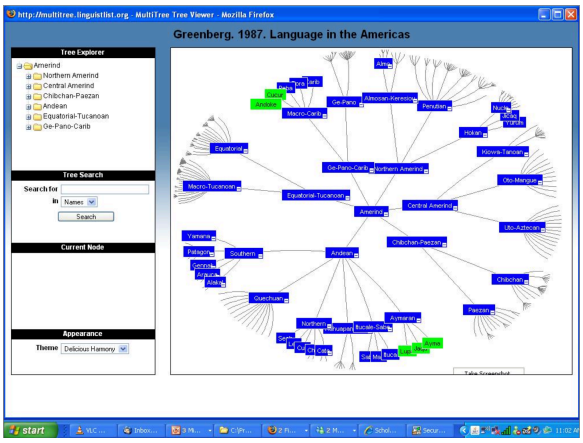

Figure 1: Multitree in the late 2000s.

The goal to collect and share historical linguistics data was successful in the sense that it made openly available scholarly information about hypotheses about the relationship between languages, as published in scientific articles. However, like most (if not all) grant awarded projects, funding to maintain the MT web app ran out and the goal to bring it to a state of completion and keep it up-to-date thereafter, did not happen.[2]

Fortunately, the developers of MT – including the Linguist List and a board of comparative historical linguists – encoded language genealogies and their scholarly metadata in a custom-made self-describing XML format. We have made these raw data available online.[3] Furthermore, as we describe in Section 4, we have created a FAIR exit strategy for the MT data as a whole.

One shortcoming of the MT data is that although they are intended to be faithful representations of the publications from which they are derived, it can be difficult to decipher a scholar's findings in graphical form. For example, not all language classifications are described as tree-like structures. Thus, the original editors of MT added comments to clarify or disambiguate their interpretations of the original resource. This is however not a new problem in typological database development.

Another shortcoming of the MT data is that its trees are not dated, i.e., they are purely topological trees with no meaningful branch lengths. Advances in computational phylogenetics, partially driven through explorations on linguistics data of increasing quality, have come a long way since MT had its day, nearly 20 years ago. Nevertheless, at the very least MT provides computationally accessible hypotheses of published language genealogies that scholars may or may not accept. In fact, the goal of MT was to collect as many hypotheses of language relatedness as possible – including phylogenies that may not be broadly accepted by the academic community, and the ever changing knowledge of the historical linguistics landscape.

With the wealth of new analyses and methods, in particular bayesian phylogenetic methods (Bowern, 2018; Cathcart, 2018) that can be used to test classification hypotheses with data as reflected today, the MT dataset may see a second spring of computational re-use. It may also gain new relevance by providing historic context for "modern" language phylogenies as collected for example in the Phlorest project (Forkel and Greenhill, 2023).

3. LL-MAP

The Language and Location: Map Annotation Project (LL-MAP) was also an NSF funded project from the mid-2000s, which was designed to integrate information about languages using a Geographical Information System (GIS). The combination of geographical, political, demographic, zoological, botanical and archaeological data, together with language maps, afforded exploratory analysis between linguistic and non-linguistic factors. For example, because LL-MAP data were linked to language family trees in MT, users could explore factors of linguistic diffusion, i.e., which properties of language were more likely to be borrowed in different geographic areas under different non-linguistic pressures. The use of GIS together with language maps, both historic and modern, also allowed users to geographically explore variables including political boundaries, demographics, climate, vegetation and wildlife, all of which are known factors reflecting population movement and interaction. Together with cultural information,

[2] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1519050

[3] https://github.com/linguistlist/multitree/tree/main/raw

including data about economics, ethnicity, and religious practices, language scientists could make hypothesis about why languages are where they are today.

LL-MAP had two interfaces. The first was over Google Maps and the second used GIS software from the Environmental Systems Research Institute (ESRI). The former provided languages as points on an interactive map; and the the latter had additional functionality, e.g., the ability to undertake spatial analysis. An example of working with LL-MAP in the browser is given in Figure 2.

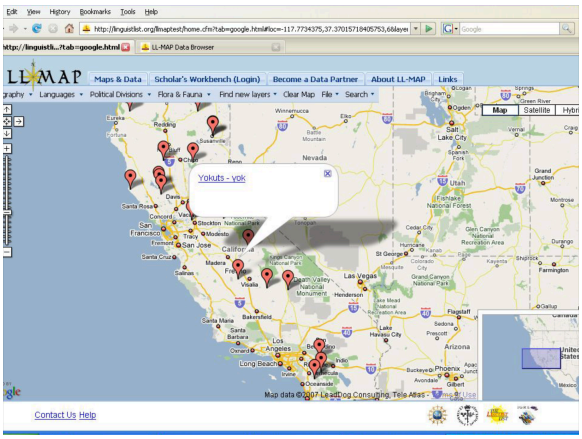

Figure 2: LL-MAP in the late 2000s.

Although the aims of LL-MAP included providing a free open source for education and research purposes, e.g., for teaching students about linguistic and cultural diversity (and by providing a framework for collaboration between linguists, historians, archaeologists, ethnographers and geneticists), LL-MAP, like MT, became a defunct web application. It was also eventually abandoned due to a lack of financial support.

4. Converting legacy data

The Cross Linguistic Data Formats (CLDF; Forkel et al. 2018) provide digital infrastructure to create FAIR cross-linguistic datasets: A data serialization format that allows not only syntactic interoperability (by being built on CSVW[4]) but also semantic interoperability by integrating a linguistically informed ontology that links to domain-specific reference catalogs, most notably Glottolog via Glottocodes (Forkel and Hammarström, 2022). Furthermore, with a well-defined conversion from CSVW to RDF[5], CLDF can be integrated in RDF-based platforms like the Semantic Web.

Thus, the task of converting legacy data – such as MultiTree and LL-MAP – to interoperable datasets is reduced to:

- Modeling the data as a set of interrelated tables
- Using CLDF components and terms where appropriate
- Implementing a conversion pipeline

Thanks to the tooling available in the CLDF ecosystem, the latter task can be achieved by using the `cldfbench` package (Forkel and List, 2020). `cldfbench` is particularly suited here because it is designed to support a workflow where:

- Legacy data serves as input
- They are enriched with expert annotations by an editorial team (e.g., adding language mappings)
- They are converted to a CLDF dataset with rich metadata such that provenance information is made transparent

4.1. MultiTree

Language trees are the core entity in the CLDF version of the MT data (LinguistList, 2023b); see for example Figure 3. Starting with CLDF 1.2, the TreeTable component encodes a standardized and transparent format of phylogenetic relationships in a tree data structure. These tree structures encode in MT hypotheses about the relationships between languages, which were extracted from the historical linguistics literature. Therefore, these phylogenies are linked (via a many-to-many relation) to entities in the CLDF SourceTable. Since MT data may carry annotations per node in a language tree, a custom association table between trees and languages was introduced (`nodes.csv`). An example of per-node annotations are the labels for nodes used in the original publications. The "normalised" trees in a CLDF dataset must have identifiers of items in the LanguageTable as label, thus source labels must be stored elsewhere. The actual tree representation in Newick format is stored in a Nexus file, linked from the TreeTable. This allows interoperability with off-the-shelf phylogenetics software to manipulate or visualize trees.

The code implementing the conversion from legacy XML data to CLDF is available on GitHub[6]. With roughly 400 lines of code we can make sense

[4] https://www.w3.org/2013/csvw/wiki/Main_Page

[5] https://www.w3.org/TR/csv2rdf/

[6] https://github.com/linguistlist/multitree/blob/main/cldfbench_multitree.py

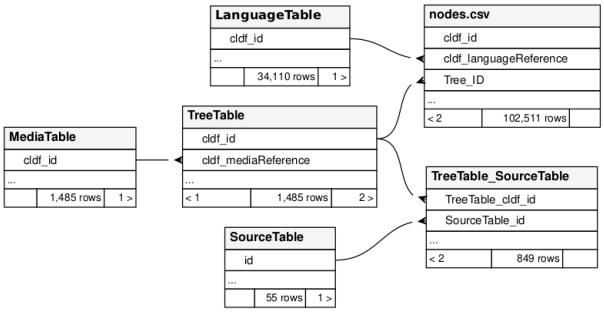

Figure 3: Entity-relationship diagram of the MultiTree data model.

of the bulk of the legacy data, linking 1485 language phylogenies extracted from 55 sources to 34,110 languoids (languages, dialects or language groups), more than half of which could be mapped to Glottolog.

As proof-of-concept, illustrating the adequacy and completeness of the CLDF dataset, we implemented a `cldfbench` command to inspect a phylogeny from the dataset from the command-line[7] (see Figure 4). The major components of the MT data – source reference, languoid identifiers and node metadata – can be accessed and re-assembled.

```
$ cldfbench multitree.show 12067
Xincan: Campbell 1997
Source: Campbell (1997) [166]
                     /-qco_jut
           /-qco-----|
           |         \-qco_yup
--xink----+-qhq
           |-qda
           \-2df
Label    Name            Type       Status
-------  --------------  ---------  --------
xink     Xincan          Subgroup
qco      Yupiltepeque    Language   Extinct
qco_jut  Jutiapa         Dialect
...
```

Figure 4: Result of the `multitree.show` command.

4.2. LL-MAP

CLDF 1.1 provides a MediaTable which is used to associate LL-MAP datasets – the Contributions table in the LL-MAP CLDF data (LinguistList, 2023a) – with:

- Scans of legacy maps
- Born-digital maps in image formats
- GIS maps in GeoJSON format

Again, maps are linked to sources (the literature from which they were extracted) and to language (sub)groups to which the geographic information pertains, modeled using standard CLDF components (see Figure 5).

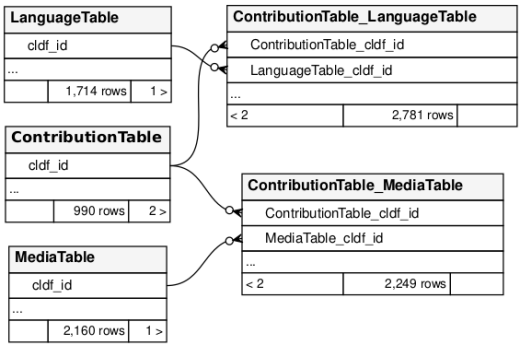

Figure 5: Entity-relationship diagram of the LL-MAP data model.

LL-MAP contributions are also often related to other maps in LL-MAP and/or to phylogenies in MultiTree. These relations have been encoded originally as "commented references" in HTML. CLDF Markdown – a CLDF extension specifying references to CLDF data objects in Markdown text – is used to keep both the relations as well as the text of this data.

The code implementing the conversion from legacy data to CLDF is available on GitHub[8]. The resulting CLDF dataset contains 990 "mapping projects", linking 2160 media files to 1714 languoids. Due to filesize constraints on GitHub as well as on Zenodo we have not made the roughly 9GB of binary media files available yet, but CLDF provides sufficient functionality to link to files provided in separate Zenodo deposits, which we will upload in the near future.

Again, we can illustrate the adequacy of the dataset via a `cldfbench` command that assembles information about one LL-MAP contribution in an HTML page which can be viewed with a browser.

4.3. Lessons learned

With the advent of content management systems, e.g., WordPress, a "through-the-web" or "editing data in the live database" paradigm for data curation became popular. This paradigm was followed for the LL-MAP and MT data because it had several advantages: roping in collaborators was easy and custom tools to edit somewhat complex GIS

[7] https://github.com/linguistlist/multitree/blob/main/USAGE.md

[8] https://github.com/linguistlist/LL-map/blob/main/cldfbench_llmap.py

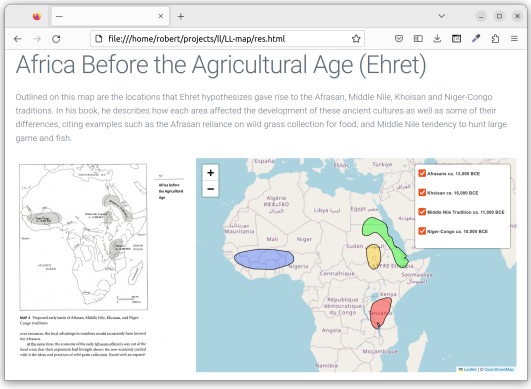

Figure 6: Result of the `llmap.show` command.

data could be provided with a web based user interface.

In this model of data curation, the web application is also used as a data publication platform. This creates multiple problems, though, once the application can no longer be maintained. Bundling of data and administrative metadata of the application (sessions, user data, authentication and authorization data) makes it impossible to "just publish data dumps". It also complicates extracting publishable data because application code, to interpret data such as workflow status information, must essentially be emulated. Lastly, while HTML is a versatile package format for the visual presentation of semi-structured data, it is a tough starting point for post-hoc structuring of data. Our conversion of legacy data to FAIR formats addresses these issues, reclaiming the two resources for the scientific record and for future reuse.

5. Summary

We have shown how to convert datasets in legacy linguistic web applications into accessible and extensible data formats via a FAIR "exit strategy". Our two case studies provide more evidence that web applications are not a viable long-term data publishing solution. But the CLDF data we managed to extract, and their re-use scenarios outlined here, show that legacy scholarly work can be saved from obsolescence and in fact provide "today's" data collections with useful context for future research.

6. Acknowledgements

LL-MAP and Multitree were supported by the US National Science Foundation. SM was supported by the Swiss National Science Foundation (PCEFP1_186841). RF was supported by the Department for Linguistic and Cultural Evolution of the MPI for Evolutionary Anthropology.

7. Bibliographical References

Claire Bowern. 2018. Computational phylogenetics. *Annual Review of Linguistics*, 4:281–296.

Chundra Aroor Cathcart. 2018. Modeling linguistic evolution: A look under the hood. *Linguistics Vanguard*, 4(1):20170043.

Robert Forkel and Simon Greenhill. 2023. Phlorest - seeing the forest and not just trees. In *International Conference on Historical Linguistics 2023*.

Robert Forkel and Harald Hammarström. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6):917–924.

Robert Forkel and Johann-Mattis List. 2020. Cldfbench. give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, pages 6997–7004, Luxembourg. European Language Resources Association (ELRA).

Robert Forkel, 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(180205):1–10.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9.

8. Language Resource References

LinguistList. 2023a. *LL-MAP: Language and Location - A Map Annotation Project*. Zenodo, 0.1. [link].

LinguistList. 2023b. *MultiTree: A digital library of language relationships*. Zenodo, 0.1. [link].