Cross-linguistic Data Formats, advancing data sharing and reuse in comparative linguistics

Robert Forkel[1]*, Johann-Mattis List[1]*, Simon J. Greenhill[12],
Sebastian Bank[1], Christoph Rzymski[1], Michael Cysouw[3],
Harald Hammarström[14], Martin Haspelmath[15], Russell D. Gray[1]

January 8, 2018

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena; [2]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra; [3]Research Center Deutscher Sprachatlas, Philipps University Marburg, Marburg; [4]Department of Linguistics and Philology, Uppsala University, Uppsala; [5]Department of English Studies, Leipzig University, Leipzig; *corresponding authors: Robert Forkel (forkel@shh.mpg.de) and Johann-Mattis List (list@shh.mpg.de)

The amount of available digital data for the languages of the world is constantly increasing. Unfortunately, most of the digital data is provided in a large variety of formats and therefore not amenable for comparison and reuse. The Cross-Linguistic Data Formats initiative proposes new standards for the four basic types of data in historical and typological language comparison (word lists, structural datasets, parallel texts, and dictionaries). The new specification for cross-linguistic data formats comes along with a software package for validation and manipulation, a basic ontology which links to more general frameworks, and usage examples of best practice.

Introduction

The last two decades have witnessed a dramatic increase in language data, not only in form of monolingual resources for the world's biggest languages, but also in form of *cross-linguistic datasets* which try to cover as many of the world's languages as possible. Creating datasets in linguistics is currently *en vogue*, and apart from traditional ways of linguistic data collection in form of etymological dictionaries, user dictionaries, and grammatical surveys, data is now being published in form of *online databases*[1] and *online appendices or supplements to published papers*, addressing

[1] The most complete list of such databases is curated at `http://languagegoldmine.com/`.

topics as diverse as cross-linguistic lexical associations [38], etymologically annotated word lists for large language families like Austronesian [20, 5] and Indo-European [13], inventories of speech sounds [51], or grammatical features compared across a large sample of the world's languages [12]. Along with the increase in the amount of data, there is also an increased interest in linguistic topics and studies. While for non-linguists linguistics was always a very specialized subject with little use beyond teaching and translation, scholars from non-linguistic disciplines such as archaeology, biology, and psychology are now beginning to realize that linguistic data has much more to offer than previously thought. Thanks to modern quantitative approaches, often inspired by evolutionary biology and bioinformatics, scholars can now investigate when particular language families started to diverge [6, 7], how robustly languages are transmitted [4], or which forces drive their change [53]. However, the use of language data is not limited to questions of language evolution alone. Cross-linguistic data have proven useful to detect structures which are universal across human populations [68, 38] or depend on environmental factors [45], specific evolutionary settings [28], or cultural habits [17].

Despite this gold rush in the creation of linguistic databases and their application reflected in a large number of scholarly publications and an increased interest in the media, linguistic data are still far away from being "FAIR" in the sense of Wilkinson et al. [64]: Findable, Accessible, Interoperable, and Reusable. It is still very difficult to *find* particular datasets, since linguistic journals do not have a policy on supplementary data or even refuse to host data on their servers. It is also often difficult to *access* data, and many papers which are based on original data are still being published without the data and authors may even refuse to share their data upon request [61, 59]. Due to idiosyncratic formats, linguistic datasets also often lack *interoperability* and are therefore not *reusable*.

Despite the large diversity of human languages, linguistic data can usually be represented by very simple data types which are easy to store and manipulate. Word lists and grammatical surveys, for example, can usually be represented by triples of *language*, *feature*, and *value*. The simplicity, however, is deceptive, as there are too many degrees of freedom which render most of the data that has been produced in the past *incomparable*. Due to the apparently simple structure, scholars rarely bother with proper serialization, assuming that their data will be easy to reuse. Although there are recent and long-standing standardization efforts, like the establishment of the *International Phonetic Alphabet* (IPA) as a unified alphabet for phonetic transcription [30], which goes back to the end of the 19th century century [32], or the more recent publication of reference catalogues for languages [24, 36] and word meanings [40], linguists often ignore these standards when compiling their datasets, or assume that they conform to the standards without verification.

While certain standards, such as the usage of unified transcription systems, are generally agreed upon but often ignored in practice, other types of linguistic data come along with a multitude of different standards which drastically exacerbate data interoperability. Thus, a large number of standards for *dictionaries* has been proposed so far (see Table 1), but no standard has found wide adoption as these are often driven primarily by computational requirements rather than by researcher requirements.

At the same time, funding agencies such as the *German Academic Research Council* emphasize that 'the use of open or openly documented formats [to enable] free public access to data deriving from research should be the norm' [21], mirroring the European Research Council's guidelines for *Open Access to Research Data* in the *Horizon 2020* programme [22]. Additionally, the *Research*

Abbreviation	Name	Comment
MDF [52]	Multi-Dictionary Formatter	More a configuration for the ShoeBox software than a real standard – too tightly coupled to one software product.
LIFT [37]	Lexicon Interchange FormaT	An XML schema for the storing of lexical information, an interchange format for MDF dictionaries – without simple tool support for reading and writing.
LMF [16]	Lexical Markup Framework	A very broad and generic ISO standard, integrating ISOcat for shareable semantics – too broad to be consistently applied to simple use cases.
LEXUS [66]	Lexus	A web-based dictionary curation tool based on LMF – now defunct.
RELISH [1]	Rendering Endangered Lexicons Interoperable through Standards Harmonization	A meta-standard, trying to bridge LIFT and LMF – an abandoned attempt to pool the resources of two marginal standards.
lemon [35]	lemon	A model for representing lexicons relative to ontologies.
ISOcat [63]	ISOcat	Discontinued because 'the original mandate to "standardize" data categories within the ISO framework was never fulfilled'.

Table 1: Attempts at standardizing dictionary data.

Data Alliance recently endorsed a *Linguistic Data Interest Group* aiming at facilitating 'the development of reproducible research in linguistics' [58]. Since the importance of cross-linguistic data is constantly increasing, it is time to re-evaluate and improve the state of standardization of linguistic data.

While we have to ask ourselves whether adding another standard might worsen the situation [67], it is also clear that the current problems of "data-FAIR-ness" in comparative and typological linguistics persist and that standardization is the only way to tackle them. What may set our attempt apart from previous trials is a focus on data re-use scenarios as motivating use cases. Previously, the focus was mostly on language documentation (i.e. oriented towards the needs of data collectors); hence, previous attempts suffered from the divide between non-comparative data collectors and comparative linguists wanting to reuse the data. Our proposal, however, builds on years of familiarity of our team with the types of data and the problems linguists face in modelling, collecting, storing, and investigating cross-linguistic data.

Results

To address the above-mentioned obstacles of sharing and re-use of cross-linguistic datasets, the *Cross-Linguistic Data Formats* initiative (CLDF) offers modular specifications for common data types in language typology and historical linguistics, which are based on a shared data model and a formal ontology.

Data model

The data model underlying the CLDF specification is simple, yet expressive enough to cover a range of data types commonly collected in language typology and historical linguistics. The core concepts of this model have been derived from the data model which was originally developed for the *Cross-Linguistic Linked Data project* (CLLD, [26]), which aimed at developing and curating interoperable

Name	URL	Ref.	Description
World Atlas of Language Structures	`wals.info`	[12]	Grammatical survey of more than 2000 languages world-wide.
Comparative Siouan Dictionary	`csd.clld.org`	[57]	Etymological dictionary of Siouan languages.
Phoible	`phoible.org`	[51]	Cross-linguistic survey of soundinventories for more than 2000 languages world-wide.
Glottolog	`glottolog.info`	[24]	Reference catalogue of language names, geographic locations, and affiliations.
Concepticon	`concepticon.clld.org`	[40]	Reference catalogue of word meanings and concepts used in cross-linguistic surveys and psycholinguistic studies.

Table 2: Examples for popular databases produced within the CLLD framework

data publication structures using linked data principles as integration mechanism for distributed resources. The CLLD project resulted in a large number of online datasets which provide linguists with a uniform "look-and-feel" despite their diverse content (see Table 2).

The main entities in this model are: (a) *Languages* – or more generally *languoids* (see [24]), which represent the objects under investigation; (b) *Parameters*, the comparative concepts (see [25]), which can be measured and compared across languages; and (c) *Values*, the "measurements" for each pair language and parameter. In addition, each triple should have at least one (d) *Source*, as cross-linguistic data is typically aggregated from primary sources which themselves are the result of language documentation based on linguistic fieldwork. This reflects the observation of Good and Cysouw [18] that cross-linguistic data deals with *doculects*, i.e. languages as they are documented in a specific primary source – rather than languages as they are spoken directly by the speakers.

In this model, each *Value* is related to one *Parameter* and one *Language* and can be based on multiple *Sources*. The many-to-many relation between *Value* and *Source* is realized via *References* which can carry an additional *Context* attribute, which is typically represented by page numbers when dealing with printed sources.

The CLDF Specification

CLDF is a package format, describing various types of cross-linguistic data. Each type is modeled via a CLDF *module*, with additional, orthogonal aspects of the data modeled as CLDF *components*. This approach mirrors the way Dublin Core metadata terms are packaged into meaningful sets using *Application Profiles* [2]: CLDF modules are profiles of cross-linguistic data types, consisting of CLDF components and terms from the CLDF ontology.

1 CLDF Ontology

The CLDF specification recognizes certain objects and properties with well-known semantics in comparative linguistics. These are listed in the CLDF Ontology, thereby making them available for reference by URIs. Wherever possible, this ontology builds on existing ontologies like the *General Ontology for Linguistic Description* (GOLD, [8]).

2 Basic Modules in CLDF

Currently, CLDF defines four basic modules which handle the most basic types of data which are frequently being used, collected, and shared in historical linguistics and typology. The *Wordlist* module handles lexical data which is usually based on a *concept list* that has been translated into a certain number of different languages. The *StructureDataset* module handles grammatical features in a very broad sense, which are usually collected to compare languages typologically. The *ParallelText* module can be used to encode texts which were translated into different languages and are split into functional units (like similar sentences or paragraphs) to render them comparable. The *Dictionary* module makes it possible to encode the lexicon of individual languages. Each of the modules defines additional components which define relations among the values across languages, inside a language, or value-internally. Some of the parameters are further standardized via reference catalogues (see below), and scholars preparing their datasets in CLDF format are encouraged to make sure that they comply to these standards to ease the reuse of their data (see Table 3).

Module	**Parameter**	**Value**	**Reference**	**Example**
Wordlist	Concept	form	Concepticon, [CLTS]	ABVD [20]
ParallelText	Functional unit	functional equivalent		BibleCorpus [49]
StructureDataset	Grammatical feature	value	[Grammaticon]	WALS [12]
Dictionary	Comparison meaning	translation equivalent	[Concepticon]	Daakaka Dictionary [56]

Table 3: Interpretation of the core data model per CLDF module. Reference catalogues in square brackets indicate that these are currently only partially implemented or in planning.

3 Package Format of CLDF

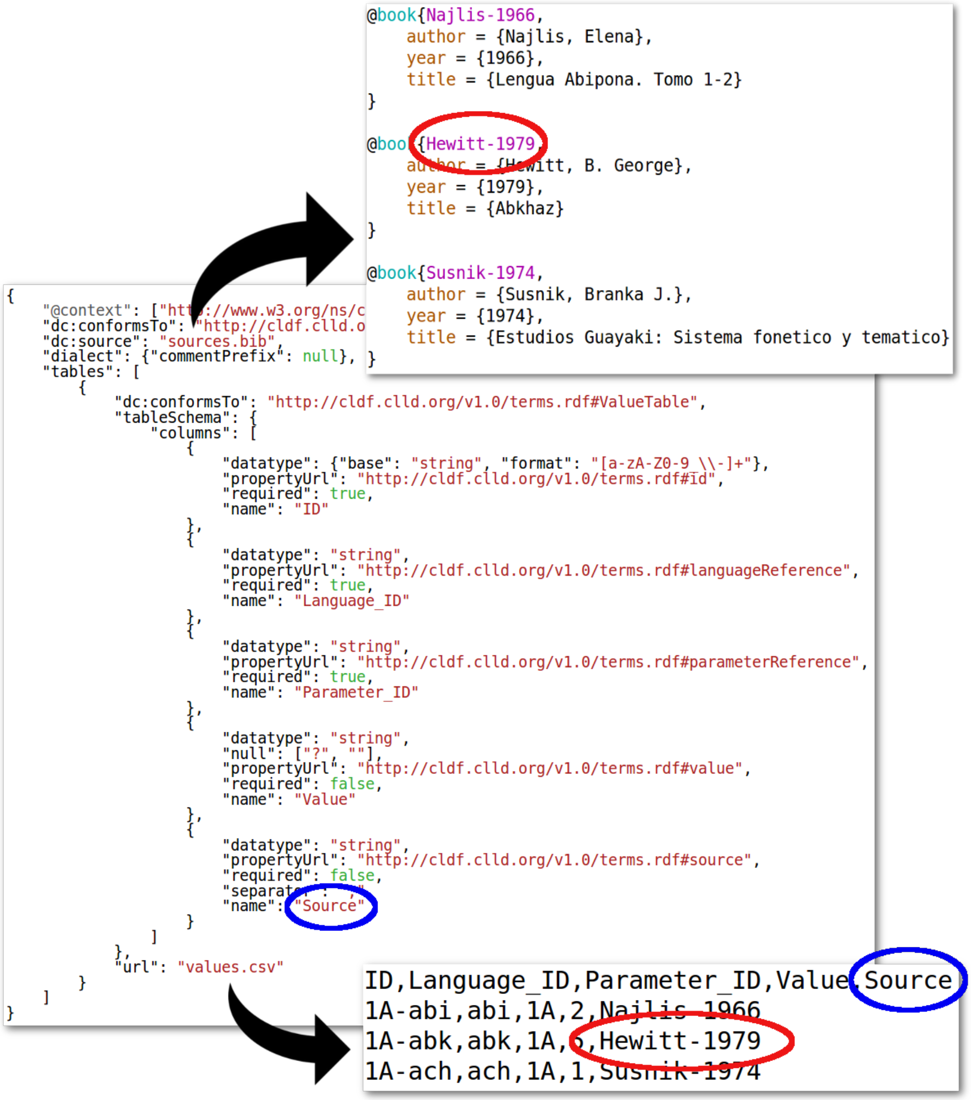
CLDF is built on the World Wide Web Consortium (W3C) recommendations *Model for Tabular Data and Metadata on the Web* [62] and *Metadata Vocabulary for Tabular Data* [55] (henceforth referred to as CSVW for "CSV on the Web"), which provide a package format allowing us to tie together multiple files containing tabular data (see Figure 1). Thus, each CLDF dataset is described by a JSON metadata file according to [55].

This means that there are standard ways of including metadata: *Common properties* on *table* or *table group* descriptions can be used to add (a) bibliographic metadata using terms from the Dublin Core namespace (`http://purl.org/dc/terms/`), (b) provenance information using terms from the PROV namespace (`https://www.w3.org/ns/prov`), (c) catalogue information using terms from Data Catalog Vocabulary (`http://www.w3.org/ns/dcat#`). Thus, by providing a way to specify such metadata in a machine-readable way, CLDF complements the efforts of the RDA Linguistics Interest Group [3].

4 Extensibility of CLDF

The CLDF specification is designed for extensibility. A CLDF dataset can comprise any number of additional tables (by simply adding corresponding table definitions in the metadata file), or by adding additional columns to specified tables Thus, we expect to see further standardization by converging usage, much like Flickr machine tags [44] evolved.

Figure 1: Using CSVW metadata to describe the files making up a CLDF dataset.



```
@book{Najlis-1966,
    author = {Najlis, Elena},
    year = {1966},
    title = {Lengua Abipona. Tomo 1-2}
}

@book{Hewitt-1979,
    author = {Hewitt, B. George},
    year = {1979},
    title = {Abkhaz}
}

@book{Susnik-1974,
    author = {Susnik, Branka J.},
    year = {1974},
    title = {Estudios Guayaki: Sistema fonetico y tematico}
}
```

```
{
    "@context": ["http://www.w3.org/ns/c
    "dc:conformsTo": "http://cldf.clld.o
    "dc:source": "sources.bib",
    "dialect": {"commentPrefix": null},
    "tables": [
        {
            "dc:conformsTo": "http://cldf.clld.org/v1.0/terms.rdf#ValueTable",
            "tableSchema": {
                "columns": [
                    {
                        "datatype": {"base": "string", "format": "[a-zA-Z0-9_\\-]+"},
                        "propertyUrl": "http://cldf.clld.org/v1.0/terms.rdf#id",
                        "required": true,
                        "name": "ID"
                    },
                    {
                        "datatype": "string",
                        "propertyUrl": "http://cldf.clld.org/v1.0/terms.rdf#languageReference",
                        "required": true,
                        "name": "Language_ID"
                    },
                    {
                        "datatype": "string",
                        "propertyUrl": "http://cldf.clld.org/v1.0/terms.rdf#parameterReference",
                        "required": true,
                        "name": "Parameter_ID"
                    },
                    {
                        "datatype": "string",
                        "null": ["?", ""],
                        "propertyUrl": "http://cldf.clld.org/v1.0/terms.rdf#value",
                        "required": false,
                        "name": "Value"
                    },
                    {
                        "datatype": "string",
                        "propertyUrl": "http://cldf.clld.org/v1.0/terms.rdf#source",
                        "required": false,
                        "separator": ";",
                        "name": "Source"
                    }
                ]
            },
            "url": "values.csv"
        }
    ]
}
```

```
ID,Language_ID,Parameter_ID,Value,Source
1A-abi,abi,1A,2,Najlis-1966
1A-abk,abk,1A,5,Hewitt-1979
1A-ach,ach,1A,1,Susnik-1974
```

This extension mechanism (and backwards compatible, frequent releases) allows us to start out small and focused on a handful of use cases and data types for which there is already tool support.

Reference Catalogues

Creating a rather lean standard format like CLDF has become far easier with the existence of reference catalogues, and the linking mechanism built into the W3C model by piggy-backing on JSON-LD – a JSON serialization of the RDF model underlying the Linked Data principles. Corresponding properties in the CLDF Ontology allow unambiguous references to standard catalogues like Glottolog [24] and ISO 639-3 – for languoids, and Concepticon [40] – for lexical concepts.

We are currently working on additional reference catalogues for phonetic transcriptions (working title *Cross-Linguistic Transcription Systems*, [41]) and grammatical features (working title *Grammaticon*, [27]) and hope to add them in future versions of the CLDF specification.

A Python API to interact with CLDF datasets: `pycldf`

In many research disciplines the Python programming language has become the de-facto standard for data manipulation (often including analyses, [50]) Thus, providing tools for programmatic access to CLDF data from Python programs increases the usefulness of a format specification like CLDF. We implemented a Python package `pycldf`[14], serving as reference implementation of the CLDF standard, and in particular supporting reading, writing and validating CLDF datasets (see `https://github.com/cldf/pycldf/tree/master/examples`).

As a matter of fact, by making use of the table descriptions in a CLDF metadata file, pycldf can do a lot more. For example, based on the datatype descriptors and foreign key relations specified in table schemas, pycldf can provide a generic conversion of a CLDF dataset into an SQLite database; thereby allowing analysis of CLDF datasets using SQL – one of the work horses of data science.

Discussion

At the beginning of the CLDF initiative we developed a list of practitioner requirements for cross-linguistic data, based on the experiences of linguists who have worked and are regularly working with cross-linguistic datasets. These practical principles are summarized in table 4 [23], and when comparing them with our first version of CLDF, it can be seen that CLDF still conforms to all of them. Furthermore, when comparing our initial requirements with the criteria for file formats and standards put forward in guidelines for research data management such as the ones proposed by the WissGrid project [43], one can also see that the perspectives are largely compatible, thus corroborating our hope that while being sufficiently specific to be of use for linguists, CLDF will also be generic enough to blend in with current best practices for research data management across disciplines.

Following a similar line of reasoning as Gorgolewski et al. [19] lay out in their proposal of a unified data structure for brain imaging data, and building on recommendations from the "Good Practices of Scientific Computing" by Wilson et al. [65], we decided to base CLDF on well-known and well-supported serialization formats, namely CSV and JSON, with their specific shortcomings being outbalanced by building on CSVW, including its concept of CSV dialects, which allows us to support more variation in tabular data files and help with adaptation of the format. CSVW and its support for foreign keys between tables also allows us to seamlessly implement the recommendation to "anticipate the need to use multiple tables, and use a unique identifier for every record" [19].

Since CSVW is specified as a JSON-LD dialect (i.e. grounded in the Resource Description Framework RDF [47]), it can be combined with an RDF *Vocabulary* or *Ontology* to provide (a) the syntax of a relational serialization format via [55], as well as (b) the semantics of the entities in the data model via the ontology. Thus, the CLDF Ontology provides answers to the two questions of "Which things do exist?" and "Which things are based on others?", which are considered crucial to assess the identification needs for data collections [43].

Abbr.	Requirement	Note
P	PEP 20	"Simple things should be simple, complex things should be possible" [54]: Datasets can be one simple CSV file, encoding language-parameter-value-triples.
R	Referencing	If entities and parameters can be linked to reference catalogues such as Glottolog or Concepticon, this should be preferred to duplicating information.
A	Aggregability	Data should be simple to concatenate, merge, and aggregate in order to guarantee their reusability.
H	Human- and machine-readability	Data should be both editable *by hand* and amenable to reading and writing by software (preferable software which typical linguists can be expected to use).
T	Text	Data should be encoded as UTF-8 text files or in formats that provide full support for UTF-8.
I	Identifiers	Identifiers should be resolvable HTTP-URLs, where possible, if not, this should be documented in the metadata.
C	Compatibility	Compatibility with existing tools, standards, and practices should always be kept in mind and never easily given up.
E	Explicitness	One row should only store one data point, and each cell should only have one type of data, unless specified in the metadata.

Table 4: Practical demands regarding cross-linguistic data formats.

When adopting CSVW as the basis of the specification, it may seem counterintuitive to model source information via BibTeX – rather than as just another CSV table, linked to with foreign keys. But given that (a) most potential users are acquainted with BibTeX, (b) Glottolog – the most extensive bibliography of language descriptions – disseminates BibTeX and (c) the many-to-many relation between values and sources would have required an additional association table, BibTeX seemed to be the right choice when maximizing maintainability of datasets.

Another design decision taken with CLDF was to not specify a single-file format. Instead of forcing users to provide their data in database formats, like SQLite [60], or in pure text formats with extensible markup, like the NEXUS format in biology [46], we opted for specifying a multi-file format – and deliberately chose to not define any packaging. Instead, we regard packaging of usually rather small sets of small text files as a problem for which multiple solutions with particular use cases have already been proposed (e.g. *zip* for compression, *bagit* [34] for archiving, etc.). We do not even have to specify a particular directory layout for the multiple files forming a CLDF dataset, because the description file references data files using URIs, thereby turning CLDF into a multi-file format almost as flexible as HTML.[2]

Since CLDF has been developed in close collaboration with researchers working on different ends of data-driven research in historical linguistics and language typology, CLDF is already being used by large linguistic projects [31] and as the data format for publishing supporting information [29]. CLDF is the native format for the forthcoming global language databases *Grambank*, *Lexibank* and *Parabank*(`http://glottobank.org/`) being developed by a consortium of research centers and universities. Further, CLDF is by now already supported by a larger number of software packages and applications, ranging from libraries for automatic sequence comparison in historical linguistics (LingPy [42]), via packages for phylogenetic analyses (BEASTLing [48]), up to interfaces for data inspection and curation (EDICTOR [39]). Since the CLDF initiative was

[2] The advantages of multi-file formats are that they describe a curation paradigm that we extracted from more than ten years of experience with databases in the CLLD project along with inobtrusive ways to enhance existing datasets.

born out of the CLLD project, it is readily integrated into the CLLD framework and will allow users to publish their data without efforts on the web, making their data *findable* by exposing data and metadata to the major search engines, and increasing thus their interoperability. It is important to note that CLDF is not limited to linguistic data alone. By embracing reference catalogues like Glottolog which provide geographical coordinates and are themselves referenced in large-scale surveys of cultural data, such as D-PLACE [33], CLDF may drastically facilitate the testing of questions regarding the interaction between linguistic, cultural, and environmental factors in linguistic and cultural evolution.

Methods

Efforts to standardize cross-linguistic data, in particular typological datasets and with the aim of comparability across datasets, have been undertaken since at least 2001, when Dimitriadis presented his *Typological Database System* [9] (see [11]). One initial step was to introduce general database principles to database design in linguistic typology [10].

Rather than standardizing data formats, the CLLD project largely tried to standardize the software stack for cross-linguistic databases. Still, the core data model which could be extracted from these database software implementations served as one of the inspirations when standard data formats were discussed at the workshop *Language Comparison with Linguistic Databases*, held 2014 at the Max Planck Institute for Psycholinguistics in Nijmegen.

The followup workshop *Language Comparison with Linguistic Databases 2* – held 2015 at the Max Planck Institute for Evolutionary Anthropology in Leipzig – saw already concrete proposals towards what now is CLDF [23]; and later this year, the workshop *Capturing Phylogenetic Algorithms for Linguistics* – held at the Lorentz Center in Leiden – brought together people interested in analysis of cross-linguistic data, thus providing a test bed for the proposals.

Apart from these bigger meetings where scholars discussed ideas of standardization, the CLDF-initiative profited a lot from the numerous Glottobank meetings (`http://glottobank.org`) co-organized by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for the Science of Human History (Jena), in which the big-picture ideas of standards for linguistic data were discussed in more concrete by smaller teams which came forward to work on specific aspects of the specification, including reference catalogues like Concepticon, the handling of etymological data, and the linking with external projects like D-PLACE.

These events formed a group representing the main institutions in the small field of large-scale comparison of cross-linguistic data, which contributed to the CLDF specification.

When a Linguistics Data Interest Group was endorsed by RDA in 2017, echoing RDA's call to 'develop and apply common standards across institutions and domains to ensure greater interoperability across datasets' in Linguistics, this matched up nicely with our progress with CLDF 1.0.

Code Availability

The CLDF specification is curated using a GitHub repository (`https://github.com/cldf/cldf`). Released versions are published and archived via ZENODO. The current version of the specification is CLDF 1.0 [15].

The `pycldf` package is maintained in a GitHub repository (`https://github.com/cldf/cldf`). Released versions are available from the Python Package Index (`https://pypi.python.org/pypi/pycldf`) and archived with ZENODO.

Author Contributions

RDG, RF, MC, HH, MH, and JML initiated the CLDF initiative. By making CLDF a key initiative for data handling at the Department of Linguistic and Cultural Evolution (MPI-SHH, Jena), RDG provided financial, administrative, and conceptual support for CLDF. RF, SJG, and JML conceptualized the specification. RF conceptualized and designed the implementation. RF, MC, SJG, HH, MH, and JML contributed to the specification. RF and SB wrote the code for the `pycldf` package. RF, JML, and CR wrote the first draft. All authors revised the first draft and agree with the final version.

Competing Interests

The authors declare no competing interests.

Acknowledgements

This research would not have been possible without the generous support by many institutes and funding agencies. As part of the CLLD project (`http://clld.org`) and the GlottoBank project (`http://glottobank.org`), the work was supported by the Max Planck Society, the Max Planck Institute for the Science of Human History, and the Royal Society of New Zealand (Marsden Fund grant 13-UOA-121, RF). JML was funded by the DFG research fellowship grant 261553824 *Vertical and lateral aspects of Chinese dialect history* (2015-2016) and the ERC Starting Grant 715618 *Computer-Assisted Language Comparison* (`http://calc.digling.org`). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041).

References

[1] Helen Aristar-Dry et al. “”Rendering Endangered Lexicons Interoperable through Standards Harmonization”: the RELISH project”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. ISBN: 978-2-9517408-7-7.

[2] Thomas Baker. *Guidelines for Dublin Core Application Profiles*. Tech. rep. 2009. URL: http://dublincore.org/documents/2009/05/18/profile-guidelines/.

[3] A. L. Berez-Kroeker et al. *The Austin Principles of Data Citation in Linguistics*. 2017. URL: http://site.uit.no/linguisticsdatacitation/austinprinciples/.

[4] Damián E. Blasi, Susanne Maria Michaelis, and Martin Haspelmath. "Grammars are robustly transmitted even during the emergence of creole languages". In: *Nature Human Behaviour* (2017), pp. 1–5. DOI: 10.1038/s41562-017-0192-4.

[5] Robert Blust and Stephen Trussell. *The Austronesian Comparative Dictionary*. 2010. URL: http://www.trussel2.com/acd/ (visited on 01/06/2018).

[6] Remco Bouckaert et al. "Mapping the origins and expansion of the Indo-European language family". In: *Science* 337.6097 (2012), pp. 957–960.

[7] Will Chang et al. "Ancestry-constrained phylogenetic analysis ssupport the Indo-European steppe hypothesis". In: *Language* 91.1 (2015), pp. 194–244.

[8] GOLD Community. *General Ontology for Linguistic Description (GOLD)*. Ontology. Department of Linguistics (The LINGUIST List), Indiana University, 2010. URL: http://linguistics-ontology.org/.

[9] Alexis Dimitriadis. *The Typological Database System*. URL: http://languagelink.let.uu.nl/tds/index.html.

[10] Alexis Dimitriadis and Simon Musgrave. "Designing linguistic databases: A primer for linguists". In: *The Use of Databases in Cross-Linguistic Studies*. Ed. by Martin Everaert, Simon Musgrave, and Alexis Dimitriadis. Vol. 41. Empirical Approaches to Language Typology [EALT]. Berlin: De Gruyter Mouton, 2009, pp. 13–75.

[11] Alexis Dimitriadis et al. "How to integrate databases without starting a typology war: The Typological Database System". In: *The Use of Databases in Cross-Linguistic Studies*. Ed. by Martin Everaert, Simon Musgrave, and Alexis Dimitriadis. Vol. 41. Empirical Approaches to Language Typology [EALT]. Berlin: De Gruyter Mouton, 2009, pp. 155–208.

[12] Matthew S. Dryer and Martin Haspelmath, eds. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: http://wals.info/.

[13] Michael Dunn, ed. *Indo-European lexical cognacy database (IELex)*. 2012. URL: http://ielex.mpi.nl/.

[14] Robert Forkel and Sebastian Bank. *pycldf 1.0.9*. Jena, 2017. DOI: 10.5281/zenodo.1119287.

[15] Robert Forkel et al. *CLDF 1.0*. Tech. rep. Jena, 2017. DOI: 10.5281/zenodo.1117644.

[16] Gil Francopoulo. *LMF Lexical Markup Framework*. Wiley, 2013. URL: http://www.lexicalmarkupframework.org/.

[17] E. Gibson et al. "Color naming across languages reflects color use". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.40 (2017), pp. 10785–10790.

[18] Jeff Good and Michael Cysouw. “Languoid, doculect, glossonym: Formalizing the notion of ‘language’”. In: *Journal of Language Documentation and Conservation* 7 (2013), pp. 331–359.

[19] Krzysztof J. Gorgolewski et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments”. In: *Nature Scientific Data* 3.160044 (2016). DOI: `10.1038/sdata.2016.44`. URL: `http://dx.doi.org/10.1038/sdata.2016.44`.

[20] Simon J. Greenhill, Robert Blust, and Russell D. Gray. “The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics”. In: *Evolutionary Bioinformatics* 4 (2008), pp. 271–283.

[21] *Guidelines on the Handling of Research Data in Biodiversity Research*. Deutsche Forschungsgemeinschaft. 2015.

[22] *H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. European Commission: Directorate-General for Research & Innovation. 2017.

[23] Harald Hammarström. *A Proposal for Data Interface Formats for Cross-Linguistic Data*. Language Comparison with Linguistic Databases 2. 2015. URL: `https://github.com/clld/lanclid2/raw/master/presentations/hammarstrom.pdf?raw=true`.

[24] Harald Hammarström, Robert Forkel, and Martin Haspelmath. *Glottolog*. Version 3.0. 2017. URL: `http://glottolog.org`.

[25] Martin Haspelmath. “Comparative concepts and descriptive categories”. In: *Language* 86.3 (2010), pp. 663–687.

[26] Martin Haspelmath and Robert Forkel. *CLLD – Cross-Linguistic Linked Data*. 2015. URL: `http://clld.org`.

[27] Martin Haspelmath and Robert Forkel. *Toward a standard list of grammatical comparative concepts: The Grammaticon*. Talk held at the database workshop of the ALT Meeting 2017. 2017. URL: `http://dynamicsoflanguage.edu.au/storage/alt-2017-database-workshop-book-of-abstracts-forkel-haspelmath-haynie-skirgard.pdf`.

[28] H. J. Haynie and C. Bowern. “Phylogenetic approach to the evolution of color term systems”. In: *Proc. Natl. Acad. Sci. U.S.A.* 113.48 (2016), pp. 13666–13671.

[29] Nathan Hill and Johann-Mattis List. “Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages”. In: *Yearbook of the Poznań Linguistic Meeting* 3.1 (2017), pp. 47–76.

[30] *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge: Cambridge University Press, 1999.

[31] Gereon Kaiping and Marian Klamer, eds. *LexiRumah*. 2017. URL: `http://www.model-ling.eu/lexirumah/`.

[32] Werner Kalusky. *Die Transkription der Sprachlaute des Internationalen Phonetischen Alphabets: Vorschläge zu einer Revision der systematischen Darstellung der IPA-Tabelle*. München: LINCOM Europa, 2017.

[33] Kathryn R. Kirby et al. “D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity”. In: *PLOS ONE* 11.7 (July 2016), pp. 1–14. DOI: 10.1371/journal.pone.0158391. URL: https://doi.org/10.1371/journal.pone.0158391.

[34] J. Kunze et al. *The BagIt File Packaging Format (V0.97)*. Tech. rep. 2016. URL: https://tools.ietf.org/html/draft-kunze-bagit-14.

[35] *lemon - The Lexicon Model for Ontologies*. URL: http://lemon-model.net/.

[36] M. Paul Lewis and Charles D. Fennig, eds. *Ethnologue. Languages of the world*. 2013. URL: http://www.ethnologue.com.

[37] *Lexical Interchange Format Standard*. XML schema in a now defunct code repository. URL: https://github.com/sillsdev/lift-standard.

[38] J.-M. List et al., eds. *CLICS: Database of Cross-Linguistic Colexifications*. Version 1.0. 2014. Archived at: http://www.webcitation.org/6ccEMrZYM. URL: http://clics.lingpy.org.

[39] Johann-Mattis List. “A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 2017, pp. 9–12.

[40] Johann-Mattis List, Michael Cysouw, and Robert Forkel. “Concepticon. A resource for the linking of concept lists”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. LREC 2016. (Portorož, May 23–28, 2016). Ed. by Nicoletta Calzolari (Conference Chair) et al. European Language Resources Association (ELRA), 2016, pp. 2393–2400.

[41] Johann-Mattis List and Robert Forkel. *Cross-Linguistic Transcription Systems*. Python library. 2017. URL: https://github.com/cldf/clts.

[42] Johann-Mattis List and Robert Forkel. *LingPy. A Python library for historical linguistics*. Version 2.6. Forthcoming. URL: http://lingpy.org.

[43] Jens Ludwig and Harry Enke. “Leitfaden zum Forschungsdatenmanagement”. Version 0.6. In: *Ergebnisse aus dem WissGrid-Projekt* 15 (2013), p. 120. URL: www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-oeffentlicher-Entwurf-Checkliste-Forschungsdaten-Management.pdf.

[44] *Machine tags*. 2007. URL: https://www.flickr.com/groups/api/discuss/72157594497877875.

[45] Ian Maddieson and Christophe Coupé. “Human spoken language diversity and the acoustic adaptation hypothesis”. In: *The Journal of the Acoustical Society of America* 138.3 (2015), 1838–1838.

[46] D. R. Maddison, D. L. Swofford, and W. P. Maddison. “NEXUS: an extensible file format for systematic information”. In: *Syst. Biol.* 46.4 (1997), pp. 590–621.

[47] *RDF Primer*. Tech. rep. 2004. URL: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[48] Luke Maurits et al. “BEASTling: A software tool for linguistic phylogenetics using BEAST 2”. In: *PLOS ONE* 12.8 (Aug. 2017), pp. 1–17. DOI: 10.1371/journal.pone.0180908. URL: https://doi.org/10.1371/journal.pone.0180908.

[49] Thomas Mayer and Michael Cysouw. “Creating a massively parallel bible corpus”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. LREC. (Reykjavik, May 26–31, 2014). Ed. by Nicoletta Calzolari (Conference Chair) et al. European Language Resources Association (ELRA), 2014, pp. 3158–3162. ISBN: 978-2-9517408-8-4.

[50] K. J. Millman and M. Aivazis. “Python for Scientists and Engineers”. In: *Computing in Science Engineering* 13.2 (2011), pp. 9–12. ISSN: 1521-9615. DOI: 10.1109/MCSE.2011.36.

[51] Steven Moran, Daniel McCloy, and Richard Wright. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://phoible.org, Accessed on 2015-10-01.) 2015.

[52] *Multi-Dictionary Formatter (MDF)*. URL: https://software.sil.org/shoebox/mdf/.

[53] Mitchell G. Newberry et al. “Detecting evolutionary forces in language change”. In: *Nature* (2017), pp. 1–4. DOI: 10.1038/nature24455.

[54] Tim Peters. *PEP 20 – The Zen of Python*. 2004. URL: https://www.python.org/dev/peps/pep-0020/.

[55] Rufus Pollock et al. *Metadata Vocabulary for Tabular Data*. Tech. rep. World Wide Web Consortium (W3C), 2015. URL: https://www.w3.org/TR/tabular-metadata/.

[56] Kilu von Prince. “Daakaka dictionary”. In: *Dictionaria* 1 (2017), pp. 1–2171. URL: http://dictionaria.clld.org/contributions/daakaka.

[57] Robert L. Rankin et al., eds. *Comparative Siouan Dictionary*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2015. URL: http://csd.clld.org/.

[58] *RDA and Linguistics*. URL: https://www.rd-alliance.org/rda-disciplines/rda-and-linguistics.

[59] Anju Saxena and Lars Borin. “Carving Tibeto-Kanauri by its joints: Using basic vocabulary lists for genetic grouping of languages”. In: *Approaches to Measuring Linguistic Differences*. Berlin: De Gruyter Mouton, 2013, pp. 175–198. ISBN: 978-3-11-030525-8.

[60] *SQLite As An Application File Format*. URL: https://sqlite.org/appfileformat.html.

[61] Marco Tamburelli and Lissander Brasca. “Revisiting the classification of Gallo-Italic: a dialectometric approach”. In: *Digital Scholarship in the Humanities* fqx41 (2017).

[62] Jeni Tennison, Gregg Kellogg, and Ivan Herman. *Model for Tabular Data and Metadata on the Web*. Tech. rep. World Wide Web Consortium (W3C), 2015. URL: `https://www.w3.org/TR/tabular-data-model/`.

[63] *The ISOcat Data Category Registry*. no longer maintained. URL: `https://tla.mpi.nl/tools/tla-tools/older-tools/isocat/`.

[64] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.160018 (2016), pp. 1–9.

[65] Greg Wilson et al. "Good enough practices in scientific computing". In: *PLOS Computational Biology* 13.6 (June 2017), pp. 1–20. DOI: `10.1371/journal.pcbi.1005510`. URL: `https://doi.org/10.1371/journal.pcbi.1005510`.

[66] Katarzyna Wojtylak. *LEXUS for creating lexica*. 2012. URL: `https://tla.mpi.nl/tools/tla-tools/older-tools/lexus/`.

[67] *xkcd - A Webcomic - Standards*. URL: `http://xkcd.com/927/`.

[68] Hyejin Youn et al. "On the universal structure of human lexical semantics". In: *Proceedings of the National Academy of Sciences of the United States of America* (2016). DOI: `10.1073/pnas.1520752113`.