Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6995–7002
Marseille, 11–16 May 2020

CLDFBench: Give Your Cross-Linguistic Data a Lift

Robert Forkel[1] and Johann-Mattis List[1]
[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena
{forkel, list}@shh.mpg.de

Abstract

While the amount of cross-linguistic data is constantly increasing, most datasets produced today and in the past cannot be considered FAIR (findable, accessible, interoperable, and reproducible). To remedy this and to increase the comparability of cross-linguistic resources, it is not enough to set up standards and best practices for data to be collected in the future. We also need consistent workflows for the "retro-standardization" of data that has been published during the past decades and centuries. With the Cross-Linguistic Data Formats initiative, first standards for cross-linguistic data have been presented and successfully tested. So far, however, CLDF creation was hampered by the fact that it required a considerable degree of computational proficiency. With `cldfbench`, we introduce a framework for the retro-standardization of legacy data and the curation of new datasets that drastically simplifies the creation of CLDF by providing a consistent, reproducible workflow that rigorously supports version control and long term archiving of research data and code. The framework is distributed in form of a Python package along with usage information and examples for best practice. This study introduces the new framework and illustrates how it can be applied by showing how a resource containing structural and lexical data for Sinitic languages can be efficiently retro-standardized and analyzed.

Keywords: cross-linguistic data, retro-standardization, data curation

1. Introduction

While the amount of cross-linguistic data is constantly increasing, this "data deluge" has the potentially detrimental effect of making it easier to neglect "old" data – since similar enough "new" data may be available. For many low-resource languages – in particular those ones which are on the brink of extinction – this means they will effectively disappear from research when existing data is of low quality and new data can no longer be obtained. Thus, making existing data ready for re-use – "lifting" it into current research environments – is vital for any cross-linguistic research. At the same time, it is also vital to make sure that new data that is being constantly produced by the linguistic community is provided in such a form that it can easily be lifted to overarching standards.

While initial efforts to increase the reproducibility of research in linguistics have been undertaken (Berez-Kroeker et al., 2018) and large datasets which aggregate linguistic resources in order to make them apt for cross-linguistic investigations have been increasingly prepared during the past two decades (Dryer and Haspelmath, 2011; Haspelmath and Tadmor, 2009; Key and Comrie, 2016), we still face a situation in which the majority of linguistic data does not conform to the principles of FAIR data in the sense of Wilkinson et al. (2016), as the data are often not *findable*, not *accessible*, not *interoperable*, and also not *reproducible*.

2. Cross-Linguistic Data Formats

As one step towards increasing the interoperability of cross-linguistic data, the Cross-Linguistic Data Formats (CLDF, `https://cldf.clld.org`) specification was published in 2018 (Forkel et al., 2018). Having started with a series of workshops in 2014, during which linguists who make active use of cross-linguistic data in their research discussed the challenges of data standardization, the first version of the CLDF specification proposed standard formats along with evaluation tools for the most basic types of data encountered in cross-linguistics research, namely *word lists*, *structural datasets*, and *dictionaries*.

One of the major design ideas of CLDF was to make active use of existing standards widely used on the web, such as, notably, CSVW (Tennison et al., 2015; Pollock et al., 2015), which is itself building on JSON-LD (W3C, 2019), and thus directly compatible with the Resource Description Framework RDF (`https://www.w3.org/RDF/`). Thus, linking the entities of a given dataset to those of another is built-in. Standards for each of the three modules of CLDF were created in such a way that common entities (such as, for example, metadata about languages) could be shared. Extensibility of the standard was built into the system, to allow for the successive inclusion of additional datatypes, such as, for example, parallel texts (Östling, 2014), inter-linear-glossed texts (Lewis and Xia, 2010), or even annotated rhyme data (List et al., 2019c).

With CLDF we have laid the foundations for interoperable, cross-linguistic data. We furthermore established the appropriateness of such data formats, by converting existing databases (see `https://clld.org/datasets.html`) to CLDF.

3. Lifting and Retro-Standardizing Linguistic Data

Converting existing databases to a standard format like CLDF adds value not only on the side of the data consumer, but also for the data creator, e.g. through standardized quality control, or available tooling for publication, such as the `clld` toolkit (Forkel et al., 2019), which serves cross-linguistic data on the web and has been used for the publication of many well-known cross-linguistic datasets, such as the *World Atlas of Language Structures Online* by Dryer and Haspelmath (2011), or the *World Loanword Database* by Haspelmath and Tadmor (2009).

Making data readable for tools could be called "syntactic interoperability". But with retro-standardization we also

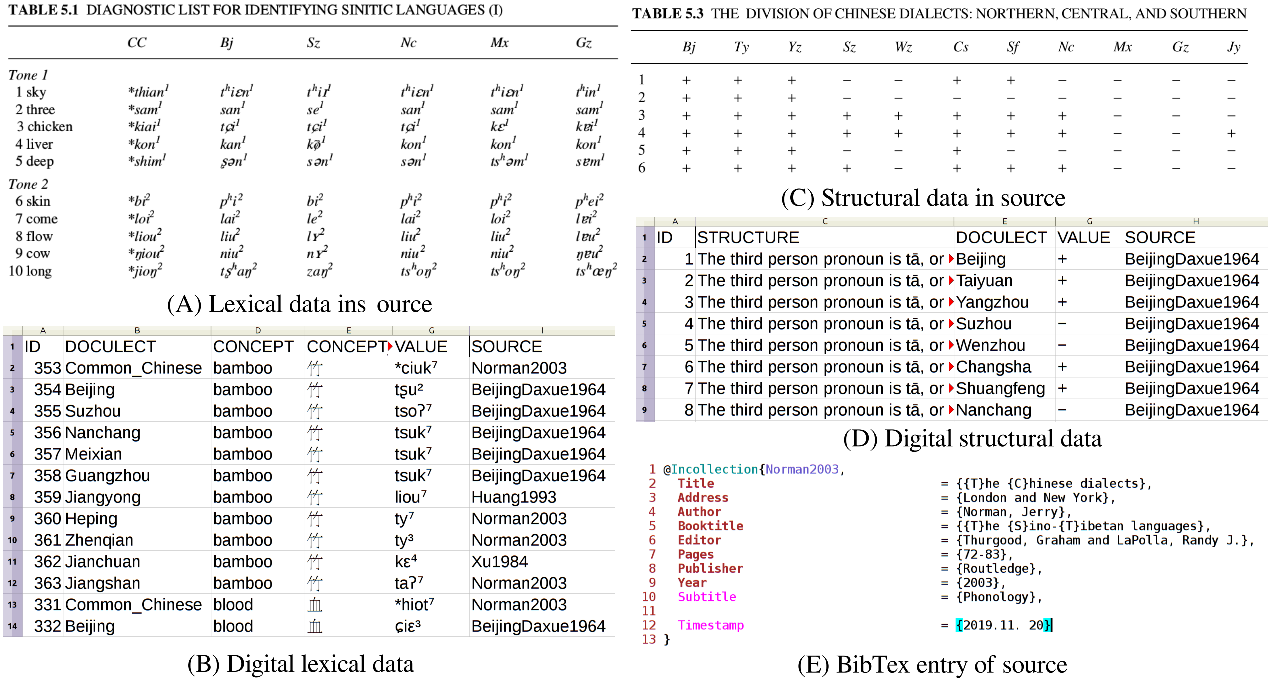


TABLE 5.1 DIAGNOSTIC LIST FOR IDENTIFYING SINITIC LANGUAGES (I)

	CC	*Bj*	*Sz*	*Nc*	*Mx*	*Gz*
Tone 1						
1 sky	**thian^1*	*t^{h}iɛn^1*	*t^{h}iɪ^1*	*t^{h}iɛn^1*	*t^{h}iɛn^1*	*t^{h}in^1*
2 three	**sam^1*	*san^1*	*se^1*	*san^1*	*sam^1*	*sam^1*
3 chicken	**kiai^1*	*tɕi^1*	*tɕi^1*	*tɕi^1*	*kɛ^1*	*kɐi^1*
4 liver	**kon^1*	*kan^1*	*kø̃^{1}*	*kon^1*	*kon^1*	*kon^1*
5 deep	**shim^1*	*ʂən^1*	*sən^1*	*sən^1*	*ts^{h}əm^1*	*sɐm^1*
Tone 2						
6 skin	**bi^2*	*p^{h}i^2*	*bi^2*	*p^{h}i^2*	*p^{h}i^2*	*p^{h}ei^2*
7 come	**loi^2*	*lai^2*	*le^2*	*lai^2*	*loi^2*	*lɐi^2*
8 flow	**liou^2*	*liu^2*	*lʏ^2*	*liu^2*	*liu^2*	*lɐu^2*
9 cow	**ŋiou^2*	*niu^2*	*nʏ^2*	*niu^2*	*niu^2*	*ŋɐu^2*
10 long	**jioŋ^2*	*tʂ^{h}aŋ^2*	*zaŋ^2*	*ts^{h}oŋ^2*	*ts^{h}oŋ^2*	*ts^{h}œŋ^2*

(A) Lexical data ins ource

ID	DOCULECT	CONCEPT	CONCEPT	VALUE	SOURCE
353	Common_Chinese	bamboo	竹	*ciuk^7	Norman2003
354	Beijing	bamboo	竹	tʂu^2	BeijingDaxue1964
355	Suzhou	bamboo	竹	tsoʔ^7	BeijingDaxue1964
356	Nanchang	bamboo	竹	tsuk^7	BeijingDaxue1964
357	Meixian	bamboo	竹	tsuk^7	BeijingDaxue1964
358	Guangzhou	bamboo	竹	tsuk^7	BeijingDaxue1964
359	Jiangyong	bamboo	竹	liou^7	Huang1993
360	Heping	bamboo	竹	ty^7	Norman2003
361	Zhenqian	bamboo	竹	ty^3	Norman2003
362	Jianchuan	bamboo	竹	kɛ^4	Xu1984
363	Jiangshan	bamboo	竹	taʔ^7	Norman2003
331	Common_Chinese	blood	血	*hiot^7	Norman2003
332	Beijing	blood	血	ɕiɛ^3	BeijingDaxue1964

(B) Digital lexical data

TABLE 5.3 THE DIVISION OF CHINESE DIALECTS: NORTHERN, CENTRAL, AND SOUTHERN

	Bj	*Ty*	*Yz*	*Sz*	*Wz*	*Cs*	*Sf*	*Nc*	*Mx*	*Gz*	*Jy*
1	+	+	+	−	−	+	+	−	−	−	−
2	+	+	+	−	−	−	−	−	−	−	−
3	+	+	+	+	+	+	+	+	−	−	−
4	+	+	+	+	+	+	+	+	−	−	+
5	+	+	+	−	−	+	−	−	−	−	−
6	+	+	+	+	−	+	+	+	−	−	−

(C) Structural data in source

ID	STRUCTURE	DOCULECT	VALUE	SOURCE
1	The third person pronoun is tā, or	Beijing	+	BeijingDaxue1964
2	The third person pronoun is tā, or	Taiyuan	+	BeijingDaxue1964
3	The third person pronoun is tā, or	Yangzhou	+	BeijingDaxue1964
4	The third person pronoun is tā, or	Suzhou	−	BeijingDaxue1964
5	The third person pronoun is tā, or	Wenzhou	−	BeijingDaxue1964
6	The third person pronoun is tā, or	Changsha	+	BeijingDaxue1964
7	The third person pronoun is tā, or	Shuangfeng	+	BeijingDaxue1964
8	The third person pronoun is tā, or	Nanchang	−	BeijingDaxue1964

(D) Digital structural data

```
@Incollection{Norman2003,
  Title                                   = {{T}he {C}hinese dialects},
  Address                                 = {London and New York},
  Author                                  = {Norman, Jerry},
  Booktitle                               = {{T}he {S}ino-{T}ibetan languages},
  Editor                                  = {Thurgood, Graham and LaPolla, Randy J.},
  Pages                                   = {72-83},
  Publisher                               = {Routledge},
  Year                                    = {2003},
  Subtitle                                = {Phonology},

  Timestamp                               = {2019.11. 20}
}
```

(E) BibTex entry of source

Figure 1: Digitization of the exemplary dataset by Norman (2003).

aim at "semantic interoperability" – addressing the scientific value and quality of the data more directly. This appears in line with the findings of the German *Rat für Informationsinfrastrukturen*, namely that data quality was neglected in the age of digitization, and consequently that maintaining data quality is not enough, but it needs to be *increased* in order to facilitate long-term re-use of data (RfII, 2019).

As an example for an increase in quality of an existing resource, consider the typical field work notes that provide word lists for one and more languages based on a questionnaire consisting of *elicitation glosses*. While such a dataset itself can be very valuable when provided as a digital resource (even if it is only distributed in form of a PDF document), there are many aspects that make it difficult to compare one resource with other resources of a similar type. First, the questionnaires used by linguists vary widely, and there is no straightforward way to compare them automatically (List et al., 2016; List, 2018). Second, information on the language varieties documented by field work resources are often hidden in the main text, and at times even only provided in form of a custom geographic map from which it is very difficult to extract the relevant information (geographic location, name of township, etc.). Third, phonetic transcriptions vary widely, and scholars use various short cuts to make it easier to document their languages, or for aesthetic reasons, given phonological considerations, and specifically older sources often do not use the International Phonetic Alphabet (IPA, 1999) as their primary transcription system (Anderson et al., 2018).

When standardizing a given dataset along these three dimensions (elicitation glosses, language varieties, transcription system), the data can be automatically enriched with all the information that has been acquired independently from other linguistic resources for the standardized items. Elicitation glosses linked to global standards can be compared to other questionnaires, in order to identify a potential overlap in data points. Along with a close identification of the language varieties that are presented by a given dataset, one could compare different documentations of identical varieties across different times, or as documented by different fieldworkers. Last but not least, the use of standardized transcription systems could offer detailed information on many aspects related to the sound inventories of the documented languages, including the cross-linguistic frequency of sounds, their distinctive features, or the spelling traditions in different transcription systems.

In turn, a retro-standardized dataset adds to the growing pool of datasets that have already been successfully lifted, and thus increases the knowledge on cross-linguistic language resources even more. The best practice recommendations and tools we present in this study can be seen as an attempt to facilitate this work drastically. By introducing *reference catalogs* as a way to store meta-linguistic data of different types, including consistent identifiers along with additional information on language varieties (Glottolog, `https://glottolog.org`, Hammarström et al., 2019), elicitation glosses (Concepticon, `https://concepticon.clld.org`, List et al., 2020), and phonetic transcription systems (CLTS, `https://clts.clld.org`, List et al., 2019a), substantial parts of linguistic datasets could already be retro-standardized *without* the format specification proposed by the CLDF initiative. However, as the detailed description of the tools which support the creation of CLDF datasets will show, a lot of the heavy lifting can be substantially facilitated with the workbench we have built around the CLDF framework.

3.1. Linking Data to Glottolog

Somewhat surprisingly, language identification is difficult and creates a high barrier for data re-use in many cases.

Glottolog has been introduced to address exactly this issue, and do so in a more effective way than ISO 639-3, for three major reasons: (1) Glottolog provides many alternative names used for languages in other catalogs, archives, or databases such as WALS or AIATSIS. (2) Glottolog is supplemented by a Python API `pyglottolog` (see `https://pypi.org/project/pyglottolog`), providing programmatic access to this data, in particular local full-text using the `Whoosh` search engine. (3) Last but not least, Glottolog data are available as CLDF dataset – allowing standardized access from computing environments other than Python.

3.2. Mapping Data to Concepticon

That elicitation glosses used in fieldwork questionnaires, comparative wordlists, and other larger linguistic data collections are notoriously difficult to handle has long been noted by linguists, although most linguists simply put up with the problem without working towards a solution. With the publication of the Concepticon project (List et al., 2016), a first attempt to link the large amount of concept lists consistently has been presented, and so far turned out to be very successful, specifically in supporting the aggregation of lexical data from different resources (List et al., 2018b), as exemplified by the Database of Cross-Linguistic Colexifications (`https://clics.clld.org`, Version 2.0, List et al., 2018a).

The first version of the Concepticon contained 160 concept list, which made up for 30,222 elicitation glosses which were in part linked to 2,495 distinct concept sets. Different contributors had been assembling these data since at least 2012. The most recent version of the Concepticon (2.3.0) has increased this number to 310 lists, 65,979 elicitation glosses, and 3,677 distinct concept sets (List et al., 2020). It is important to note that the concrete work on the Concepticon project has not been increasing since its first publication. The reason why the project could keep, if not speed up, its original pace was due to improved approaches to data curation, consistency checking, and, specifically, semi-automated approaches for *concept mapping*, that were later refined by experts.

The essential idea of this concept mapping algorithm, which can be applied to concept lists in 29 different languages (of which, however, apart from English only German, Spanish, Russian, Portuguese, French, and Chinese are well supported), is to make active use of the elicitation behavior of linguists that has been accumulated in the Concepticon project. The Concepticon gloss for "aubergine", for example, is "AUBERGINE", but quite a few concept lists elicit this concept as English *eggplant*, this information is readily employed in the mapping process and greatly facilitates the search for standard concept sets defined by the Concepticon.

In addition to the concept mapping algorithm shipped along with the `pyconcepticon` Python package, a local web-based mapping procedure is available, which is implemented in JavaScript and uses a slightly modified mapping algorithm that allows also to check for spelling errors on a word-by-word basis across seven languages (see List et al. 2018b for a description of the algorithm and `https://digling.org/calc/concepticon` for the most recent online version of this mapping procedure).

Although the Concepticon mapping algorithms are not employing any machine learning architecture so far, but rather based on a simple stemming procedure combined with a handcrafted decision-tree-like search, it has turned out to be extremely efficient, and it also improves with each concept list being added.

3.3. Transforming Transcriptions into CLTS

Contrary to what many scholars think, the International Phonetic Alphabet can barely be considered a *standard* for transcription, and it is also not intended as such by the International Phonetic Association. What it constitutes is a mere attempt to provide a transcription system that is theoretically capable of describing the major phonetic contrasts in the languages of the world (IPA, 1999).

Linguistic practice usually acknowledges the IPA and makes occasional use of it, but the degree to which scholars adhere to the most recent updates of the IPA differs greatly among linguistic subfields, especially also, since some fields use their own transcription systems, such as, among others, the Uralic Phonetic Alphabet (UPA), the North-American Phonetic Alphabet (NAPA), or the specific, barely explicitly mentioned traditions of transcribing South-East Asian languages (Anderson et al., 2018).

In addition to transcription *systems*, cross-linguistic *datasets* that document phonetic and phonological aspects of languages, such as, for example, sound inventories (Moran et al., 2014; Nikolaev et al., 2015; Maddieson et al., 2013), also differ, at times drastically, in the way they make use of the IPA recommendations.

As a result, phonetic transcriptions are still largely incomparable across resources. Since scholars interpret transcription systems loosely and often prefer to stick to informal conventions of transcription that evolved over time in their specific peer groups, it is impossible to interpret a given phonetic transcription without knowing the author, the time, and target languages it was used to document.

The goal of the Cross-Linguistic Transcription Systems initiative (List et al., 2019a) was to provide a reference catalog that could help to unify linguistic datasets that provide data in phonetically transcribed form. Typical examples for these data are the numerous fieldwork notes for various languages of the world (Allen, 2007; Mitterhofer, 2013), larger collections of comparative word lists (Greenhill et al., 2008; Heggarty et al., 2019), or individual grammars and glossaries.

CLTS offers a standardized set of currently 8754 sounds which are uniquely defined by a feature system that is largely based on the feature system employed by the IPA. These sounds are linked to 15 transcription datasets and 5 different phonetic alphabets. The recommended standard, labelled "BIPA" ("broad", i.e., broad-coverage, IPA) has the advantage of providing an explicit solution for many issues resulting from an inconsistent combination of transcription symbols (e.g., if aspiration should be marked before or after palatalization, as in [k^{hj}] vs. [k^{jh}]). In addition, CLTS offers a Python library that can be used to *generate* new sounds that are virtually defined as part of the feature system, as well as to identify the most likely sound in the CLTS system, if a transcription does not follow the standards of any of the transcription systems available in the CLTS.

Converting data according to the CLTS recommendations can be further facilitated with help of *orthography profiles* (Moran and Cysouw, 2018). Orthography profiles can be seen as lookup tables for sound segments provided in any phonetic transcription written from left to right. If only the lookup table is provided, an orthography profile converts a given sound sequence into sound segments, by segmenting different sounds with a space (e.g., [t^{h} ɔ x t^{h} ə r] instead of [t^{h}ɔxt^{h}ər]). Additionally, they can also be used to *convert* the items in the lookup table into another transcription system by providing additional columns with transcription targets. In this form, orthography profiles can be used quite efficiently to retro-standardize the transcriptions provided in linguistic resources. The LingPy software package (`http://lingpy.org`, List et al. 2019) also offers a function to seed an initial orthography profile from a given dataset, which has proven very efficient for the preparation of larger comparative datasets (Sagart et al., 2019).

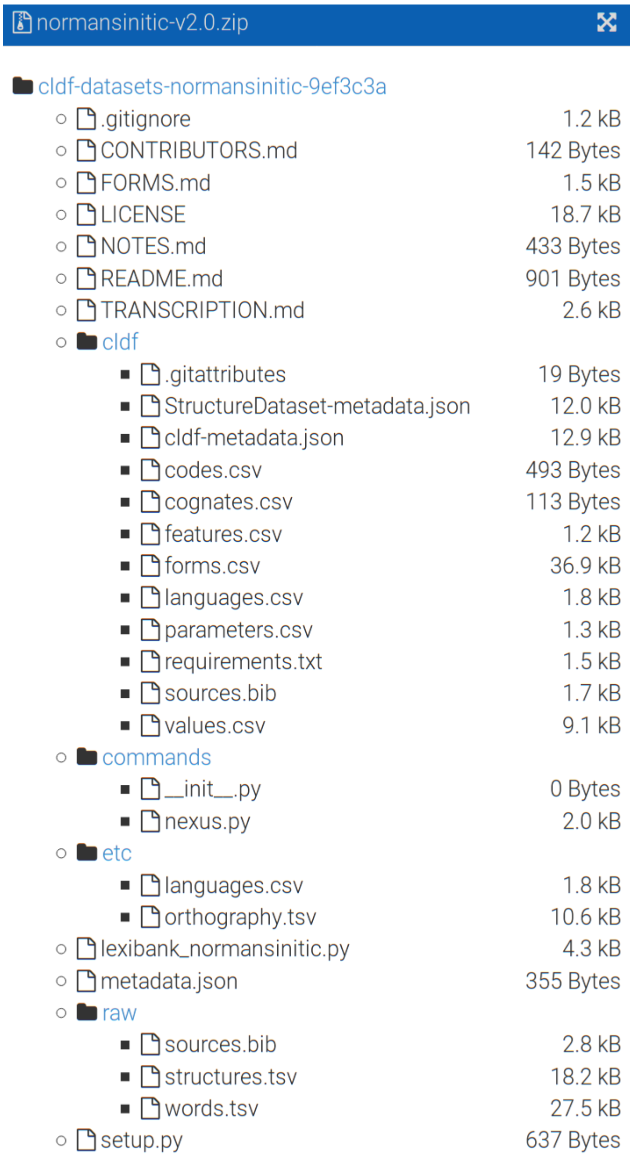
4. CLDFBench

4.1. Basic Idea

The requirements for “good” CLDF datasets laid out above, the existence of large amounts of published linguistic data, and the experience that was gained over a decade of data curation within the CLLD project (`https://clld.org`) were the motivating factors behind a new tool, the Python package `cldfbench` (`https://pypi.org/project/cldfbench`). Technically, `cldfbench` provides a layer on top of the `pycldf` package that provides low-level access to read, write, and validate CLDF datasets (`https://pypi.org/pycldf`). For the practice of data curation and retro-standardization, `cldfbench` provides a controlled and efficient way of creating and curating CLDF datasets, well-suited for integration with version control and long term archiving of data and code.

With CLDF adding linguistically relevant semantics to the CSV package format, and `pycldf` providing functionality to read and write CLDF datasets from Python, `cldfbench` addresses the concrete use cases of creating CLDF from existing resources.

`cldfbench` provides an efficient and transparent workflow for curating legacy data which allows for the replicable conversion of legacy formats to CLDF, and provides a data layout which transparently delineates (1) the *source data*, (2) *enhancements and configuration data*, such as linkings of languages to Glottolog, mappings of elicitation glosses to Concepticon, or orthography profiles (which provide a mapping of a source transcription system to the standardized B(road)IPA transcription system provided by CLTS), and (3) derived CLDF data (see also Figure 2).



```
normansinitic-v2.0.zip

cldf-datasets-normansinitic-9ef3c3a
  ○ .gitignore                                   1.2 kB
  ○ CONTRIBUTORS.md                           142 Bytes
  ○ FORMS.md                                     1.5 kB
  ○ LICENSE                                     18.7 kB
  ○ NOTES.md                                  433 Bytes
  ○ README.md                                 901 Bytes
  ○ TRANSCRIPTION.md                             2.6 kB
  ○ cldf
      ■ .gitattributes                         19 Bytes
      ■ StructureDataset-metadata.json          12.0 kB
      ■ cldf-metadata.json                      12.9 kB
      ■ codes.csv                             493 Bytes
      ■ cognates.csv                          113 Bytes
      ■ features.csv                             1.2 kB
      ■ forms.csv                               36.9 kB
      ■ languages.csv                            1.8 kB
      ■ parameters.csv                           1.3 kB
      ■ requirements.txt                         1.5 kB
      ■ sources.bib                              1.7 kB
      ■ values.csv                               9.1 kB
  ○ commands
      ■ __init__.py                             0 Bytes
      ■ nexus.py                                 2.0 kB
  ○ etc
      ■ languages.csv                            1.8 kB
      ■ orthography.tsv                         10.6 kB
  ○ lexibank_normansinitic.py                    4.3 kB
  ○ metadata.json                             355 Bytes
  ○ raw
      ■ sources.bib                              2.8 kB
      ■ structures.tsv                          18.2 kB
      ■ words.tsv                               27.5 kB
  ○ setup.py                                  637 Bytes
```

Figure 2: File structure of a `cldfbench` package released with Zenodo.

4.2. The Bigger Picture of Data Curation

The `cldfbench` framework is built to operate in a bigger environment for research data curation, with the ultimate goal of making the curation of FAIR data easy. By providing a workflow to convert legacy data to CLDF, `cldfbench` makes datasets **I**nteroperable. Datasets curated with `cldfbench` are suited perfectly for version control with tools like `GIT` (`https://git-scm.com/`) and by extension for collaborative curation on platforms like GitHub (`https://github.com`). If datasets are hosted on GitHub, continuous quality control can be implemented using services such as Travis-CI (`https://travis-ci.org/`). Thus, `cldfbench` encourages *transparent*, *replicable* data curation – which is essential for making constantly evolving data **R**eusable.

Datasets curated on GitHub are easily “hooked up” with Zenodo (`https://zenodo.org`) for archiving. Thus, datasets curated with `cldfbench` can easily be made **F**indable and **A**ccessible. Using the collection management functionalities such as *Communities* on Zenodo (https://zenodo.org/communities) further enhances findability, and additionally enables post-publication review and quality assessment.

4.3. Workflows for Retro-Standardization

A typical workflow for retro-standardization within the curation framework supported by cldfbench consists of five steps. In a first step (1), the data need to be *digitized* (unless they have already been digitized or have been originally submitted in digital form). In a second step (2), the data are converted to CLDF, following specific procedures depending on the data type (see our usage example below). In a third step (3), once the data are considered good enough to be shared, the dataset is versionized. In a fourth step (4), which can be automatized if users have linked their GitHub accounts with Zenodo, the version is archived with Zenodo. In order to make sure that the data are easy to find by colleagues, a fifth step (5) involves the further characterization of the data through Zenodos community system.

4.4. Usage Example

In order to illustrate the advantages of our CLDF curation workflow with cldfbench, we have selected a considerably small language resource offering both lexical and structural data. The data stems from an article by Norman (2003). In this article, the author shares data on Sinitic (Chinese) dialect varieties along with the unclassified Tibeto-Burman language Bai (collected from his own field work and occasional secondary sources). The data are presented in tables that can be easily digitized, since the amount of data is considerably small, providing 14 structural features for 11 Sinitic varieties, and 40 word forms (cognate morphemes) for 10 Sinitic varieties plus Bai (Jianchuan dialect).

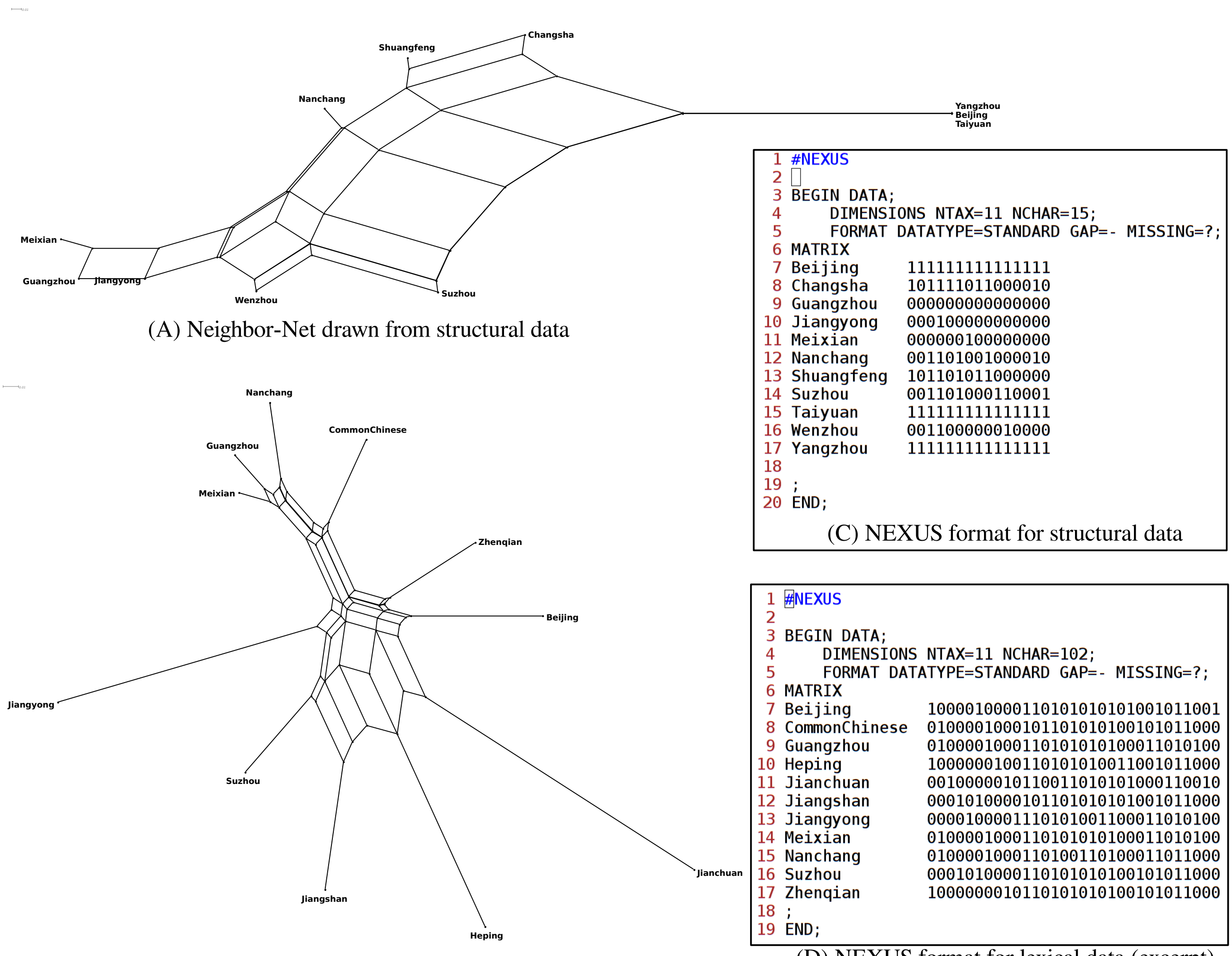
Our digitization step (step 1 in our workflow) consists in extracting the data tables (lexical and structural data) from the article and rendering them in *long-table format*. Representing the data in this form is not a necessary requirement by cldfbench but has the advantage that it facilitates the parsing of the text files when converting the data to CLDF in the second step. Apart from extracting the data from the sources, we also need to render the original sources that the author used to prepare the data, which is done by providing a BibTeX file. Figure 1 shows how the original lexical data (A) and the original structural data (B) are rendered in tabular form (B and D), and how the sources are rendered as bibliographic entries in BibTeX format. When converting the data to CLDF format (step 2 in our workflow), we first need to provide the links to our reference catalogs. To link the languages in the data to Glottolog, we create a list of all the 16 language varieties that occur in the dataset, along with geographic locations, Glottocodes, and additional information. We did not have to link the concept list underlying the lexical data to Concepticon, since this was already done in the past (see https://concepticon.clld.org/contributions/Norman-2003-40), but we created an orthography profile with help of LingPy's preliminary orthography profile creation method which can be invoked directly via cldfbench (as a sub-command of the lexibank handler for word list data). With this information, the data can be conveniently converted to CLDF with help of cldfbench.

The retro-standardization of a given dataset with help of cldfbench is done in form of a Python package. A typical Python package for cldfbench consists of a metadata file (metadata.json), a license file (LICENSE), a setup file to install the data as a Python package registered with cldfbench (setup.py), and a main Python script that takes care of the data lifting (in our case called lexibank_normansinitic.py). Raw data are stored in a specific folder (raw/), as are additional information pertaining to reference catalogs, such as Glottocodes or orthography profiles (etc/), or additional, user-provided commands (commands/). The lifted data in CLDF format itself is stored in the folder cldf. Optionally, specific information on the contributors who curated a given dataset can be provided in form of Markdown files (CONTRIBUTORS.md), or additional information on the dataset can be shared (NOTES.md). When invoking the CLDF conversion via cldfbench, not only the CLDF dataset is created, but also additional information. Thus, when dealing with wordlists, for example, information on the transcription system (TRANSCRIPTION.md) or the treatment of word forms in the data (FORMS.md) is automatically created upon converting the data to CLDF. Figure 2 shows a typical document tree of a cldfbench package, as rendered once submitted to Zenodo (https://zenodo.org).

Once a cldfbench package has been created, the conversion to CLDF can be invoked via the commandline (see https://github.com/cldf/cldfbench for details), similarly to user-defined commands provided along with the cldfbench package. Installation of our example dataset, which is curated on GitHub (https://github.com/cldf-datasets/normansinitic) can be done in a straightforward way with the help of the version control tool git. Installing the data, following our online instructions, will automatically install all necessary dependencies.

In order to emphasize the advantage of converting one's data into the standard CLDF formats, we added a custom command to convert the datasets into the NEXUS format (Maddison et al., 1997), a widely used standard format to represent evolutionary data in biology, which serves as the basic input format for many software packages. In Figure 3, we show splits graphs of the two datasets, computed with help of the Neighbor-Net algorithm (Bryant and Moulton, 2004) as implemented in the SplitsTree package (Huson, 1998) along with the corresponding NEXUS files of both datasets. While the NEXUS file for the structural data could be generated directly from the dataset, the lexical data had to be automatically searched for cognates before. For this purpose, standard algorithms from LingPy were employed (List, 2014). As can be easily seen from the two Neighbor-Nets, the structural data fails to discriminate the different dialect varieties considerably, while the lexical data results in a more or less balanced network that reflects the major dialect groups as they have been proposed in the literature.

The cldfbench package of Norman's dataset used in this



(A) Neighbor-Net drawn from structural data

(B) Neighbor-Net drawn from lexical data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=15;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing     111111111111111
Changsha    101111011000010
Guangzhou   000000000000000
Jiangyong   000100000000000
Meixian     000000100000000
Nanchang    001101001000010
Shuangfeng  101101011000000
Suzhou      001101000110001
Taiyuan     111111111111111
Wenzhou     001100000010000
Yangzhou    111111111111111

;
END;
```

(C) NEXUS format for structural data

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=11 NCHAR=102;
    FORMAT DATATYPE=STANDARD GAP=- MISSING=?;
MATRIX
Beijing        100001000011010101010101001011001
CommonChinese  010000100010110101010100101011000
Guangzhou      010000100011010101010100011010100
Heping         100000010011010101010011001011000
Jianchuan      001000000101100110101010100110010
Jiangshan      000101000010110101010101001011000
Jiangyong      000010000111010100110001101010 0
Meixian        010000100011010101010100011010100
Nanchang       010000100011010011010001101100 0
Suzhou         000101000011010101010100101011000
Zhenqian       100000001011010101010100101011000
;
END;
```

(D) NEXUS format for lexical data (excerpt)

Figure 3: Neighbor-Net analysis of the datasets, based on their export to NEXUS format as required by the SplitsTree software package.

demonstration was released as version 2.0.0 and has been archived with Zenodo (List and Forkel, 2019). In order to experiment with the data, and to try and replicate the CLDF creation process or the additional conversion to NEXUS format, all users have to do is to download the data (directly from GitHub or from Zenodo) and follow the detailed instructions from the `cldfbench` package.

5. Conclusion

The original publication of the first version of the CLDF standard was meant as basic tool for scholars to produce FAIR data. While we could illustrate the suitability of CLDF on many occasions so far, and specifically demonstrate it with the release of Version 2.0 of the CLICS database, where we managed to aggregate cross-linguistic polysemy data from 15 different datasets, it has so far been rather tedious, specifically for scholars less proficient in programming, to make use of the CLDF standards in their own work. With `cldfbench`, we hope to make a first step into the direction of potential users who might want to try to lift their own or other published datasets to a higher level of cross-linguistic comparability. First tests of `cldfbench` have been carried out successfully, as reflected specifically in the recent release of CLICS3 (`https://clics.clld.org`), in which we managed to double the data basis underlying CLICS2 (List et al., 2018a; List et al., 2018b), which is now based on more than 2000 languages and more than 2000 concepts (List et al., 2019d; Rzymski et al., 2020), and has also been successfully used to test hypotheses on emotion concepts in the languages of the world (Jackson et al., 2019). In the future, we hope to further increase the amount of retro-standardized datasets. Specifically structural data is often published in form of tables in books and journals, where the data is accessible, but by no means interoperable. By assembling these data in retro-standardized form and making them broadly accessible on Zenodo, we hope to encourage colleagues to join us in our efforts of rendering cross-linguistic datasets more comparable.

Supplementary Data and Code

The `cldfbench` Python package is curated on GitHub (`https://github.com/cldf/cldfbench`), archived on Zenodo (Version 1.0, `https://doi.org/10.5281/zenodo.3518698`), and available from the Python package managment system

PyPi (`https://pypi.org/project/cldfbench`) along with its dependencies. The example dataset by Norman (2003) is curated on GitHub (`https://github.com/cldf-datasets/normanisinitic`) and archived with Zenodo (Version 2.0, `https://doi.org/10.5281/zenodo.3552559`).

Acknowledgements

JML was supported by the ERC Starting Grant 715618 “Computer-Assisted Language Comparison” (`http://calc.digling.org`). We thank Simon Greenhill, Christoph Rzymski, Mei-Shin Wu, and Tiago Tresoldi for testing the `cldfbench` package in different stages of its development.

6. Bibliographical References

Allen, B. (2007). *Bai Dialect Survey*. SIL International, Dallas.

Anderson, C., Tresoldi, T., Chacon, T. C., Fehn, A.-M., Walworth, M., Forkel, R., and List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Berez-Kroeker, A. L., Gawne, L., Smythe Kung, S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., and Woodbury, A. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1):1–18.

Bryant, D. and Moulton, V. (2004). Neighbor-Net. *Molecular Biology and Evolution*, 21(2):255–265.

Matthew S. Dryer et al., editors. (2011). *The World Atlas of Language Structures online*. Max Planck Digital Library, Munich.

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(180205):1–10. https://doi.org/10.1038/sdata.2018.205.

Forkel, R., Bank, S., and Rzymski, C. (2019). *CLLD: A toolkit for cross-linguistic databases (Version 5.0.0)*. Max Planck Institute for the Science of Human History, Jena. https://doi.org/10.5281/zenodo.3437148.

Greenhill, S. J., Blust, R., and Gray, R. D. (2008). The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283.

Hammarström, H., Haspelmath, M., and Forkel, R. (2019). *Glottolog (Version 4.1)*. Max Planck Institute for the Science of Human History, Jena.

Haspelmath, M. and Tadmor, U. (2009). *World Loanword Database*. Max Planck Digital Library, Munich.

Heggarty, P., Shimelman, A., Abete, G., Anderson, C., Sadowsky, S., Paschen, L., Maguire, W., Jocz, L., A., M J. A., Wägerle, L., Appelganz, D., do Couto e Silva, A. P., Lawyer, L. C., Cabral, A. S. A. C., Walworth, M., Michalsky, J., Koile, E., Runge, J., and Bibiko, H.-J. (2019). Sound comparisons: A new online database and resource for research in phonetic diversity. In Sasha Calhoun, et al., editors, *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 280–284, Canberra. Asutralasian Speech Science and Technology Association.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73.

IPA, I. P. A. (1999). *Handbook of the International Phonetic Association. A guide to the use of the international phonetic alphabet*. Cambridge University Press, Cambridge.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Mucha, P. J., Forkel, R., Greenhill, S. J., and Lindquist, K. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Key, M. R. and Comrie, B. (2016). *The intercontinental dictionary series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Lewis, W. D. and Xia, F. (2010). Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *LLC*, 25:303–319.

List, J.-M. and Forkel, R. (2019). cldf-datasets/normansinitic: Structural and lexical data for the paper by Norman (2013) on Chinese dialect classification (Version 2.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.3552559.

List, J.-M., Cysouw, M., and Forkel, R. (2016). Concepticon. A resource for the linking of concept lists. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2393–2400. European Language Resources Association (ELRA). https://doi.org/10.5281/zenodo.47143.

List, J.-M., Greenhill, S., Anderson, C., Mayer, T., Tresoldi, T., and Forkel, R. (2018a). *CLICS: Database of Cross-Linguistic Colexifications. (Version 2.0)*. Max Planck Institute for the Science of Human History, Jena.

List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., and Forkel, R. (2018b). CLICS2. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306. https://doi.org/10.1515/lingty-2018-0010.

List, J.-M., Anderson, C., Tresoldi, T., Rzymski, C., Greenhill, S., and Forkel, R. (2019a). *Cross-Linguistic Transcription Systems. Version 1.3.0*. Max Planck Institute for the Science of Human History, Jena.

List, J.-M., Greenhill, S., Tresoldi, T., and Forkel, R. (2019b). *LingPy. A Python library for quantitative tasks in historical linguistics*. Max Planck Institute for the Science of Human History, Jena.

List, J.-M., Hill, N. W., and Foster, C. J. (2019c). Towards a standardized annotation of rhyme judgments in Chinese historical phonology (and beyond). *Journal of Language Relationship*, 17(1):26–43. https://doi.org/10.17617/2.3149513.

List, J.-M., Rzymski, C., Tresoldi, T., Greenhill, S., and Forkel, R. (2019d). *CLICS3: Database of Cross-Linguistic Colexifications (Version 3.0)*. Max Planck

Institute for the Science of Human History, Jena. http://clics.clld.org/.

List, J. M., Rzymski, C., Greenhill, S., Schweikhard, N., Pianykh, K., Tjuka, A., Wu, M.-S., and Forkel, R. (2020). *Concepticon. A resource for the linking of concept lists (Version 2.3.0)*. Max Planck Institute for the Science of Human History, Jena.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. DÃ¼sseldorf University Press, DÃ¼sseldorf.

List, J.-M. (2018). Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences*, 5(10):1–14. https://doi.org/10.5281/zenodo.1474750.

Maddieson, I., Flavier, S., Marsico, E., CoupÃ©, C., and Pellegrino., F. (2013). LAPSyD: Lyon-Albuquerque Phonological Systems Database. In *Proceedings of Interspeech*.

Maddison, D. R., Swofford, D. L., and Maddison, W. P. (1997). NEXUS: an extensible file format for systematic information. *Syst. Biol.*, 46(4):590–621, Dec.

Mitterhofer, B. (2013). *Lessons from a dialect survey of Bena: Analyzing wordlists*. SIL International.

Moran, S. and Cysouw, M. (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin.

Steven Moran, et al., editors. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Nikolaev, D., Nikulin, A., and Kukhto, A. (2015). *The database of Eurasian phonological inventories*. RGGU, Moscow.

Norman, J. (2003). The Chinese dialects. In Graham Thurgood et al., editors, *The Sino-Tibetan languages*, pages 72–83. Routledge, London and New York.

Östling, R. (2014). Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–1127, Stroudsburg.

Pollock, R., Tennison, J., Kellogg, G., and Herman, I. (2015). Metadata Vocabulary for Tabular Data. Technical report, World Wide Web Consortium (W3C).

RfII, R. f. I. (2019). Herausforderung Datenqualiät. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel. Technical report, RfII, Göttingen.

Rzymski, C., Tresoldi, T., Greenhill, S., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., Chang, S., Lai, Y., Morozova, N., Arjava, H., HÃ¼bler, N., Koile, E., Pepper, S., Proos, M., Epps, B. V., Blanco, I., Hundt, C., Monakhov, S., Pianykh, K., Ramesh, S., Gray, R. D., Forkel, R., and List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies. *Scientific Data*, 7(13):1–12.

Sagart, L., Jacques, G., Lai, Y., Ryder, R., Thouzeau, V., Greenhill, S. J., and List, J.-M. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322. https://doi.org/10.1073/pnas.1817972116.

Tennison, J., Kellogg, G., and Herman, I. (2015). Model for Tabular Data and Metadata on the Web. Technical report, World Wide Web Consortium (W3C).

W3C, W. W. W. C. (2019). JSON-LD 1.1. A JSON-based serialization for Linked Data. Technical report, World Wide Web Consortium. https://www.w3.org/TR/2019/WD-json-ld11-20191112/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3.