FAIRification of LeiLanD

Eric Sanders[1], **Sara Petrollino**[2], **Gilles R. Scheifer**[3],
Henk van den Heuvel[1], **Christopher Handy**[4]
[1]CLS/CLST - Radboud University Nijmegen, [2]LUCL - Leiden University,
[3]Department of Social Sciences - University of Luxembourg,
[4]Informatisering en Facilitaire Zaken - Leiden University
eric.sanders@ru.nl

Abstract

LeiLanD (Leiden Language Data) is a searchable catalogue initiated by the Leiden University Centre for Linguistics (LUCL) with the support of CLARIAH. The catalogue contains metadata about language datasets collected at LUCL and other institutes of Leiden University. This paper describes a project to FAIRify the datasets increasing their findability and accessibility through a standardised metadata format CMDI so as to obtain a rich metadata description for all resources and to make them findable through CLARIN's Virtual Language Observatory. The paper describes the creation of the catalogue and the steps that led from unstructured metadata to CMDI standards. This FAIRification of LeiLanD has enhanced the findability and accessibility of highly diverse collection of language datasets.

Keywords: corpus collection, metadata conversion, FAIR

1. Introduction

LeiLanD is a large collection of descriptions of Leiden databases. Its idea stemmed from the desire to have an overview of language data collected by researchers at Leiden University[1] and to understand the curation needs and availability of the datasets. The overview was created through a Qualtrics[2] questionnaire administered to individual researchers by a team of student assistants. Individual interviews produced a first overview of the datasets collected at LUCL (and beyond), along with rich, semi-structured descriptions about the content of the datasets. The original questionnaire was drafted following a bottom up approach: the type of information requested for each dataset was selected on the basis of researchers' needs, and these were cross-checked with existing metadata standards for language data. The resulting general overview was information rich but unstructured, lacking in standard vocabularies and terminology conventions.

In order to structure and standardise the metadata, they were then mapped to the Dublin Core standard (Weibel and Koch, 2000) where possible. However, Dublin Core metadata definitions were found insufficient to account for the great diversity of metadata content collected through the questionnaires, and therefore the whole catalogue had to be reformatted to a richer metadata standard.The Component Metadata Infrastructure (CMDI) and its hierarchical and modular setup (Windhouwer and Goosen, 2022) offered the framework to accomplish this. This CMDIfication of the metadata was considered an important step in making LeiLanD's dataset more FAIR. The FAIR principles[3] were formulated in 2016 (Wilkinson et al., 2016) to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. CMDI is an interoperable format to share metadata which is used in CLARIN's Virtual Language Observatory[4] (VLO) (Van Uytvanck et al., 2012) to make datasets more findable and (its metadata) more accessible. The goal of this paper is to describe this process.

In the next section we give a short overview of LeiLanD, Section 3 describes what was done to reformat the metadata to CMDI, Section 4 shows how the CMDIfied metadata were integrated into the VLO. In the final section we present our conclusions.

2. LeiLanD

2.1. Initial Setup

LeiLanD[5] is a user-friendly, searchable catalogue which currently contains metadata for 146 language datasets collected between 1960 and 2020. The type of data collected at LUCL and at Leiden University is very diverse, ranging from written language data and recorded speech sounds,

[1] The original project was funded by LUCL with a grant for the Language Diversity in the World research profile area awarded to Felix Ameka, Nivja de Jong and M. Carmen Parafita Couto (2017); further financial support was provided by CLARIAH-PLUS, WP3 (2021).

[2] www.qualtrics.com

[3] see also https://www.go-fair.org/fair-principles/

[4] https://vlo.clarin.eu/

[5] https://leiland.lucdh.nl/

LREC-COLING 2024, pages 7101–7106
20-25 May, 2024.

to EEGs. Table 1 shows an overview of the types of data available in the LeiLanD collection.

Table 1: Types of data in LeiLanD

Broad data types	Number of datasets
Written	63
Spoken	88
Signed	6
Singing	2
EEG	4
Grammar judgments	2

Datasets represent major (spoken) European languages such as Dutch and English, lesser studied languages of Africa, Asia and The Americas, sign languages and languages that are now extinct. These monolingual, bilingual and multilingual datasets are annotated in various ways, and may include basic translations, transliterations and glossing, POS-tagging, reconstructed cognate forms and linked data annotations. Datasets in the catalogue represent various disciplines of linguistics including computational linguistics, language description and documentation, psycholinguistics, sociolinguistics, historical linguistics, language acquisition and many more. The catalogue is expected to grow as new datasets are collected and added to LeiLanD.

LeiLanD provides all sort of metadata, and for some categories it contains richer descriptions than those recommended by Dublin Core and CMDI standards. For instance, Dublin Core's and CMDI's recommended practices for "language" include only ISO 639-2 or ISO 639-3 language codes, whereas LeiLanD additionally links each language to its Glottocode, a unique and stable identifier which accounts for all types of languages, dialects and language families (Forkel and Hammarström, 2022; Hammarström et al.). Next to the standard metadata such as author and contact person for each dataset, geographical and temporal provenance, types of annotations (understood as any enrichment of the data, e.g. transcriptions, translations, glossing, POStagging, etc.), modality (i.e. signed, spoken, written), linguality type (monolingual, bilingual, multilingual), language proficiency (L1, L2, L3), gender and age of the speakers, average number of speakers for each language dataset, format and size of the dataset, and domains of linguistic research represented by a dataset, LeiLanD offers additional information about the methods of data collection (for example whether data was gathered through sociolinguistic surveys, narratives, elicitation, written texts and so on), and information over software used or developed specifically for a dataset. Accessibility options are also indicated: of 146 datasets, 28 are archived in online repositories. Access to datasets which are not directly available online can be requested on the LeiLanD web interface [6].

2.2. Website

The LeiLanD website provides multidimensional search over twelve categories in the metadata content of the collection, through intuitive dropdown boxes that allow any arrangement of data. Figure 1 shows the search options of the LeiLanD website.

As the number of projects stored in LeiLanD increases, retrieval scales automatically, as our twelve categories reduce nearly any complex searches to a manageable number of results. This system is completely dynamic, allowing for modification/addition/deletion of items in these categories as well as the categories themselves. We expect that most new projects will fit comfortably within the twelve metadata categories already available in the system.

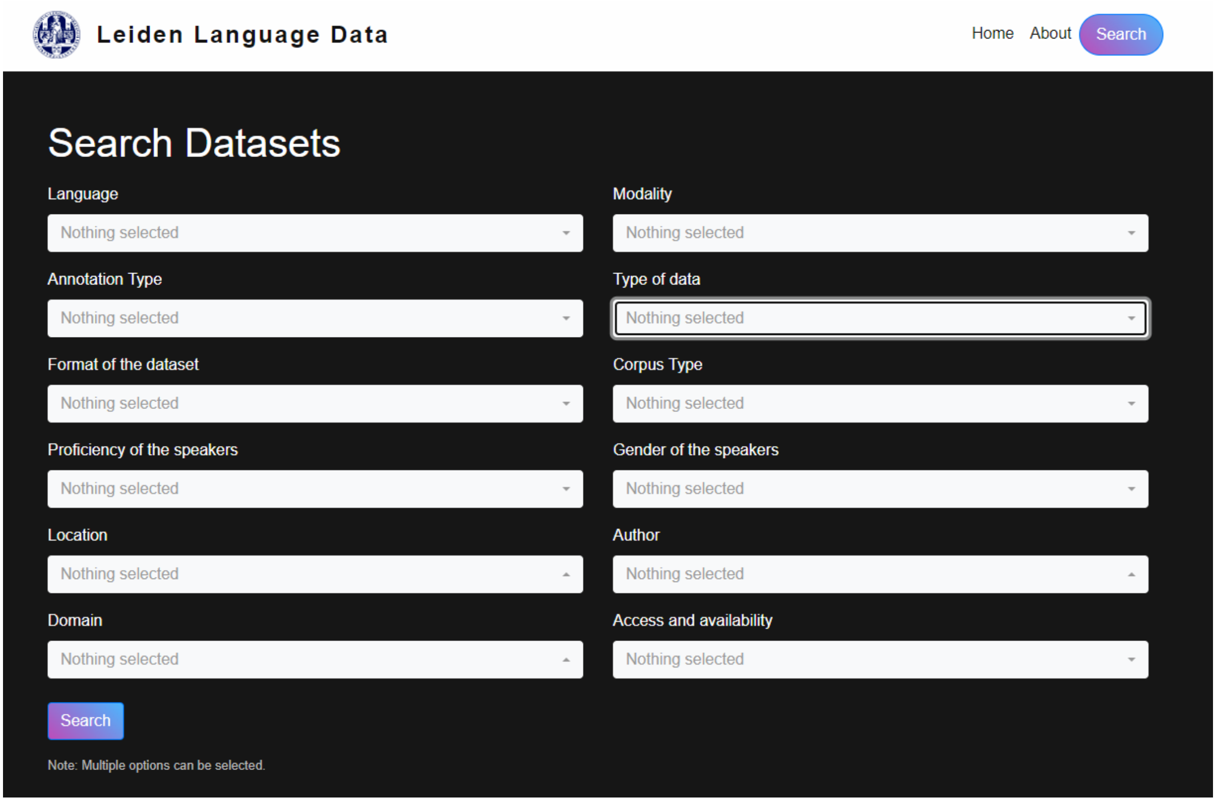
The concept for the LeiLanD website developed through several stages over a number of years, and its realisation as a usable software product likewise went through several computer programming languages. Initial designs were easily implemented in PHP. As new records were added to the database, the PHP version became cumbersome to maintain, and a more robust application was created in Java. We recently ported LeiLanD to Python/Django, preserving the interface style of the Java version while also adding new information fields and several novel features. One major advantage of migrating LeiLanD to Python is that future extension is accomplished easily through the inclusion of various data processing and data visualisation libraries within the Python ecosystem, allowing for new methods for manipulating data and metadata contained in the LeiLanD database. Now we can move from retrieving static data to dynamic analysis of data on the server and client side in addition to the base LeiLanD functionality, and interface easily with other online language projects such as dictionaries and linguistic parsing utilities. The Python/Django version of LeiLanD also makes it easy to add functions for creating, editing and deleting records through the Django administrative interface and custom management commands.

The entire LeiLanD codebase is hosted open source on GitHub[7]. We hope to see new and extended instances of LeiLanD implemented for other types of linguistic databases in the future.

Data conversion between formats occurs through the following pipeline:

[6] For a complete overview see https://leiland.lucdh.nl/

[7] https://github.com/handyc/leiland


Leiden Language Data

Home About Search

Search Datasets

Language: Nothing selected

Modality: Nothing selected

Annotation Type: Nothing selected

Type of data: Nothing selected

Format of the dataset: Nothing selected

Corpus Type: Nothing selected

Proficiency of the speakers: Nothing selected

Gender of the speakers: Nothing selected

Location: Nothing selected

Author: Nothing selected

Domain: Nothing selected

Access and availability: Nothing selected

Search

Note: Multiple options can be selected.

Figure 1: Screenshot of the search page on the LeiLanD website

Excel XLS → CSV → Django ORM → SQL DB → Django ORM → CSV → Excel XLS

Researchers can view and import datasets in familiar word processing environments while also allowing the LeiLanD system to archive datasets in a more robust and scalable database format.

3. Mapping and Reformatting to CMDI

The LeiLanD metadata were collected via questionnaires sent to the creators/owners of the databases. Metadata were mapped to Dublin Core metadata definitions where possible, but this method was discovered to be insufficient for describing all metadata elements. Metadata were stored in a MS Excel spreadsheet.
In order to provide access to the LeiLanD metadata in the VLO, the metadata of the datasets had to be provided in a suitable format for the VLO, the original metadata had to be reformatted to CMDI format. For this we used the CMDI profile ”CorpusCollection”[8] that was created as a generic profile for metadata of linguistic corpora and datasets.
Our primary concern here was to make an appropriate mapping from the LeiLanD metadata to the CorpusCollection profile. For many metadata elements there was a straightforward mapping. For example, the values in the LeiLanD field *proficiency of the speakers* are “L1, L2, L3” and these could be easily matched with the ”native, non-native” values in the CMDI category *lingualityNativeness*. However, this mapping introduced some data loss, reducing the available scale from three to two categories. Other metadata elements that we could not incorporate in the CMDI profile were the replies to the following questions from the Qualtrics survey:

- Q11 Were the participants’ proficiencies tested?
- Q13 Curation needs
- Q49 Are there any other ethical comments relating to the database?
- Q18 Resource history: - (in case it’s relevant) When was it transcribed?

It should, however, be noted that this metadata was not included in the LeiLanD metadata either, so its absence is strictly speaking, not a result of the metadata mapping operation.
Other metadata elements required some degree of creativity. For example, correspondence with the CMDI field *DCType* and its relevant subtypes (Dataset:treebank; Dataset:table;Dataset:lexicon; Dataset:list-lemmas;Dataset:list-POStags; Image:still-map; Image:still-graph; Image:still-picture; Sound:speech; Sound:music, etc.) had to

[8] Identifier: clarin.eu:cr1:p_1493735943947

be established by matching the LeiLanD field *Type of data* to *DCType*, and by using the values found in another LeiLanD field, the *Modality* field. The original LeiLanD field *Type of Data* includes values such as “Audiovisual, Audio, Photo, Written” while the field *Modality* contains the values “Spoken, Written, Signed”. The matching of LeiLanD *Type of data* to CMDI *DCType* could be done in a fairly straightforward fashion: values such as “Audio” were translated into CMDI “Sound:”. However, for the more refined matching with the *subtype* values, the relevant information had to be drawn from the LeiLanD *Modality* field. If the value in the *Modality* field was ”Spoken”, we inferred that the audio data was about spoken languages, and we accordingly selected the CMDI *subtype* ”sound:speech”. For other *DCType* and *subtype* values the mapping was more straightforward: for instance the LeiLanD value “Audiovisual” had a direct matching with the CMDI value “Image:moving-film”.

Several elements required splitting the original LeiLanD metadata to map them onto two CorpusCollection elements or combining two elements together to map to one CorpusCollection element. For example, the original LeiLanD metadata stored age and number of speakers in a single field titled *Age and Number of participants*, and the data were used to inform, accordingly, both CMDI categories *lingualityAgeGroup* and *speechCorpus:numberOfSpeakers*. Data matching the CMDI field *domain* were drawn from three LeiLanD metadata fields: *Subject*, *Subject (keywords)* and *Other disciplines*.

Adaptations of the LeiLanD metadata were made in the Excel spreadsheet. We created a custom Python script to ingest these metadata from a CSV file (from Excel) and output XML CMDI files according the CorpusCollection profile. All CMDI files were validated using the Linux tool xmllint.

Table 2 shows the different metadata fields as they were categorised in LeiLand, Dublin Core and CMDI.

4. Integration in Collection Bank

In order to make the CMDI metadata files harvestable for the VLO, we decided to store the data in the Collection Bank[9]. The Collection Bank application is an intermediary for storing CMDI metadata files of corpora, datasets and collections into the Centre for Language and Speech Technology’s (CLST, Nijmegen) OAI-PMH node[10] which is harvested on a regular basis by CLARIN’s VLO. The Collection Bank is a web-based database for storage of metadata of data collections. It was developed by the technical support group of the Humanities Lab at the arts faculty of Radboud University Nijmegen. It was developed to store the metadata of language and speech resource collections at Radboud University, but it is open to collections from other sites as well. A CMDI profile ’CorpusCollection’ was created and stored in the CLARIN Component Registry[11] to which all collections in the Collection Bank have to adhere. The metadata record entries in the Collection Bank have an Persistent Identifier (PID), which is presented in the LeiLanD entry of the database. Vice versa, the LeiLanD URL of the database is also available in the Collection Bank metadata entry of that database. New collections can be entered in the Collection Bank manually using the ’add’ page of the website, but especially for this project an XML import function was added to the application. During the integration of the LeiLanD CMDI files to the Collection Bank, a few small errors were found in both the CorpusCollection profile as in the Collection Bank website, that could be resolved easily. All collections in the Collection Bank are harvested periodically – since the LeiLanD metadata have been uploaded to the Collection Bank, they are available in the VLO. The following protocol was developed for cases in which new databases are added to LeiLanD: the database and all its metadata will be added to a structured file. A script will store the data to LeiLanD and at the same time create a CMDI-XML with the data in the CorpusCollection profile. This can be uploaded to the Collection Bank and the database will automatically be harvested and appear in the VLO. This can all be done in one place.

5. Conclusion

This paper gives an overview of the Leiden Language Data catalogue (LeiLanD), currently comprising metadata for nearly 150 language resources, from its inception in 2017, to its recent adaptations of the migration of the website from PHP to Java and then to Python, and the mapping of the metadata to CMDI in order to make it available in the VLO. Site functionality and layout of the website have been kept the same, but implementation of new features and software extensions, in addition to the creation and maintenance of datasets, is now much easier.The most recent software of the LeiLanD website was made available in Github, thus enhancing its findability and accessibility.

Similarly its metadata was FAIRified. Accomplishing that objective required converting to CMDI the unstructured metadata originating from the questionnaire administered to resource creators. Since

[9] https://applejack.science.ru.nl/collbank

[10] CLST, https://www.ru.nl/en/cls/clst is a CLARIN C Centre for metadata, see https://www.clarin.eu/content/overview-clarin-centres

[11] https://catalog.clarin.eu/ds/ComponentRegistry

we wanted to incorporate the metadata in the Collection Bank, we used an existing profile created for Collection Bank resources. In this way LeiLanD's metadata were FAIRified in the sense that they are now better findable, accessible and more interoperable. Mapping from the original metadata to the CMDI was sometimes straightforward and sometimes challenging, but we managed to fit all metadata into the profile to our satisfaction.

6. Acknowledgements

We would like to thank Erwin Komen for realising the necessary adaptations to the Collection Bank.

7. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Robert Forkel and Harald Hammarström. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6):917–24.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 4.7.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 1029–1034. European Language Resources Association (ELRA).

Stuart L Weibel and Traugott Koch. 2000. The dublin core metadata initiative. *D-lib magazine*, 6(12):1082–9873.

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.

Menzo Windhouwer and Twan Goosen. 2022. Component metadata infrastructure. In *CLARIN: The Infrastructure for Language Resources*, pages 191–222, Berlin, Boston. De Gruyter.

8. Appendix

A. Metadata Categories

Table 2: Fields used for metadata in respectively LeiLanD, Dublin Core and CMDI

LeiLanD	**Dublin Core term**	**CMDI CorpusCollection element**
Title of the dataset		title
Description	dc.description	description
Type of data	dc.type	resource.DCType & resource.modality
Number of Speakers		resource.speechCorpus:numberOfSpeakers resource.writtenCorpus.numberOfAuthors
Size	size. info	resource.totalSize.size resource.totalSize.sizeUnit
Annotations	additional.metadata	resource.annotation.type
Format of the dataset	dc.format	resource.media.format
Date created from	dc.dateCreated	provenance.temporalProvenance.startYear
Date created until	dc.dateCreated	provenance.temporalProvenance.endYear
Location of data collection	dc.coverage.placeName	provenance.geographicalProvenance. country.CountryName
Country Code	dc.language.rfc4646	provenance.geographicalProvenance. country.CountryCoding
Corpus Type	dc.subject	linguality.lingualityType
Proficiency of the speakers	dc.subject	linguality.lingualityNativeness
Language name	dc.language.rfc4646	Language.languageName
ISO language code	dc.language.iso	Language.ISO639.iso-639-3-code
Access rights	dc.accessRights	access.availability
Publisher	dc.publisher	access.website
Contact person	contact.person	access.contact.person
access:Contact:email		access.contact:email
Author	dc.contributor.author	resourceCreator.person