A core metadata schema for L2 data

Jennifer-Carmen Frey[1], Alexander König[2], Egon W. Stemle[1] and Magali Paquot[3]

[1]Institute for Applied Linguistics, Eurac Research, Italy, [2]CLARIN ERIC, The Netherlands, [3]CECL UCLouvain, Belgium

Introduction

> "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.
>
> National Information Standards Organization (2004): Understanding Metadata

> "Metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable.
>
> Higgins, 2007

- Comprehensive, domain-specific metadata paramount for findability, accessibility, interoperability and reusability (cf. FAIR principles for data stewardship, Wilkinson et al. 2016) of L2 data
- Extensive amount of L2 studies call for more comparability and standardization
- First attempt made by Granger/Paquot 2017

Methodology

1 **Original draft of Granger/Paquot 2017**

Core metadata schema v1.0 (presented in Gothenburg)

Administrative, design, learner, text & annotation metadata for L2 data

2 **Adaptation based on experience from applying it to existing L2 data (CECL, PORTA, SWELL,…)**

Core metadata schema v1.1 (presented at LCR 2022)

Additional items for FAIR compliance, revised proficiency metadata, components for tasks, annotators and transcribers

3 **Adaptation after feedback from LCR community**

Core metadata schema v2.0 (in preparation)

Incorporating ***best practice*** for consideration in new projects

Revised task metadata on situational and task characteristics, language background and exposure

Core components

- Universally applicable components
- Used to flexibly describe different types of learner corpora
- Corpus metadata can be linked to other components

Domain-specific modules

Domain-specific metadata modules can provide additional core metadata considered important for specific sub-domains of L2 studies, e.g.:

- Data collected from educational contexts
- Sign language corpora
- Multimodal corpora
- …

Example for school module

class_id
teacher_id
grade_level
school_language (medium of instruct.)
…

The core metadata schema

Situational and task metadata

Situation

- Number and type of interlocutors/addressees
- Medium and mode
- Circumstances (real-time, prepared, potentially revised and edited)
- Setting (education, work, family, leisure)
- Communicative purpose (expository, descriptive, persuasive, narrative…)
- Register (essay, summary, conversation, …)
- Topic domain (domestic, education, politics, art, sports, religion…)

Task

- Task instructions and additional materials
- Duration (in minutes)
- List of any reference tools allowed

Learner metadata

- Anonymous learner ID
- Socio-demographics
- Language background
- Target language and other languages: learning context, exposure, proficiency
- Other individual differences (motivation, attitude, aptitude, etc.)

Language background

- Home languages
- Parent languages
- Strongest languages
- Perceived L1(s)
- Languages spoken with friends
- Languages of education: kindergarten, primary, secondary and higher education
- Other used languages

Target language/other known languages: learning context, exposure, proficiency

- Learning context (instructed, immersive, naturalistic)
- Age of onset of acquisition
- Usage context (family, leisure, friends, public, work, education)
- Language proficiency
- Months spent in target language country

Corpus metadata

Administrative metadata

- PID, version, publisher, author, licence, contact mail, availability, access requirements, documentation, reference article,…
- *Meet FAIR requirements!*

Design metadata

- Target language(s), corpus size, data collection setting, …

Text metadata

- Unique identifier for language production
- Reference to any files related to it
- Linked to other components describing the learner, the situational context, annotations, etc.
- Version (if various versions exist)
- Time and place of creation/collection
- Target language
- Proficiency rating for text, CEFR conversion
- Score of official language testing if collected within language testing framework
- Token count

Annotator/transcriber metadata

- Anonymous annotator ID
- L1 and other spoken languages
- Target language competence
- Type (expert, trained, student, teacher, crowdsourcing,...)

Annotation metadata

- ID, Name or type of annotation
- Tools used for annotation and their version no.
- Automatic annotation or not
- Evaluation/correction of annotated data
- Reference to documentation
- Reference to manual annotator if relevant

Expected outcomes

- Better comparability between resources and results
- Data aggregation and composite research
- Data re-useability
- Raised awareness for FAIR issues
- Community engagement

Future perspectives

- CMDI profile in CLARIN component registry
- Development of an open-source online interface to fill in and export metadata in various formats
- Domain-specific modules provided by experts from the field

Sharing (metadata) is caring!

- Retrieve the schema!
- Send us comments and feedback!
- Use it for your own corpora!
- Tell your colleagues and students about it!

References

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., & Trippel, T. (2012). CMDI: a component metadata infrastructure. In Proceedings of the workshop describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources. LREC 2012, May 22, 2012, Istanbul, Turkey. (pp. 1-4). European Language Resources Association.

Granger, S. & Paquot, M. (2017). Towards standardization of metadata for L2 corpora. *Invited talk at the CLARIN workshop on Interoperability of Second Language Resources and Tools*, 6-8 December 2017, University of Gothenburg, Sweden.

Higgins, S. (2007). What are metadata standards? Digital Curation Centre. Standards Watch Papers. National Information Standards Organization (2004). Understanding metadata. Washington DC, United States: National Information Standards Organization.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Paquot, M., König, A., Stemle, E. Frey, J.C. (2023). Core Metadata Schema for Learner Corpora v1.1, https://doi.org/10.14428/DVN/4CDX3P

CLARIN

eurac research

CECL Centre for English Corpus Linguistics

CLARIN K CENTRE Learner Corpus Research