LC-meta: Learner corpus metadata

Standard and implementation(s)

Herb Lange

Research meeting, June 4th 2026

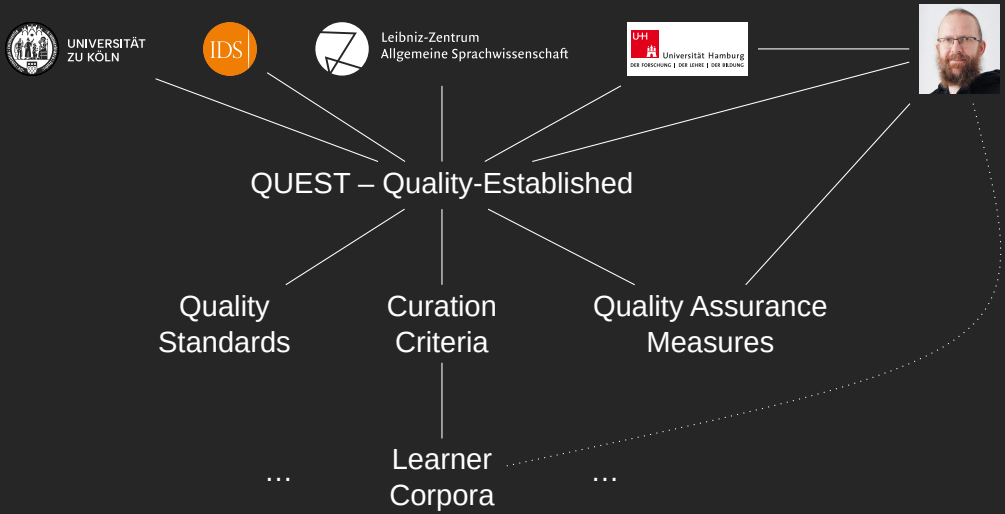
SPRÅKBANKENTEXT

Background

Why do I care?

SweLL

Why do I care?


UNIVERSITÄT ZU KÖLN
IDS
Leibniz-Zentrum Allgemeine Sprachwissenschaft
UHH Universität Hamburg
QUEST – Quality-Established
Quality Standards
Curation Criteria
Quality Assurance Measures
…
Learner Corpora
…

Why do I care?

Metadata Formats for Learner Corpora: Case Study and Discussion

Herbert Lange
IDS Mannheim
lange@ids-mannheim.de

December 2022: Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)

Learner corpus research

Written and spoken data produced by learners has always been a key resource for the study of second language acquisition (SLA). [...] Learner corpora gave rise to a flurry of studies, which have come to be grouped under the umbrella term of 'learner corpus research' (LCR). This new research strand emerged in the late 1980s as an offshoot of corpus linguistics [...]

(Granger et al. 2015, p 1)

Metadata

> *Metadata can broadly be defined as 'data about data'. When conducting a learner corpus study, metadata is crucial at every single step of the research process, from (1) study design, (2) corpus selection or compilation to (3) data analysis, and interpretation of the results.*

(Paquot et al. 2024, pp 280-281)

> *To put it differently, "metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable" (Higgins, 2007).*

(Paquot et al. 2024, p 282)

FAIR principles

Metadata is considered a key element of the FAIR principles, essential for maximising optimal reuse of research data. According to the FAIR principles, data and metadata should be Findable (for both humans and computers), Accessible (with clear guidance on how to access the data, possibly including authentication and authorization), Interoperable (with applications and workflows for analysis, storage and processing) and Reusable (containing any information needed by other researchers to effectively interpret and disseminate findings from subsequent use of the data) (Wilkinson et al., 2016).

(Paquot et al. 2024, pp 281-282)

FAIR principles

R1.3: (Meta)data meet domain-relevant community standards

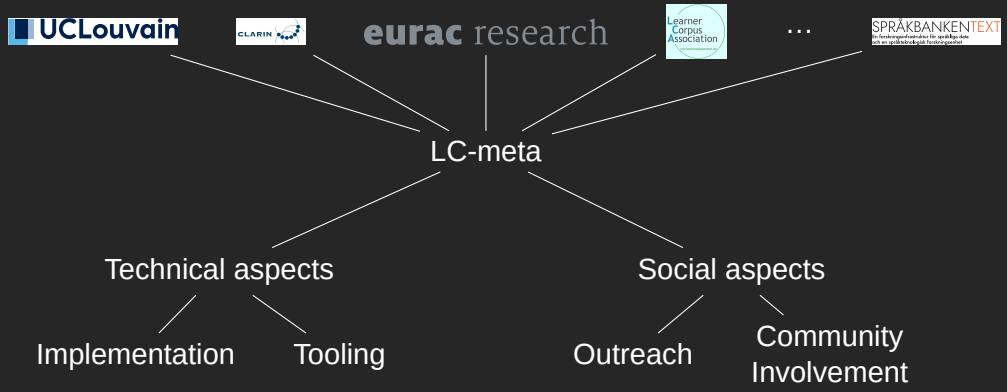
(https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/

> *To date, however, Learner Corpus Research (LCR) has not developed community standards or best practices for data collection, archiving and sharing (cf. König et al., 2021; Stemle et al., 2019; Volodina et al., 2018).*

(Paquot et al. 2024, p 282)

LC-meta

Working Group



Short History

- December 2017: Presentation: “Towards standardization of metadata for L2 corpora”
- September 2022: Presentation: first version of the Metadata schema at LCR conference
- May 2024: Metadata definition: “Core Metadata Schema for Learner Corpora (version 2)”
- June 2024: Research paper: “The Core Metadata Schema for Learner Corpora (LC-meta) Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research“

Short History

- November 2024: Call for interest: “Working Group on Metadata in Learner Corpus Research”
- September 2026: Presentation: “The Core Metadata Schema for Learner Corpora: developing user-friendly resources and addressing key challenges”

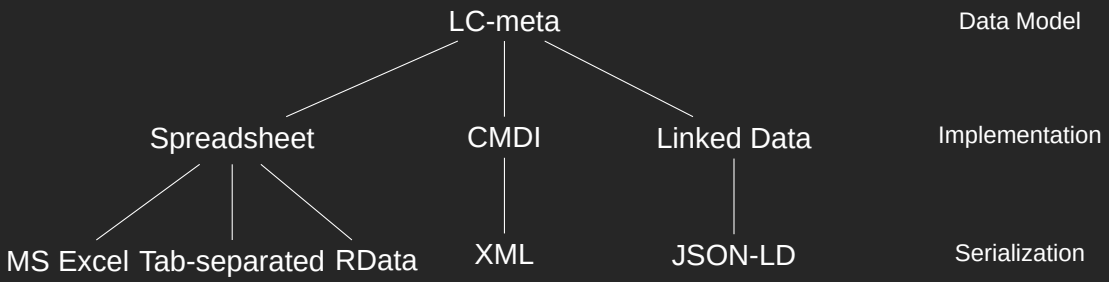
General Structure

- 168 elements, obligatory or optional, repeatable or singular
- eight interrelated categories/components
- five components considered obligatory (e.g. to meet FAIR principles) and three optional (but considered essential towards the development of data management best practices in learner corpus research)
- intended to maintain flexibility and simplicity within the schema
- “text” being used as an umbrella term for all types of learner language production

Design Principles

- Each metadata element follows the same organising principle
- Fixed vocabularies where possible: CEFR levels, language codes, …, but challenge to determine whether to use fixed attributes for specific metadata fields or keep them open-ended
- variable names and fixed attributes that are commonly used in the research community, but no closed definitions at this stage (because e.g. lack of widely agreed-upon definitions for terms such as L1, L2, and proficiency)

Interlude: Data model vs. implementation vs. serialization



CMDI implementation

- Metadata standard within CLARIN
- profiles, components, cues
- Machine-readable
- Hierarchical
- Support for cardinality, vocabularies, conditions

CMDI editor

- Web-based
- Full CMDI support
- Readily usable (with minor tweaks)

Case study: SweLL metadata

Challenges:

- Both structured and unstructured metadata spread over several files
- Some values corpus-specific (Swedish)

Other efforts

- Lime survey
- Linked data
- Office of the Chief Statistician at the World Bank metadata editor
- VLO for LCR

Questions?

Thank You

Literature

- Granger, Sylviane, et al. “Introduction: Learner Corpus Research – Past, Present and Future.” The Cambridge Handbook of Learner Corpus Research, edited by Sylviane Granger et al., Cambridge University Press (CUP), 2015, pp. 1–6, https://doi.org/10.1017/CBO9781139649414.001.
- Granger, Sylviane & Paquot, Magali “Towards standardization of metadata for L2 corpora” https://web.archive.org/web/20190919040434/https://sweclarin.se/sites/sweclarin.se/files/event_atachements/Granger_Paquot_Metadata_G%C3%B6teborg_final.pdf.
- Lange, Herbert “Metadata Formats for Learner Corpora: Case Study and Discussion” Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022), 2022, https://doi.org/10.3384/ecp190011.
- Paquot, Magali, et al. “The Core Metadata Schema for Learner Corpora (LC-Meta).” International Journal of Learner Corpus Research, vol. 10, no. 2, 2024, pp. 280–300, https://doi.org/10.1075/IJLCR.24010.PAQ.
- Paquot, Magali, et al. “Core Metadata Schema for Learner Corpora (Version 2).” V1, Open Data @ UCLouvain, 2024, https://doi.org/10.14428/DVN/AAUEM2.