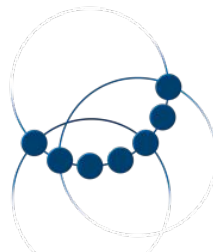


The Core Metadata Schema for Learner Corpora: Developing user-friendly resources and addressing key challenges

Magali Paquot, Jennifer-Carmen Frey, Alexander König,
Herbert Lange, Hubert Naets, Egon Stemle, & Julien Zakhia Doueih



Outline

- **The Core Metadata Schema for Learner Corpora** (Paquot et al., 2024)
 - What it is
 - Uses and users
 - LC-meta working group
- **From schema to user-friendly resources and interoperable metadata**
 - Online questionnaire intended to support the systematic and standardized collection of learner-related variables
 - Definition of metadata format representations
 - CMDI
 - Linked data
- **Challenges and future directions**

The Core Metadata Schema for Learner Corpora

LC-meta (Paquot et al., 2024)

- proposal of a set of universal metadata fields
- typical and relevant to all types of corpora used in LCR
- based on iterative community-feedback -> community standard
- 168 variables (72 obligatory)
 - Repeatable variables, conditional variables
- 8 interrelated components: administrative details, corpus design, learner-related variables, text-related variables, situational and task-related metadata, and where applicable, annotation, annotator, and transcriber metadata

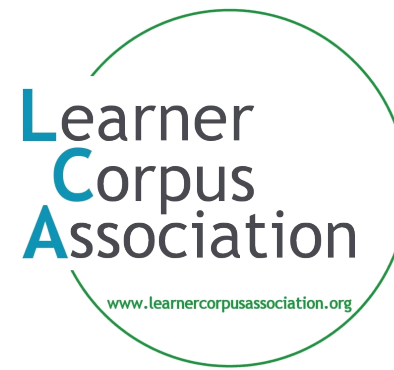
Obligatory	Optional
Unique identifier Age at text production Gender Target language Learning context L1	Highest level of education Socio-economic status Country Region Learner impairment Age of onset of acquisition

Uses and users

- Describe existing corpora
 - Common learner-related variables
 - Better discovery and comparison
- Collect new data
 - Young German Learner English (YGLE) corpus (Werner et al., 2024)
 - STAR corpus (De Cock et al., 2025)
- Document & share corpora
- Compare & combine datasets
 - Identify compatible / comparable learner populations and variables
 - DAKODA corpora (Portmann et al., forth.)
- Corpus creators, researchers, repositories/infrastructures, learner corpus community

LC-meta working group

- Maintain and further develop the schema
- Listen & respond to community needs & feedback
- Align LC-meta with evolving L2 research practices and standards
- Coordinate dissemination, training and community engagement
- Foster community uptake and sustainable development



From enthusiasm to reality ...

Strong community engagement

- 10-15 active members
- 4 task forces addressing key priorities: metadata format representation, user perspective, community feedback, and visibility/adoption

But implementation comes with challenges

- Excel is not a sustainable solution for collecting, managing and sharing metadata
- Need to make LC-meta accessible and usable across the entire ecosystem, from individual corpus compilers to large research infrastructures

How can we turn a community-developed schema into a genuinely **user-friendly, interoperable and sustainable resource**?

From schema to user-friendly resources
and interoperable metadata

The corpus compiler's perspective

Learner variables

- **From metadata fields to learner-facing questions**

Transforming variable names in the metadata model into questions learners can understand and answer

- **Towards systematic data collection**

Developing an online questionnaire to support the standardized collection of learner-related variables at the data collection stage

Key challenge: How can we bridge the gap between a complex metadata model, the practical needs of corpus compilers, and the learner experience of completing a questionnaire?

From metadata fields to learner-facing questions

learner_L1
learner_parent_L1
learner_country
learner_language_education_primary
learner_target_language_onset
learner_target_language_instruction
learner_L2_language_onset
learner_L2_language_proficiency
learner_educational_background
learner_impairment

Are you currently attending language classes in [the target language]?

DOCUMENTATION

Purpose : To determine whether the learner was receiving formal language instruction at the time of data collection.

From metadata fields to learner-facing questions

learner_L1	→
learner_parent_L1	
learner_country	
learner_language_education_primary	
learner_target_language_onset	
learner_target_language_instruction	
learner_L2_language_onset	
learner_L2_language_proficiency	
learner_educational_background	
learner_impairment	

What is your first language? What is your mother tongue?

What language(s) did you first acquire in early childhood?

By this, we mean the language(s) you were first exposed to and learned in your home environment during childhood. These are often, but not necessarily, the language(s) you are most proficient in or use most frequently.

DOCUMENTATION

Purpose: To collect the language(s) that the learner considers their first language(s). If you use a different conceptualisation of first language in your corpus, or decide to use a different question to operationalise this variable, please document this decision clearly in the corpus documentation.

Towards systematic data collection: technical implementation (for now ...)

Lime Survey

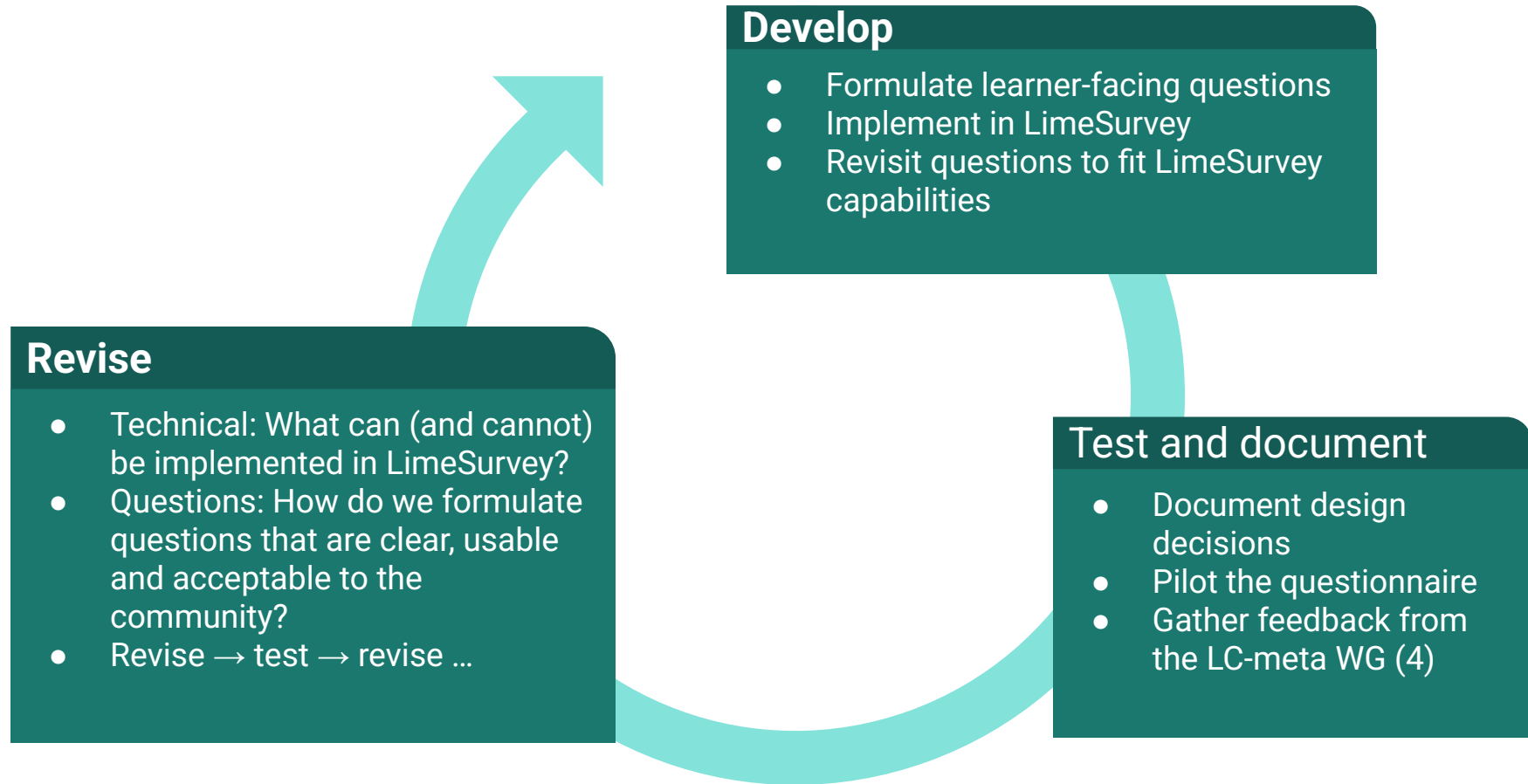
Advantages	Challenges
Open-source and freely available Widely used in research and higher education Flexible questionnaire design Conditional logic and branching GDPR/privacy features Easy data export	Generic survey tool $\neq$ metadata interface Interface constraints can influence question design Questionnaire maintenance when LC-meta evolves

Iterative process: schema $\rightarrow$ question $\rightarrow$ interface $\rightarrow$ revision

Schema's complex structure (blocks of repeatable questions, etc.) $\rightarrow$ JavaScript (but higher technical expertise required, higher risk of breaking after LimeSurvey updates, etc.)

Turning LC-meta into a questionnaire

A very time-consuming experience: bear with us!



Target release: end of 2026

The research infrastructure's perspective

FAIR principles as a guiding framework

- Make metadata interoperable, reusable and machine-actionable
- Support discovery, exchange and reuse across tools and infrastructures
- Enable metadata to be understood and processed beyond individual projects

A key infrastructure question

- How can LC-meta be represented in formats that are compatible with existing research infrastructures and technologies?

More than one representation is possible

- Our approach: two formats (CMDI and RDF).
- Why? Different formats serve different purposes and infrastructures

CMDI – An XML based format

Established metadata format within CLARIN infrastructure → Visibility, discoverability and interoperability within CLARIN

- Modular and hierarchical (XML) → Reduces duplication and machine-readable by design
- Sufficiently expressive for the LC-meta schema
 - supports optionality, cardinality and conditional elements
 - controlled vocabularies help ensure consistency and avoid input errors
- Supported by a mature ecosystem of tools: creation/editing and validation

CMDI – Present and Future

Currently

- **Complete translation** of LC-meta V2 into a CMDI profile
- **Alpha** version of LC-meta CLARIAH CMDI forms

Short term

- **Extensive testing** on various corpora
- **Further integration** with CLARIN infrastructure and services. Goal: improved discoverability

Long term

- - **Establish** Learner Corpus Research as a recognised community within CLARIN
- - **Facilitate** the discovery, interoperability and reuse of learner corpus metadata

A RDF-based format

Specialised format for knowledge representation, here used to describe learner corpora

- Represents **complex, nested data** and **potentially cyclic** data
- **Combines** easily data from multiple sources → **Extends** the metadata with corpus-specific properties or external vocabularies
- **Represents missing information** explicitly including the reason
- **Accommodates** legacy and incomplete corpora
- Most importantly: **Single format for data and metadata**

Based on widely used internet (W3C) standards such as:

RDF (*Resource Description Framework*), OWL (*Web Ontology Language*), SPARQL

A RDF-based format – Present and Future

Currently

- First **beta version** released internally
- Metadata can be handled directly with standard linked data tools

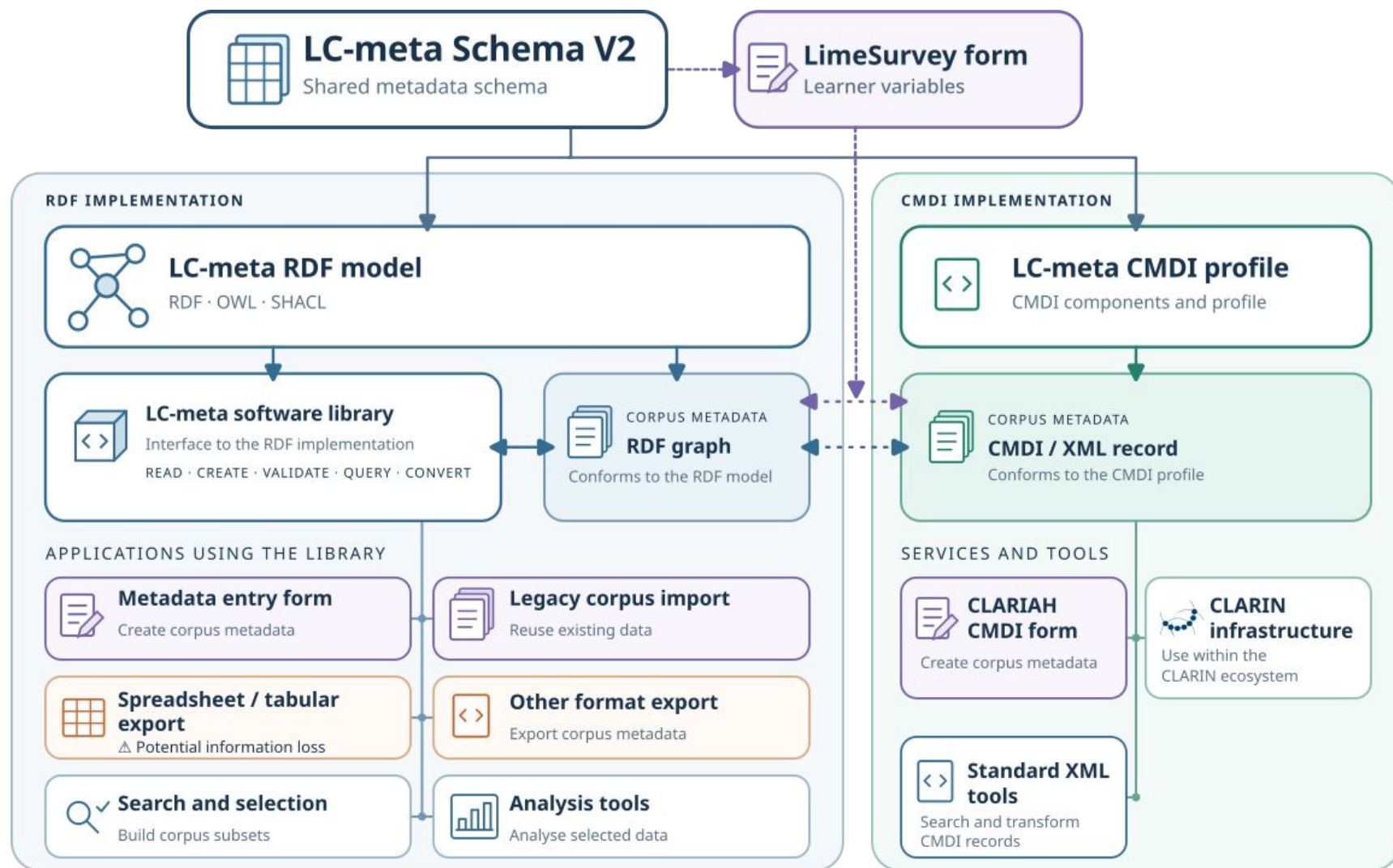
Short term

- **Extensive testing** on various corpora
- Creation of a **dedicated software library** that makes it easier to manage the format and which can be integrated into various tools

Longer-term aim

- Going beyond metadata, same format for metadata, data and annotations
- Development of specialized data acquisition, conversion, and exploration tools

The LC-meta ecosystem



Challenges and future directions

Slow progress, limited resources...

- **No dedicated funding or personnel** to support the development and maintenance of LC-meta
- Working group activities and task forces rely largely on **voluntary contributions**
- Progress towards user-friendly resources has therefore been slower than anticipated
- This creates a bottleneck for the work of other task forces:
 - community feedback and validation
 - visibility and adoption
 - training and dissemination
- Wider uptake of LC-meta ultimately depends on resources that corpus compilers and research infrastructures can easily use/understand

...but fundamental work!

Future priorities for LC-meta WG

- Develop user-friendly tools and resources that enable the community to fully adopt and use LC-meta
- Ensure long-term maintenance, governance and technical support
- Support training, documentation and community uptake
- Integrate LC-meta with existing research infrastructures and metadata standards

Long-term vision

- **An integrated learner corpus portal:** From a shared metadata schema to a sustainable, interoperable ecosystem supporting the collection, documentation, sharing and long-term reuse of learner corpus data and metadata

How can we, as a community, make sure that this fundamental work receives the sustainable support it needs?

Thank you for your attention!
And special thanks to all the LC-meta WG members for their work and
patience!

From metadata fields to learner-facing questions

learner_L1
learner_parent_L1
learner_country
learner_language_education_primary
learner_target_language_onset
learner_target_language_instruction
learner_L2_language_onset
learner_L2_language_proficiency
learner_educational_background
learner_impairment

In which country have you lived for most of your life?

Purpose: To identify the participant's primary linguistic and cultural environment, either over their lifetime or in recent years, which may influence language use and learning.

Use the following alternative if you want a recent residence perspective relevant to language exposure:

In what country have you spent most of the last 5 years?

if time, illustrate with one example

copy-paste from LimeSurvey

Beyond the spreadsheet: modeling learner corpora as connected data

Allow the corpus domain to be modelled in a more connected and flexible way, and to capture corpus designs and relationships in greater detail

- L2 and L1 profiles and language productions
- Learners and native-speakers and their language productions
- Cross-sectional and longitudinal corpora
- One or multiple producers and languages
- Original productions and successive transformations