Towards FAIR Metadata for Specialised Corpora: A Community-Informed Empirical Study of Schema Development in Two Communities

Egon W. Stemle[†]
Institute for Applied Linguistics
Eurac Research
Bozen-Bolzano, Italy
egon.stemle@eurac.edu

Alexander König
CLARIN ERIC
The Netherlands
alex@clarin.eu

Nannan Liu
SUNY at Binghamton
United States
nliu9@binghamton.edu

Jennifer-Carmen Frey[†]
Institute for Applied Linguistics
Eurac Research
Bozen-Bolzano, Italy
jennifer.frey@eurac.edu

Hubert Naets[‡]
CKL2CORPORA
UCLouvain
Belgium
hubert.naets@uclouvain.be

Magali Paquot[‡]
CKL2CORPORA
F.R.S.-FNRS
UCLouvain
Belgium
magali.paquot@uclouvain.be

Mariachiara Russo
University of Bologna
Italy
mariachiara.russo@unibo.it

Abstract

This paper investigates the development of domain-specific metadata schemata to support FAIR data management within specialised research communities. We report on two case studies, the Core Metadata Schema for Learner Corpora (LC-meta) and the metadata schema developed for interpreting corpora within the Unified Interpreting Corpus (UNIC) project. We address how metadata are described and standardised in these projects, the principles that informed the schema design, and the infrastructural challenges that arise when semantically rich, community-specific metadata are to be represented within generic discovery frameworks. Our findings underline the importance of domain-oriented approaches to metadata standardisation, iterative design and stakeholder involvement, and suggest a more active role for CLARIN in supporting long-term sustainability, technical mediation, and shared stewardship.

1 Introduction

The *FAIR Guiding Principles for scientific data management and stewardship* (Wilkinson et al., 2016) have become a cornerstone of contemporary data management, promoting Open Science by ensuring that research outputs are Findable, Accessible, Interoperable, and Reusable. Existing infrastructures, such as CLARIN B-centres (Hinrichs & Krauwer, 2014) and data repositories like Zenodo[1] and Figshare[2], facilitate *F*indability and *A*ccessibility to a certain extent, but achieving true *I*nteroperability and *R*eusability remains a challenge (Bhardwaj et al., 2025; Frey et al., 2020). One important reason is that the generic metadata standards adopted by these repositories are designed for a broad range of content types

[†]Eurac Research CLARIN Centre (ERCC), K-centre for CMC and Social Media Corpora (CKCMC)

[‡]C-centre & K-centre for Learner Corpora (CKL2CORPORA)

[1]https://zenodo.org/

[2]https://figshare.com/

and therefore often lack the granularity required to capture the methodological, linguistic and contextual specificities of specialised corpora.

These limitations become particularly visible in federated metadata environments, whose primary objective is to aggregate heterogeneous resources from different repositories or infrastructures through a shared discovery layer, as, for example, in CLARIN's Virtual Language Observatory (VLO) (Van Uytvanck et al., 2012) and related cross-repository discovery services. Federation relies on a common descriptive baseline that enables cross-domain indexing and querying, which typically requires metadata to be normalised and reduced to generic elements. While effective for high-level discovery, this approach introduces a structural tension with the needs of specialised research communities, whose corpora depend on fine-grained, domain-specific metadata to support particular methodological paradigms and reuse scenarios. Such distinctions are often difficult to express within generic schemata and risk being flattened or rendered inaccessible in federated settings. In practice, this creates a recurring dilemma, as participation in infrastructures such as CLARIN is essential for visibility and sustainability. However, reliance on generic metadata representations can undermine Interoperability and Reusability by leaving crucial contextual information undocumented or inaccessible.

Domain-specific metadata standards have already been developed for some CLARIN communities, for example, for parliamentary corpora (Erjavec & Pančur, 2021) and cultural heritage data (Riley, 2024). However, many research communities still lack such standards, even though effective Interoperability and Reusability require the metadata schemata to be tailored to the specific characteristics of the data and their intended reuse contexts. Developing such schemata is itself a demanding and resource-intensive endeavour, requiring expertise that spans both the research domain and technical infrastructures, as well as careful negotiation between community-specific needs and more general, cross-domain approaches.

To address some of these challenges, illustrate possible approaches to solutions and highlight persistent problems, this study examines two community efforts to develop domain-specific metadata standards. The first is the Core Metadata Schema for Learner Corpora (LC-meta) (Paquot et al., 2024), developed collaboratively by researchers and infrastructure experts through a community-informed, iterative process. The second is the metadata schema for interpreting corpora (Liu & Russo, 2025) developed within the Unified Interpreting Corpus (UNIC) project and implemented on the UNIC platform.[3] Methodologically, UNIC builds on the path opened by LC-meta and extends it under different empirical, technical and ethical conditions.

By analysing these examples, the paper aims to highlight strategies, challenges and lessons learned in designing metadata standards that balance domain-specific needs with broader Interoperability goals. In addition, it aims to open a broader discussion on the role of CLARIN in supporting research communities, in particular by facilitating the sharing of expertise, promoting methodological convergence, and providing infrastructure and guidance for the development and maintenance of domain-specific metadata standards.

2 Metadata Standards and the Case for Domain-Specific Adaptation

Metadata, often described as 'data about data' (Greenberg, 2005), initially referred to simple descriptive information, such as titles, authors, or dates, that helped identify and organise datasets. Over time, the concept has expanded to encompass a rich system of structured information that documents not only the content of a dataset, but also its provenance, internal structure, semantics, and conditions of creation and use (Gilliland, 2008). Modern metadata captures definitions, rules, contextual information, and the methods used to generate derived data, enabling users to understand, interpret, and evaluate the significance and reliability of digital resources. In this sense, metadata is no longer a peripheral administrative tool but a critical component of research infrastructure, underpinning data transparency, Interoperability, and meaningful reuse across disciplines.

The development of metadata schemata has been central to data stewardship initiatives in digital hu-

[3] https://unic.dipintra.it/ – Code available under the Apache 2.0 License at https://github.com/F-technology-srl/unic

manities and linguistics. Frameworks such as the Text Encoding Initiative (TEI),[4] Dublin Core[5], and the Component MetaData Infrastructure (CMDI)[6] provide formalised structures for describing digital resources, offering Interoperability across disciplines and data types. These general frameworks establish a common foundation for resource description, enabling sharing and integration beyond the original context of creation. At the same time, they often allow some level of domain-specific adaptation, so that communities can extend schemata to capture features relevant to their own research.

However, in practice, these adaptations are often limited, and the general nature of these frameworks can make it difficult to fully represent the specific characteristics and needs of particular research domains. In the CLARIN ecosystem, metadata quality is often uneven across corpora and resource types (Fišer et al., 2018), and some user groups report dissatisfaction with the searchability and transparency of metadata in CLARIN's VLO (Lušicky & Wissik, 2017). These findings highlight the persistent gap between the availability of metadata schemata and their actual usability in supporting resource discovery and reuse.

The effectiveness of metadata schemata depends not only on adherence to formal standards but also on the degree to which they enable accurate, context-sensitive interpretation and reuse of resources (Greenberg, 2005). In other words, metadata must be both technically correct and conceptually aligned with the practices, expectations, and values of the relevant research community. As Velterop and Schultes (2021) argue, 'metadata need to comply with standards that are relevant for the domains and disciplines involved …[and] they need to be properly applied drawing on the controlled vocabularies created by or in close consultation with the practising domain experts'.

Domain-specific metadata schemata are therefore essential. By capturing the unique characteristics, requirements, and interpretive frameworks of resources within a particular field, they help ensure that data can be discovered, understood, and reused accurately. Beyond supporting Interoperability within a research community, such schemata contribute to broader Open Science objectives, for example, by enhancing transparency, enabling replication, and facilitating comparative or meta-analytic studies, thereby improving overall study quality (e.g., Paquot (2024)).

3 Case Study I - Core Metadata Schema for Learner Corpora (LC-meta)

In the following, we present our first case study (see Section 4 for the second) that has developed a domain-specific metadata schema and show how metadata is described and standardised. We discuss design principles and underlying values, as well as the technical implementation of the metadata framework. This will provide valuable insights and feed into the development of a blueprint in Section 5 for community-informed, domain-specific metadata.

LC-meta (Paquot et al., 2024) is a domain-specific metadata schema designed for the Learner Corpus Research community. Its primary goals are to support FAIR resource creation, ensure comparability across resources, facilitate cumulative research, and promote best practices in the field.

The schema includes 168 elements that were identified as broadly relevant across learner corpus research. These elements are either obligatory, to ensure FAIR compliance and comparability, or optional, to support best practices and facilitate future research. The schema was developed through a collaborative, community-informed, iterative process, consisting of three phases: a draft version based on reviewing existing resources to prompt community feedback, iterative refinements based on testing and community feedback, all leading to the publication of a FAIR resource and dissemination and community entrustment via the creation of a dedicated working group with experts from various institutions.

3.1 General Structure of the Schema: Eight Interrelated Components

LC-meta includes 168 partly obligatory, partly optional elements, organised into eight interrelated main components:

(1) Administrative metadata,
(2) Corpus design metadata,

[4]https://tei-c.org/guidelines/
[5]https://www.dublincore.org/specifications/dublin-core/dces/
[6]https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

(3) Learner metadata,
(4) Text metadata,
(5) Situational and task-related metadata,
(6) Annotation metadata,
(7) Annotator metadata,
(8) Transcriber metadata.

The component-based structure is intended to support detailed and structured metadata descriptions while remaining compatible with existing language resource infrastructures, in particular component-based approaches such as CMDI.

Rather than describing all information in a single, flat record, LC-meta separates metadata components by function and scope. This enables clearer documentation, avoids unnecessary duplication, and facilitates linking between different aspects of a corpus, such as participants, texts, annotations, and contributors to transcription and annotation. The components are designed to be combined flexibly, allowing the schema to accommodate a wide range of learner corpus designs.

The eight components are explicitly interrelated through references and persistent identifiers. Except for administrative and corpus design metadata, all components are repeatable. This makes it possible to represent multiple learners, texts, situational contexts or annotation layers within a single corpus. Administrative, corpus design, learner, text, situational and task-related metadata form the core descriptive backbone of learner corpora and are therefore obligatory. By contrast, the annotation, annotator and transcriber components are optional. These latter components focus on data provenance and promote best practices in well-documented data reporting.

As the aim of LC-meta was to offer a comprehensive set of metadata fields that apply to all types of corpora in the field, elements specific to certain research contexts (e.g. corpora for sign languages or corpora collected in educational settings) are not included in the core metadata schema but should be described within separate components yet to be developed.

Below, we briefly describe the eight components of the core metadata schema, provide examples of the elements contained in each and explain the rationale for the proposed components.

(1) Administrative metadata provides high-level information about the learner corpus as a resource. It includes elements required for identification (corpus name, acronym and persistent identifier), citation, versioning, documentation, and access conditions. These metadata support core infrastructure functions such as persistent identification, discovery in repositories, and long-term maintenance. As such, this component, which consists mainly of obligatory elements, aligns closely with metadata typically required for depositing resources in repositories and is crucial for ensuring compliance with FAIR principles (cf.König et al., 2021; Lindström Tiedemann et al., 2018). By explicitly modelling administrative information as a separate component, LC-meta supports consistent data management practices and facilitates integration with existing repository and catalogue systems.

(2) Corpus design metadata documents the conceptual and methodological framework underlying a learner corpus. Elements in this category concern, for example, the language(s) represented in the corpus, whether texts are single- or multi-authored, the mode(s) of communication and other characteristics such as longitudinality of the data as a whole. While more specific metadata might be encoded document-by-document at the text level, this component describes the overall scope of the resource, including the types of data collected, the linguistic focus, and key design decisions made during corpus compilation. This component plays a crucial role in interpretability and comparability across learner corpora. Making design assumptions explicit helps users find relevant resources in a register, understand the structure of a resource as a whole and assess whether a given corpus is comparable to other resources and/or (re-)usable for other research questions.

(3) Learner metadata describes the individuals who contributed language data to the corpus. The component contains basic socio-demographic variables such as the participant's age, gender, and first language, as well as more domain-specific items, including proficiency levels and more detailed information on a learner's language background, multilingual profiles, motivation, or the presence of learning

difficulties. Instead of treating participant information as attributes of texts, LC-meta models learners as independent entities that can be linked to one or more texts. This approach supports common use cases in learner corpus research, such as longitudinal designs or corpora in which multiple texts are associated with the same learner. It also avoids redundancy by allowing learner information to be recorded once and reused across multiple data objects. At the same time, this component also provides the possibility to include information on a participant's experience with and exposure to one or various other languages by filling out sub-components that are themselves repeatable, enabling the modelling of multilingual profiles with exposure to various languages and by leveraging the structural ideas behind CMDI (Broeder et al., 2012).

(4) Text metadata describes individual units of learner language production. As such, the term text is broadly defined to include written, spoken, and multimodal data. This repeatable component contains, for each text, a unique and anonymous identifier for the text itself, as well as descriptive information about the Administrative metadata: it provides high-level information about the learner corpus as a resource. text. It also contains references to associated files (e.g., handwritten text in PDF format, sound files, and videos), as well as links to other relevant components, that are linked through their identifiers. Within the schema, text metadata thus functions as a central linking point: Each text is associated with learner metadata, situational and task-related metadata, and — where applicable — annotation and processing metadata. This design supports fine-grained descriptions at the text level while maintaining coherence across the corpus and enabling selective reuse of metadata components.

(5) Situational and task-related variables describe the communicative context in which learner language data were produced. This includes both general situational characteristics and, in instructed learning settings, details about the specific tasks used to elicit language samples. The variables proposed in this section are based on Biber and Conrad (2019, p. 40) and include, for example, the number of and relationship to addressees, the mode and medium of communication, the communicative purpose and the broad topic domain and register. By modelling situational and task-related information as a separate component, LC-meta allows multiple texts to be linked to the same contextual description where appropriate. This is particularly useful for avoiding the workload of entering metadata for corpora that include repeated tasks or standardised elicitation settings. The component thus supports both detailed contextualisation and efficient metadata management.

(6) Annotation metadata describes the linguistic or other annotations applied to the corpus data. Each annotation type is treated as a separate entity, allowing different annotation layers to be documented independently. This component records information such as the nature of the annotation (manual or automatic), the tools or methods used (including distinct version numbers), and the extent of quality control or evaluation. While it is not yet common practice in LCR to provide this information as machine-actionable metadata, the explicit modelling of annotation layers supports transparency, reproducibility, and reuse, and aligns with the increasing recognition of annotations as valuable research outputs in their own right.

(7, 8) Annotator and transcriber metadata describe the human agents involved in the transcription and annotation of learner language data. Although optional, these components enable provenance tracking and support documentation of annotation practices, which can be relevant for assessing data quality and reliability. Making annotators and transcribers explicit also supports attribution and acknowledgement of the work involved in preparing language resources for reuse.

3.2 Metadata Elements: Design Principles and Key Characteristics

While LC-meta is organised into components at a higher level, all individual metadata elements follow a common internal structure and set of organising principles. These principles were chosen to balance standardisation, usability, and flexibility, while remaining compatible with FAIR requirements and existing practices in corpus and language resource infrastructures.

Main element-structure and documentation: each metadata element is represented by a clearly defined field name, accompanied by a short, human-readable description. In addition, each element is documented by additional columns specifying conditionality, recommended fixed values, examples and further guidance, obligatoriness and repeatability for the element. Together, this makes explicit the information that can be recorded, and how and under which circumstances it should be provided.

Conditional metadata and guided data collection: some metadata elements are only relevant if another element has been answered in a particular way. Beyond its descriptive function, this design choice anticipates future technical implementations. In particular, conditionality could be used to configure adaptive metadata collection interfaces, such as web-based questionnaires, that dynamically guide respondents based on their previous answers. This reduces users' cognitive load and helps prevent inconsistent or irrelevant metadata entries.

Fixed attributes, controlled vocabularies and openness: wherever feasible, LC-meta recommends the use of fixed attributes, that is, closed lists of predefined values. These are intended to support Interoperability, comparability and machine-actionability. At the same time, the schema deliberately avoids enforcing controlled vocabularies when no widely accepted standard exists. For many community-relevant variables, consensus is still emerging. In such cases, metadata fields are left open-ended, while the schema provides recommended attributes and illustrative examples in a separate *further details and examples* column.

Use of examples and dual representations: the *further details and examples* column provides clarifications, usage notes, and concrete examples without making them normative. In some cases, LC-meta adopts a dual-representation strategy: a required field with a controlled vocabulary captures broad, comparable categories, while an optional open field allows users to provide more fine-grained or locally meaningful distinctions. For example, register is represented both by an obligatory field with a small set of broad categories and by an optional free-text field in which corpus compilers can specify more precise register types. This approach allows the schema to support both cross-corpus comparison and detailed documentation of individual resources.

Obligatory and optional metadata elements: similar to the main components, not all metadata elements are mandatory. Obligatory elements are intentionally limited in number and focus on information that is essential for basic interpretability, comparability, and reuse. Many other elements are optional and need only be recorded once at the corpus or design level, reducing the overall documentation burden. Inevitably, some (already existing) learner corpora might not be able to provide all obligatory elements and will have to fill in missing values. Nevertheless, we believe that establishing such a shared minimal core is particularly important for cumulative research and for integrating infrastructure. It provides guidance to newcomers, supports harmonisation across projects and increases the long-term value of corpora; even when primary data cannot be shared and only metadata can be made available. This combination of obligatory and optional fields also responds to desiderata identified by the community and the research literature (e.g. Tracy-Ventura et al. (2020); Wulff (2023)), thereby providing guidance on best practices for corpus compilation.

Repeatability and representation of complex realities: finally, metadata elements are characterised as non-repeatable or repeatable, allowing multiple values to be recorded where appropriate. This allows for capturing complex research designs while avoiding formal restrictions, yielding both flexibility and extensibility of the schema.

3.3 Technical Implementation

LC-meta was initially released as a structured spreadsheet with the explicit goal of later conversion into machine-actionable formats. For successful community uptake of the schema, the working group has identified the need for additional technical tools that enable user-friendly metadata administration, as well as tools for searching corpora based on metadata and survey-templates to collect metadata from

study participants. However, due to the workload and the financial limitations for outsourcing such a task, their technical implementation remains future work.

3.4 Community-Informed Development Process

In order to assure domain relevance and community uptake, the development of the LC-meta went through three main phases: In 2014, a comprehensive review of the metadata used in existing learner corpora[7] and other resources, such as the Talkbank website[8], the TEI and the Corpus Encoding Standard[9] resulted in a first draft proposal for core metadata for learner corpus research by Granger and Paquot[10]. The draft was then presented at the CLARIN *Workshop on Interoperability of Second Language Resources and Tools*[11] to collect community feedback and direct input. The draft proposal was then tested by a group of interested domain experts, including one of the initial authors and other researchers with experience in language research infrastructures and corpus creation from different institutions. The group evaluated the applicability of the initial draft to learner corpora with different corpus designs, developing a revised version which was again presented at two domain-specific conferences (Frey et al., 2023; König et al., 2022) for further community feedback. Additionally, feedback was gathered via mailing lists, online platforms and online questionnaires. After clarifying, evaluating and incorporating – where appropriate – all gathered feedback, LC-meta was published as a versioned, findable and openly accessible resource[12] with a detailed description in Paquot et al., 2024. After this, the schema was entrusted to the community by establishing a working group within the domain-specific Learner Corpus Association (LCA)[13], which collaborates closely with the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA)[14], and took over responsibility for further developing, maintaining and disseminating the metadata schema.

4 Case Study II - Metadata Schema for Interpreting Corpora within the Unified Interpreting Corpus (UNIC) Project

In the following, we present our second case study of a domain-specific metadata schema and show how metadata is described and standardised. We discuss design principles and underlying values, as well as the technical implementation of the metadata framework. This will provide valuable insights and feed into the development of a blueprint in Section 5 for community-informed, domain-specific metadata.

The metadata schema for interpreting corpora within the Unified Interpreting Corpus (UNIC) project (Liu & Russo, 2025) is based on the value-sensitive design paradigm (Friedman & Hendry, 2019), an approach in which conceptual design features and technical implementations embed social and moral values from the start, alongside usability and performance. While inspired by the component structure of LC-meta, the UNIC metadata schema consists of general and domain-specific elements within ten components to meet the needs of the translation and interpreting research community:

(1) Corpus metadata,
(2) Event metadata,
(3) Actor (as an umbrella term for speaker and signer) metadata,
(4) Source text metadata,
(5) Interpreter metadata,
(6) Target text metadata,
(7) Transcriber metadata,
(8) Annotator metadata,
(9) Annotation metadata,
(10) Alignment metadata.

[7] See the Learner Corpora around the World webpage: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[8] https://talkbank.org/

[9] https://www.cs.vassar.edu/CES/

[10] http://hdl.handle.net/20.500.12124/61

[11] https://sweclarin.se/swe/workshop-interoperability-l2-resources-and-tools

[12] https://doi.org/10.14428/DVN/AAUEM2

[13] https://www.learnercorpusassociation.org/working-group-on-metadata-in-learner-corpus-research/

[14] https://uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora.html

4.1 Design Principles

Aside from FAIR principles, the UNIC schema is also explicitly aligned with CARE principles (collective benefit, authority to control, responsibility and ethics; Carroll et al., 2020) because interpreting corpora often involve sensitive communicative settings, such as healthcare, asylum procedures, or institutional decision-making and may include audiovisual recordings and personal data. The CARE Principles were originally a framework for Indigenous Data Governance, ensuring that data practices respect the rights, interests and well-being of Indigenous peoples, not just technical openness or efficiency. The Principles are often extended to apply more broadly to vulnerable groups. The ethical considerations led to design decisions that affect both metadata content and exposure, including the introduction of opt-out mechanisms, fine-grained access levels and the separation of publicly visible descriptive metadata from access-controlled data download. The UNIC schema remains closely aligned with CLARIN infrastructures, as the majority of metadata elements reuse concepts already defined within CLARIN registries, and the CMDI profile is available in the Component Registry.[15]

4.2 Technical Implementation

A further methodological distinction between LC-meta and UNIC is the technical implementation of the schema. UNIC implemented its schema directly within an open-source platform (see footnote 3), which supports metadata and data registration, validation, filtering and search. The platform also acts as a communication channel between corpus creators and (re)users. The authors conducted surveys with computational and corpus linguists as well as interpreting students, and user feedback indicates greater perceived usefulness of the UNIC filters than those on the VLO. The authors are also collecting information on platform registrations and data use to further refine UI features and schema usability. In this respect, UNIC operationalises the value-sensitive design principle, in which technical quality and usability are closely linked, and schema development should be informed not only by conceptual and empirical considerations but also by interaction with real users.

4.3 Schema Development Process

While addressing a different research domain, the UNIC schema adopts a similar community-informed, empirically grounded, and iterative approach as LC-meta and extends it to accommodate the specific characteristics of interpreting studies. Similar to the LC-meta project, the starting point for schema development was a systematic review of existing practice. In the case of UNIC, this review was conducted at a larger empirical scale and across a wider range of modalities. A total of 1,898 publications were screened, resulting in the identification of 125 interpreting corpora described in 382 publications, covering 38 spoken and eight signed languages.

This corpus survey in the form of an ‘overlap and gap’ analysis served two purposes. First, it allowed the identification of metadata elements that recur across interpreting corpora (‘overlaps’) and therefore constitute candidates for core metadata. Second, it made gaps visible between the information in the data and that captured in existing metadata records. Elements that occurred in more than half of the corpora were prioritised as mandatory, while less frequent elements were retained as optional. This distinction mirrors the LC-meta principle of defining a minimal obligatory core that ensures comparability, while preserving flexibility for domain-specific variation.

The UNIC metadata schema demonstrates that conceptual, technical and empirical investigations can be applied to a specific domain while remaining compatible with CLARIN’s overarching metadata ecosystem. As with LC-meta, the UNIC metadata also highlights several common challenges in developing and maintaining such infrastructure, challenges that are likely to recur in many similar community-driven initiatives, which we discuss in section 6.

[15] https://catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu%3Acr1%3Ap_1747312582431&registrySpace=public

5 A Blueprint for Community-Informed Domain-Specific Metadata Development

The development of the Core Metadata Schema for Learner Corpora (LC-meta) and the Unified Interpreting Corpus (UNIC) metadata schema represents systematic efforts within CLARIN-related communities to operationalise FAIR, and, in the case of UNIC, also CARE principles through a community-informed, empirically grounded methodology rather than via top-down standardisation. Taken together, both initiatives suggest a reusable four-step blueprint for other research communities.

First, schema development should begin with a systematic review of existing practice. Metadata variables should be extracted from a wide range of relevant corpora and related publications within the target domain, including corpus documentation, corpus reports and state-of-the-art syntheses of the field (e.g., review articles, methodological overviews or handbook chapters). Based on this review, an initial draft of the metadata schema can be developed.

Second, an iterative and participatory design process should be initiated. Draft versions of the schema should be shared with the research community through workshops, conference presentations, mailing lists, and publicly accessible online documents. Feedback should be systematically collected and regularly integrated into the schema development process, ensuring that concrete community needs drive revisions. This approach will help ensure that successive schema versions reflect actual usage rather than assumptions not grounded in empirical practice.

Third, the schema should explicitly distinguish between obligatory and optional metadata elements, as well as between closed and open vocabularies. These design principles reflect the view that standardisation should establish a minimal common ground necessary for comparability and reuse, while still leaving room for methodological diversity and innovation. Rather than enforcing fixed definitions for potentially contested concepts, the schema should prioritise flexibility and recommend the use of accompanying documentation to clarify and contextualise specific metadata values.

Fourth, the schema should be conceived as a modular and extensible framework. No single schema can capture the full diversity of a domain from the outset. A shared core should therefore be complemented by additional modules developed by sub-communities with specialised expertise. Crucially, the long-term maintenance and evolution of a domain-specific schema should be embedded in appropriate governance structures within the domain's research community and connected to relevant scholarly associations, research centres, and research infrastructures.

Provided that some structural challenges are addressed with CLARIN's support (see Section 6), this combination of empirical grounding, iterative community engagement, and modular extensibility provides a viable blueprint for developing additional domain-specific metadata schemata that are both widely adoptable and responsive to community needs. Together, LC-meta and UNIC illustrate a methodological path for specialised metadata development within CLARIN. LC-meta provides the conceptual and procedural foundation; UNIC builds on this foundation under different empirical, technical and ethical conditions. For CLARIN, this highlights a key role: not to impose uniform metadata semantics, but to support communities in applying, adapting and extending a shared methodological framework. At the same time, developing domain-specific metadata schemata is highly time-consuming and resource-intensive, with limited prospects for external funding, underscoring the critical importance of CLARIN's support in enabling and sustaining these efforts.

In the next section, we propose concrete roles for CLARIN to support communities in developing domain-specific metadata schemata through technical assistance, infrastructure, tooling, governance frameworks that promote long-term sustainability and, last but not least, financial support.

6 How to Support Community-Specific Metadata within a FAIR CLARIN Ecosystem

As mentioned throughout the previous sections, the projects discussed above provide strong examples of obstacles and challenges that are likely to recur in similar community-driven initiatives.

6.1 Fostering Awareness for and Supporting the Creation of Domain-Specific Metadata Schemata

Despite the existence of general metadata schemata such as TEI, Dublin Core, and CMDI, the granular needs of subfields like interpreting studies or learner corpus research require more fine-grained, domain-specific schemata. LC-meta and UNIC underscore this gap. TEI, Dublin Core and CMDI offer broad, flexible structures, but they lack dedicated components for interpreting-specific variables (e.g., transcript-recording alignment, interpreter professional status) or learner-specific information (e.g., age of onset, multilingual background, learning context). Domain-specific schemata like LC-meta and UNIC enable richer, more nuanced metadata tailored to the theoretical and empirical needs of their fields. As such, the schemata complement broader frameworks. They provide the precision and context required to move from general Interoperability to discipline-optimised Reusability and research impact.

At the same time, the two projects also show that ideas and techniques can be reused and adapted across communities, even when their individual requirements differ. This paper examined how FAIR metadata can be realised for specialised corpora whose requirements go beyond generic metadata standards alone. Drawing on the development of LC-meta and the UNIC metadata project, we have argued that domain-specific metadata schemata are not peripheral to FAIR data management, but constitute a necessary condition for achieving Reusability in practice. The central question, therefore, is not whether such schemata should be developed, but how their development could be embedded within a shared research infrastructure such as CLARIN.

Our analysis reinforces a key insight that emerges from both LC-meta and UNIC. Metadata standards are not imposed on a community, but emerge through iterative, collaborative practice. FAIRness, in this sense, is not achieved by enforcing uniformity but by fostering coordination between communities, infrastructures, and services. One of CLARIN's core strengths lies precisely in its capacity to raise awareness and support such coordination by providing shared technical foundations while respecting the epistemic and methodological diversity of its user communities.

6.2 Building and Maintaining Infrastructure

In the case of UNIC, significant funds could be invested in outsourcing the design and implementation of an open-source platform to an external contractor, as funding for the project was provided under the HORIZON-MSCA Action 2022. This approach enabled the fast development of a feature-rich system tailored to the needs of this specific community. At the same time, this has led to a dependence on a codebase whose complexity exceeds what can generally be maintained within an academic research environment. As initial funding dries up, the long-term maintenance of such platforms becomes increasingly challenging. Even when the software is released under an open-source licence, continuity cannot be taken for granted. Bringing new developers into the project requires significant onboarding effort, including familiarisation with the architectural decisions, technology stack, and the domain-specific assumptions embedded in the code. Within a university context, where technical staff turnover is common and long-term software engineering positions are rare, this effort competes with core research and teaching responsibilities. Moreover, the incentive structures within academia are not well-suited for sustained infrastructure maintenance. Contributions to codebases, refactoring, documentation, or technical debt reduction are rarely recognised as scholarly output, despite being essential for the viability of research infrastructures. Consequently, even motivated community members may find it difficult to justify the time investment required to assume responsibility for such a platform. These challenges are not the exception. They are symptomatic of a structural gap between short-term project funding and long-term infrastructural needs.

CLARIN's added value lies in its position as a coordinating infrastructure with a long-term perspective, cross-community reach, and experience in sustaining shared services beyond individual project lifecycles. In this respect, CLARIN could act as an infrastructural stabiliser by providing a minimal, yet reliable framework within which community-specific platform development can be anchored and sustained after the end of project funding. Further support could include long-term hosting arrangements, integration and alignment with persistent identifier mechanisms and support with CLARIN authentication

and authorisation mechanisms.

6.3 Standardisation, Exchange and Usability

Discussions within the CLARIN community make clear that CMDI currently serves as the only commonly agreed meta-standard. At the same time, CMDI was never intended to serve as a single, semantically exhaustive metadata schema. Instead, it provides a flexible framework for defining profiles, understood as community- or domain-specific schema instantiations. This design reflects a structural tension that is intrinsic to CLARIN's mission, namely the need to ensure Interoperability across heterogeneous resources while enabling communities to describe their data with sufficient semantic precision.

A recurring concern among community members is the usability of richly annotated data. Community-specific metadata, because of its complexity and domain orientation, does not easily integrate into generic discovery services such as the VLO without significant information loss. At the same time, reducing such metadata to a minimal common denominator undermines precisely those aspects that make specialised corpora interpretable and reusable within their research contexts. The challenge is therefore not only technical but conceptual: how can metadata representations of different granularities coexist without losing functionality or visibility?

Our metadata model offers one possible response to this challenge. By treating CMDI-based descriptive metadata as the authoritative, infrastructure-facing layer, identified through persistent identifiers and exposed via repositories, CLARIN can provide a stable reference point for all other metadata representations. Richer, community-maintained metadata can then be understood as supersets that extend this authoritative layer, while simplified representations, such as citation metadata, EOSC-compatible HTML summaries, or exports to national discovery services, function as subsets. The persistent identifier serves as the common anchor linking all representations and ensuring their coherence.

This means that while community platforms like UNIC may remain internally technologically diverse, CLARIN could support the definition of stable interfaces for metadata exchange, harvesting, and discovery. By ensuring that key outputs, such as CMDI-compatible metadata, search endpoints, or landing pages, remain Interoperable with CLARIN services, CLARIN can decouple long-term visibility and Reusability from the internal evolution of a platform's codebase. CLARIN B- and C-centres can act as custodians of authoritative CMDI records, ensuring stability and persistence while allowing semantic innovation at higher descriptive layers.

6.4 Financial Support

Developing and maintaining domain-specific metadata standards is a demanding and resource-intensive endeavour that requires sustained expertise at the intersection of research practice and technical infrastructures, as well as careful negotiation between community-specific needs and more general, cross-domain approaches. Complexity is further exacerbated by the fact that many linguistic research communities are inherently multilingual. Their data, research questions and user groups typically span multiple languages, institutions and countries. As a consequence, no single institution, national community, or language group can reasonably be expected to develop and maintain a domain-specific metadata standard on its own. At the same time, no single language or national community is the exclusive beneficiary of such a standard. The benefits are distributed across institutions and countries, while the costs of development, coordination and long-term maintenance are concentrated. Project-based funding schemes tend to support the development of new software or standards, but rarely provide adequate mechanisms for their long-term maintenance and coordination, as illustrated by the experience of projects such as UNIC.

The main tension arises from the structure of existing funding schemes. Local or national funding programmes, for which individual institutions could realistically apply, are typically ill-suited to support the development of domain-specific metadata schemata, and even less so trans-national coordination efforts whose benefits extend beyond a single country. Conversely, European-level funding instruments, while more appropriate in scope, are usually targeted at substantially larger initiatives and infrastructural developments. In this context, CLARIN ERIC could assume an intermediary role by aggregating funding at the infrastructure level and redistributing it in a targeted manner to support coordinated, community-driven metadata work across Knowledge Centres.

Many challenges do not stem from a lack of expertise, but from its fragmentation across countries, institutions, and projects. CLARIN is therefore well-positioned to step in as a broker of shared stewardship by providing targeted, sustained support to Knowledge Centres that coordinate domain-specific, cross-institutional metadata work. By financially enabling these Centres to facilitate community exchange, documentation, and alignment with infrastructural components, CLARIN can help address structural incentive gaps and ensure that domain-specific metadata standards remain both viable and Interoperable over time. In conclusion, supporting FAIR metadata for specialised corpora requires CLARIN to embrace its role as an infrastructure mediator rather than a semantic authority. By enabling layered metadata representations, supporting community-driven schema development, and providing stable points of reference through CMDI and persistent identifiers, CLARIN can help bridge the gap between general Interoperability and domain-optimised Interoperability and Reusability. This approach allows specialised corpora to become truly FAIR, now and in the future.

7 Conclusion

This paper has examined how FAIR metadata can be realised for specialised corpora whose descriptive requirements exceed what can reasonably be expressed through generic metadata standards alone. Drawing on the development of LC-meta and the subsequent extension of its methodology in the UNIC project, we have argued that domain-specific metadata schemata are not peripheral to FAIR data management, but constitute a necessary condition.

The two case studies show that such schemata are most effective when they are developed through community-informed, empirically grounded, and iterative processes, and when their long-term maintenance is embedded in governance structures that connect research communities with infrastructures. At the same time, they also make visible a number of recurring structural challenges, most notably the tension between generic federation and semantic precision, the gap between short-term project funding and long-term infrastructural needs, and the fragmentation of expertise across institutions and countries.

Against this background, CLARIN's role should not be understood as that of an authority imposing uniform metadata standards across all domains. Rather, CLARIN is best positioned to act as an infrastructure mediator that enables the articulation, coordination, and long-term sustainability of community-specific metadata work. By supporting layered metadata representations, stable interfaces, community-driven governance, and targeted coordination through Knowledge Centres, CLARIN can help bridge the gap between general Interoperability and domain-optimised Reusability.

In this sense, the contributions of LC-meta and UNIC reach beyond the two communities examined here. Together, they point towards a viable methodological and infrastructural model for future domain-specific metadata development within CLARIN and illustrate why such work deserves more systematic institutional, technical, and financial support.

References

Bhardwaj, R. K., Nazim, M., & Verma, M. K. (2025). Research data management and FAIR compliance through popular research data repositories: An exploratory study. *Data Technologies and Applications*, *59*(3), 472–492. https://doi.org/10.1108/DTA-12-2022-0477

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Broeder, D., Windhouwer, M., Van Uytvanck, D., Trippel, T., & Goosen, T. (2012). CMDI: A Component Metadata Infrastructure. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini & T. Trippel (Eds.), *Proceedings of the Describing Language Resources with Metadata Workshop at LREC 2012*. European Language Resources Association (ELRA). Retrieved February 1, 2026, from http://lrec.elra.info/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. *Data Science Journal*, *19*(1), 1–12. https://doi.org/10.5334/dsj-2020-043

Erjavec, T., & Pančur, A. (2021). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.4133

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Retrieved February 1, 2026, from https://aclanthology.org/L18-1210/

Frey, J.-C., König, A., Stemle, E., Falaise, A., Fišer, D., & Lüngen, H. (2020). The FAIR Index of CMC Corpora. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the prism of digital humanities*. L'Harmattan. https://hdl.handle.net/10863/15984

Frey, J.-C., König, A., Stemle, E. W., & Paquot, M. (2023). A core metadata schema for L2 data. *Book of Abstracts of EUROSLA32*. https://hdl.handle.net/10863/42215

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Gilliland, A. J. (2008). Setting the Stage. In *Introduction to Metadata* (Second Edition, pp. 1–19). Getty Research Institute. Retrieved February 1, 2026, from https://www.getty.edu/publications/resources/virtuallibrary/0892368969.pdf

Greenberg, J. (2005). Understanding Metadata and Metadata Schemes. *Cataloging & Classification Quarterly*, *40*(3–4), 17–36. https://doi.org/10.1300/J104v40n03_02

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 1525–1531. Retrieved February 1, 2026, from http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf

König, A., Frey, J.-C., & Stemle, E. W. (2021). Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. *Information*, *12*(5), 199. https://doi.org/10.3390/info12050199

König, A., Frey, J.-C., Stemle, E. W., Glaznieks, A., & Paquot, M. (2022). Towards standardizing LCR metadata. *Book of Abstracts of the 6th Lerner Corpus Research Conference*. https://hdl.handle.net/10863/42994

Lindström Tiedemann, T., Lenardič, J., & Fišer, D. (2018). L2 learner corpus survey: Towards improved verifiability, reproducibility and inspiration in learner corpus research. In I. Skadin & M. Eskevich (Eds.), *CLARIN Annual Conference Proceedings, 2018* (pp. 146–150). CLARIN. Retrieved February 1, 2026, from https://www.clarin.eu/event/2018/clarin-annual-conference-2018-pisa-italy

Liu, N., & Russo, M. (2025). A value-sensitive metadata schema for interpreting corpora: Implementation on the Unified Interpreting Corpus (UNIC) platform. *Interpreting. International Journal of Research and Practice in Interpreting*, *27*(2), 157–196. https://doi.org/10.1075/intp.00123.liu
OA: https://hdl.handle.net/11585/1013497.

Lušicky, V., & Wissik, T. (2017). Discovering Resources in the VLO: A Pilot Study with Students of Translation Studies: CLARIN Annual Conference 2016 (L. Borin, Ed.). *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, 63–75. Retrieved February 1, 2026, from http://www.ep.liu.se/ecp/136/005/ecp17136005.pdf

Paquot, M. (2024). Learner corpus research: A critical appraisal and roadmap for contributing (more) to SLA research agendas. *Corpus Linguistics and Linguistic Theory*, *20*(3), 567–590. https://doi.org/10.1515/cllt-2024-0014

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, *10*(2), 280–300. https://doi.org/10.1075/ijlcr.24010.paq

Riley, J. (2024). *Seeing Standards: A Visualization of the Metadata Universe*. Platform for Experimental Collaborative Ethnography. Retrieved April 22, 2024, from https://worldpece.org/content/seeing-standards-visualization-metadata-universe-0

Tracy-Ventura, N., Paquot, M., & Myles, F. (2020). The Future of Corpora in SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (1st ed.). Routledge. https://doi.org/10.4324/9781351137904

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1029–1034). European Language Resources Association (ELRA). Retrieved April 2, 2026, from https://aclanthology.org/L12-1227/

Velterop, J., & Schultes, E. (2021). An Academic Publishers' GO FAIR Implementation Network (APIN) (A. De Kemp, Ed.). *Information Services & Use*, *40*(4), 333–341. https://doi.org/10.3233/ISU-200102

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018–160018. https://doi.org/10.1038/sdata.2016.18

Wulff, S. (2023). Corpus Research. In A. Chaouch-Orozco, E. Puig-Mayenco, J. Rothman, J. Cabrelli, J. González Alonso & S. M. Pereira Soares (Eds.), *The Cambridge Handbook of Third Language Acquisition* (pp. 683–695). Cambridge University Press. https://doi.org/10.1017/9781108957823.027