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
University of Bologna
Italy
nannan.liu@unibo.it

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

Magali Paquot
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 the community-informed creation of the Core Metadata Schema for Learner Corpora (LC-meta) initiative and the metadata schema developed for interpreting corpora within the Unified Interpreting Corpus (UNIC) project. We address how metadata are described and standardised, as well as the principles for schema design and the technical implementation of metadata frameworks. Our findings underline the importance of stakeholder involvement, iterative design, domain-oriented approaches to metadata standardisation, and a demand for technical support.

1 Introduction

The FAIR Principles (Wilkinson et al., 2016) have become a cornerstone of data management, promoting Open Science by making resources Findable, Accessible, Interoperable, and Reusable. Although existing repositories such as CLARIN B Centres (Hinrichs & Krauwer, 2014) make *F*indability and *A*ccessibility relatively easy to achieve, realising *I*nteroperability and *R*eusability remains a challenge (Frey et al., 2020): the generic metadata standards these repositories employ are meant for a broad range of content types and too coarse for domain-specific corpora. Individual communities require customised metadata schemata to reflect their unique characteristics and reuse contexts.

Even though "an 'overwhelming' number of metadata standards for the 'cultural heritage' field exist" (Riley, 2024) and also for CLARIN communities, e.g. Parla-CLARIN (Erjavec & Pančur, 2021), many research communities still lack *their* domain-specific metadata standard.

This study examines two community efforts to develop such standards: 1. the Core Metadata Schema for Learner Corpora (LC-meta) (Paquot et al., 2024), developed collaboratively by researchers and infrastructure experts in a community-informed iterative process; 2. the metadata schema for interpreting corpora (Liu & Russo, 2024) implemented on the UNIC platform[1], developed in the value-sensitive paradigm based on conceptual, empirical, and technical investigations.

In the following, we will address: How are domain-specific corpora currently described with metadata across communities? What principles and values underpin metadata schema development? These questions aim to bridge conceptual ideals and practical implementations, and through examination of two real-world schema development projects, we highlight both shared challenges and divergent solutions emerging from domain-specific needs and infrastructure contexts.

[1] https://unic.dipintra.it/ - Available under the Apache 2.0 License at https://github.com/F-technology-srl/unic

2 Metadata Schema – but Domain-Specific

Metadata was often defined in very general terms as "data about data" (Greenberg, 2005) but has evolved to encompass complex systems of description that encode the provenance, structure, semantics, and accessibility of digital objects (Gilliland, 2008). Today, metadata contains information on the definitions of the data, rules that govern the data's inner logic and contextual information. It may also include information on the process used when creating derived data to better understand its meaning and significance.

The development of metadata schemata has been central to data stewardship initiatives in digital humanities and linguistics. Text Encoding Initiative (TEI)[2], Dublin Core (DC)[3], and Component MetaData Infrastructure (CMDI)[4] provide formal structures, yet often require—and allow for—domain-specific adaptation and schema customisation to meet disciplinary relevance and usability while preserving interoperability.

Despite the theoretical advancement of metadata frameworks, practical challenges persist, for example, metadata quality is often uneven across corpora and resource types (Fišer et al., 2018); and specific user groups are dissatisfied with searchability and metadata transparency in CLARIN's VLO interface (Lušicky & Wissik, 2017). Such findings underscore the gap between schema availability and usability.

On the other hand, the effectiveness of metadata schemata depends on the extent to which they enable accurate, context-sensitive resource discovery and reuse (Greenberg, 2005). To this end, "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" (Velterop & Schultes, 2021). A domain-specific metadata schema can then capture and represent the unique characteristics, requirements, and interpretation of the data and resources within a particular domain.

3 Case Studies

3.1 LC-meta

LC-meta was developed through a collaborative, community-informed iterative process. The development process consisted of three phases:

Granger and Paquot conducted a comprehensive review of existing learner corpora and consulted resources such as the Learner Corpora around the World webpage[5], the Talkbank website[6], the TEI and the Corpus Encoding Standard[7] to identify common metadata variables and best practices in related fields like first language acquisition and corpus linguistics and proposed an initial draft of the core metadata schema. Their draft[8] was presented at the CLARIN *Workshop on Interoperability of Second Language Resources and Tools*[9] for feedback and direct input.

The authors, a group of people specialising in learner corpus research and/or research infrastructures from three different institutions, gathered to test the initial draft proposal, trying to apply it to several learner corpora from different institutions and with different corpus designs, which led to changes that were presented at two domain-specific conferences to collect feedback from the wider community. Further feedback was gathered via mailing lists and online platforms.

After all feedback had been clarified, evaluated and, where appropriate, incorporated, LC-meta was released[10]. Subsequently, a working group was established within the relevant domain-specific Learner Corpus Association (LCA)[11] to further develop, maintain and promote the metadata schema. For this

[2] https://tei-c.org/guidelines/

[3] https://www.dublincore.org/specifications/dublin-core/dces/

[4] https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

[5] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[6] https://talkbank.org/

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

[8] https://hdl.handle.net/20.500.12124/61

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

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

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

working group to be maximally representative of the research community, it will be further developed collaboratively with both the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA)[12] and the LCA.

3.2 UNIC

UNIC's approach to metadata standardisation built on LC-meta's ideas for structuring the metadata design and refinement phases, but also integrated conceptual, empirical, and technical dimensions tailored to its specific needs: The schema was grounded in both the FAIR and CARE[13] principles, emphasising data sovereignty, ethical access, and collective benefit. This was especially important in sensitive contexts such as healthcare and asylum interpreting.

The authors collected metadata from 125 spoken and signed language interpreting corpora. Empirically, they conducted 'overlap and gap analyses' to identify essential schema elements, and technically, automatic FAIRness evaluation revealed weaknesses in existing infrastructures. Further, they gathered feedback from students and researchers through surveys and software demos on the perceived usefulness of filters, which correspond to the compulsory metadata elements.

4 Findings

Despite the existence of general metadata schemata like TEI, Dublin Core and CMDI, the granular needs of subfields like interpreting studies or learner corpus research require more fine-grained domain-specific schema. The LC-meta and UNIC schemata underscore this gap: TEI, DC and CMDI offer broad, flexible structures, but 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).

LC-meta includes roughly 170 elements, which may be obligatory ($\sim$85) or optional ($\sim$85), and may be repeatable or singular. These elements are organised into eight interrelated main components: 1. Administrative metadata; 2. Corpus design metadata; 3. Learner metadata; 4. Text metadata; 5. Situational and task-related metadata; 6. Annotation metadata; 7. Annotator metadata; 8. Transcriber metadata;

The UNIC metadata includes 22 obligatory and 73 optional elements in ten components: 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.

The schema was implemented on the UNIC online platform (see Footnote 1), which supports metadata registration, filtering, search, and communication between corpus creators and (re)users. User feedback indicates greater perceived usefulness of the UNIC filters than those on the VLO. The authors collect data use plans when users register on UNIC and further feedback to refine UI features and metadata categories.

Domain-specific schemata like LC-meta and UNIC enable richer, more nuanced metadata tailored to the theoretical and empirical needs of their fields. The schemata are complementary–not redundant–to broader frameworks. They provide the precision and context required to move from general interoperability to discipline-optimised reusability and research impact. On the other hand, it is still possible to reuse and adapt ideas and techniques from another community despite individual requirements.

5 Conclusion

Although general-purpose metadata schemata and infrastructures like TEI, DC and CMDI offer a solid foundation, they often fall short when addressing the specific needs of specialised corpora. Our case studies show that domain-specific metadata schemata–such as LC–meta and the UNIC schema—are essential for ensuring accurate description, discovery, and reuse within particular research contexts.

Rather than replacing existing standards, these schemata extend them, enabling more precise, community-relevant metadata. To support this, CLARIN should take a more active role—similar to its

[12] https://uclouvain.be/en/research-institutes/ilc/clarin-knowledge-centre-for-learner-corpora.html
[13] https://www.gida-global.org/care

approach with the Resource Families—by facilitating community-led schema development and offering the technical infrastructure needed to integrate these efforts with platforms like the VLO.

By doing so, CLARIN can help bridge the gap between general interoperability and domain-optimised reusability, advancing FAIR data practices across diverse linguistic research communities.

References

Erjavec, T., & Pančur, A. (2021). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*, (Issue 14). https://doi.org/10.4000/jtei.4133

Fišer, D., Lenardič, J., & Erjavec, T. (2018, May). 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 April 9, 2025, from https://aclanthology.org/L18-1210/

Frey, J.-C., König, A., Stemle, E., Falaise, A., Fišer, D., & Lüngen, H. (2020, May). The FAIR Index of CMC Corpora. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the prism of digital humanities*. L'Harmattan. https://api.zotero.org/users/332053/publications/items/R7Z5VHA2/file/view

Gilliland, A. J. (2008, September). Setting the Stage. In *Introduction to Metadata* (Second Edition, pp. 1–19). Getty Research Institute. 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 August 20, 2019, from http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf

Liu, N., & Russo, M. (2024, October). A core metadata schema for interpreting corpora - Implementation on the Unified Interpreting Corpus (UNIC) platform. https://www.clarin.eu/sites/default/files/02_Liu-Russo_CLARIN2024_Metadata.pdf

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.) [Place: Linköping Publisher: Linköping University Electronic Press]. *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, 63–75.

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 [tex.ids= paquot-EtAl:2024:CoreMetadataSchemac publisher: John Benjamins]. *International Journal of Learner Corpus Research*, *10*(2), 280–300. https://doi.org/10.1075/ijlcr.24010.paq

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

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, I. 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