CLARIN Annual Conference Proceedings

2025

Edited by

Cristina Grisot and Thalassia Kontino

30 September – 2 October 2025
Vienna, Austria

Please cite as:
CLARIN Annual Conference Proceedings, 2025. ISSN 2773-2177 (online).
Eds. Cristina Crisot and Thalassia Kontino.
Vienna, Austria, 2025.

Programme Committee

Chair:

- Cristina Grisot, University of Zurich and at the Swiss National Center for Data Services for the Humanities DaSCH – chair of the leading sub-committee (CH)

PC Subcommittee:

- Costanza Navarretta, University of Copenhagen (DK)
- Vincent Vandeghinste, Dutch Language Institute, the Netherlands KU Leuven (BE)
- Joshua Wilbur, Univerity of Tartu (EE)
- Tanja Wissik, Austrian Academy of Sciences before Andreas Witt (AT)

Members:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies (IS)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Gijsbert Rutten, Leiden University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- German Rigau, HiTZ, the Basque Center for Language Technology (ES)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Gunn Inger Lyse Samdal, University Library, University of Bergen (NO)
- Sara Stymne, Uppsala University (SE)
- Marko Tadič, University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Andreas Witt, University of Mannheim (DE)
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University (ZA)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Starkaður Barkarson, IS
- Tomaž Erjavec, SI
- Cristina Grisot, CH
- Eva Hajičová, CZ
- Krister Lindén, FI
- Monica Monachini, IT
- Tanja Wissik, AT
- Costanza Navarretta, DK
- Maciej Piasecki, PL
- Stelios Piperidis, GR
- Gijsbert Rutten, NL
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Vincent Vandeghinste, BE
- Tamás Váradi, HU
- Joshua Wilbur, EE
- Andreas Witt, DE
- Friedel Wolff, ZA
- Martin Wynne, UK

Subreviewers:

- Ilze Auzina, LV
- Megan Bushnell, UK
- Riccardo Del Gratta, IT
- Maria Gavriilidou, GR
- Marissa Griesel, ZA
- Penny Labropoulou, GR
- Rooweither Mabuya, ZA
- Daniele Melaccio, IT
- Respect Mlambo, ZA
- Muzi Matfunjwa, ZA
- Deon du Plessis, ZA
- Tõnis Nurk, EE
- Michael Rießler, EE
- Mmasibidi Setaka, ZA
- Benito Trollip, ZA

CLARIN 2025 submissions, review process and acceptance

- Call for abstracts: 18 December 2024 first call published on CLARIN website, disseminated, and submission system open
- Submission deadline: 11 April 2025
- In total 59 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 12 June 2025
- Notifications to authors: 16 June 2025
- 42 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/content/call-extended-abstracts-clarin-annual-conference-2025.

Preface

We present with great pleasure the proceedings of the CLARIN Annual Conference 2025. The conference is held in Vienna from September 30 to October 3, 2025.

CLARIN conferences have been organised since 2012 and are the place where the people from the various countries participating in the CLARIN pan-European research infrastructure meet with each other and with the users of the infrastructure. Together, we form the CLARIN community. This year, we are pleased to welcome 222 on-site participants and 115 online participants to the conference.

For this years' edition, 59 extended abstracts were submitted from a total of 22 different countries. After peer reviewing conducted by the members of the National Coordinators Forum, 42 papers were accepted (0.71 success rate): 16 oral presentations and 38 poster presentations. We have organised the 16 oral presentations (and the written proceedings) thematically into three categories: 'Resources and usage', 'CLARIN and beyond', and 'Design and construction of CLARIN infrastructure'.

We would like to acknowledge the work of the PC Committee, the National Coordinators and the other sub-reviewers in ensuring the quality of the papers.

A new addition to this year's conference is the Panel *Language Technology: AI, Industry, and Research Infrastructures in Collaboration*, which will address the challenges and opportunities of AI-powered language technologies in fostering collaboration between academia and industry.

We are also very pleased to have two keynote speakers. Andreas Baumann from the University of Vienna (Austria) will talk about *Language change as an epidemiological phenomenon: why interdisciplinary linguistic research inevitably needs open data*. Lonneke van der Plas from the Univesità della Svizzera italiana (Switzerland) will give a talk titled *Creativity and AI*. We hope that both these talks will provide us with inspiration and reflection in an era where language technology and artificial intelligence developments have become omnipresent.

We hope you enjoy your stay in Vienna, that you have an inspiring conference and enriching interactions with other members of the CLARIN community.

Cristina Grisot, CLARIN-CH
Programme Committee Chair
National Coordinators Forum

Thalassia Kontino
CLARIN ERIC

Table of Contents

Posters

Interfacing CLARIN with H2IOSC: Metadata Interoperability through Ontology-based Mediation

Daniele Melaccio
ILC-CNR
Pisa, Italy
name.surname@cnr.it

Federico Boschetti
ILC-CNR
Pisa, Italy
name.surname@cnr.it

Monica Monachini
ILC-CNR
Pisa, Italy
name.surname@cnr.it

Abstract

We present an ontology-based approach for integrating CLARIN language resources into the H2IOSC semantic framework, promoting interoperability across the Social Sciences and Humanities. Building on CMDI and the CLARIN Concept Registry, our work extends CMD2RDF beyond syntactic conversion to formal semantic integration. CMDI metadata is mapped to CIDOC CRM and SSHOCro, with extensions for CLARIN-specific needs. This ontology-first strategy keeps CMDI fully operational while ensuring theoretical consistency and long-term sustainability. The resulting semantic layer enhances discoverability, enables cross-disciplinary integration, and contributes to a scalable knowledge graph aligned with FAIR principles and the EOSC vision.

1 Introduction

H2IOSC, the Humanities and Cultural Heritage Open Science Cloud,[1] brings together the Italian nodes of four European research infrastructures—CLARIN-IT, DARIAH-IT, E-RIHS.it, and OPERAS-IT—into a collaborative cluster that offers an open, multidisciplinary environment for advanced research on complex digital data and cultural objects.

In this context, interoperability is a cornerstone: it enables collaboration, data sharing, and resource integration across disciplines and technologies, and aligns H2IOSC with the EOSC FAIR principles. H2IOSC adopts an ontology-based methodology for integrating each RI's domain, consistently with CLARIN's long-standing FAIR trajectory (de Jong et al., 2018).

Within CLARIN, language-related datasets and tools are described through CMDI (Windhouwer & Goosen, 2022), supported by the CLARIN Concept Registry (CCR; Schuurman and Windhouwer 2015) (Section 2). Earlier work such as CMD2RDF (Windhouwer et al., 2017) showed the feasibility of serializing CMDI into RDF, but structural conversion alone is insufficient for cross-infrastructure interoperability. In H2IOSC, CLARIN CMDI descriptions are therefore embedded in a formal ontology to ensure semantic stability across infrastructures (Section 3).

CLARIN-IT maps CMDI profiles to CIDOC CRM (Bekiari et al., 2024) and SSHOCro (Bekiari et al., 2022) (Section 4). The mapping is non-invasive—CMDI remains fully operational—while the ontology adds a semantic layer that enables broader discovery and integration. Section 4.1 outlines the workflow and Section 4.2 illustrates it with a case study.

2 CLARIN Domain and Focus

CLARIN is an extensive federation of repositories, service centres, and centres of expertise that facilitate sustained access to digital language resources and advanced tools for linguistic analysis. Certified CLARIN centres describe hosted resources using CMDI, a standardized yet flexible metadata framework, while the CLARIN Concept Registry (CCR) ensures semantic consistency through controlled

[1] Official website https://www.h2iosc.cnr.it/

vocabularies with persistent identifiers. Together, they provide the structural and semantic scaffolding that underpins the CLARIN ecosystem.

Researchers typically explore resources through the Virtual Language Observatory (VLO)[2], a unified discovery interface that aggregates metadata from all CLARIN centres. Complementing the VLO, the CLARIN Resource Families[3] organize datasets into intuitive, user-oriented classes such as Parallel Corpora, Treebanks, or Spoken Corpora, offering a thematic overview that supports both newcomers and experienced users. On the services side, the CLARIN Language Resource Switchboard (LRS)[4] catalogs available linguistic tools (Zinn, 2018) and enables users to connect datasets with appropriate web applications for further processing. Viewed together, these three catalogues—VLO, Resource Families, and LRS—offer complementary perspectives on the same ecosystem: discovery, classification, and deployment.

2.1 CLARIN Metadata Framework: CMDI and CCR

CMDI is a modular, XML-based metadata schema that enables centres to describe linguistic resources with both structural consistency and descriptive flexibility. Metadata profiles are built from reusable components, making it possible to represent resource-specific features—whether textual corpora, lexica, or annotation workflows—while maintaining overall coherence across the infrastructure. This modularity is a distinctive strength of CMDI: it allows local adaptation to specific needs while preserving interoperability across centres.

The CCR complements CMDI by providing a registry of authoritative, controlled vocabularies. Each concept in the CCR is assigned a persistent identifier, ensuring that metadata values are standardized and semantically unambiguous across CMDI profiles. This not only guarantees cross-centre interoperability but also significantly enhances resource discoverability.

Taken together, CMDI and CCR form the semantic backbone of CLARIN. CMDI provides the structural framework for resource description, while the CCR anchors metadata values in a shared semantic space. This combination ensures that CLARIN's resources are not only richly described and consistently interpretable but also well-prepared for ontology-based integration within initiatives like H2IOSC.

3 Interoperability through ontology-based mediation

H2IOSC integrates data, tools, and services from E-RIHS.it, DARIAH-IT, CLARIN-IT, and OPERAS-IT through a Common Semantic Framework (CSF), inspired by SSHOC (SSHOC Consortium, 2021a, 2021b, 2021c). The CSF adopts a layered design: a top-level conceptual backbone which, following the SSHOC example, is based on CIDOC CRM and the SSHOC reference ontology (SSHOCro); below this upper backbone, more specific domain ontologies enrich the framework, capturing the richness of disciplinary perspectives while preserving overall coherence. CIDOC CRM and SSHOCro serve as key reference models, providing widely recognised semantics for cross-domain interoperability.

This design secures interoperability not only at technical or syntactic levels but—crucially—at the semantic level. Ontology-based mediation harmonizes resource descriptions, supports extensibility, and provides a stable foundation for cross-disciplinary collaboration.

3.1 The H2IOSC Marketplace as catalogue and orchestrator

The CSF's principles are put to use in the H2IOSC Marketplace through a dedicated data model—implemented within H2IOSC to support its dual role as catalogue and orchestrator. Structured around attributes, properties, and controlled vocabularies, this model is explicitly aligned with the CSF and its reference ontologies (CIDOC CRM and SSHOCro). Thus, the ontology supplies the semantic backbone for consistency and traceability, while the Marketplace is supplied with a data model that acts as a practical layer for cataloguing and workflow execution.

The Marketplace enables discovery, classification, and interaction with resources across infrastructures, and orchestrates workflows that combine datasets and tools into coherent pipelines (Sichera et al.,

[2]Virtual Language Observatory accessible at https://vlo.clarin.eu/
[3]Overview available at https://www.clarin.eu/resource-families
[4]CLARIN Switchboard accessible at https://switchboard.clarin.eu/

2024, 2025). Descriptions authored once in the ontology can be projected into the Marketplace without semantic drift and reused across heterogeneous platforms. A representative workflow and proof-of-concept pipelines are currently being developed to illustrate this orchestration capability.

Grounding the Marketplace in the ontology-based CSF avoids lock-in to a specific technical format and strengthens conceptual stability, adaptability, and long-term sustainability.

4 Mapping to CIDOC CRM and SSHOCro

Among the various ontological frameworks available for semantic interoperability, CIDOC CRM and SSHOCro were selected to meet the requirements of H2IOSC, drawing on the experience gained within the SSHOC project. CIDOC CRM, recognized as an ISO standard (ISO 21127), is widely adopted in the cultural heritage and humanities sectors for its event-centric modeling of entities, activities, and contextual relationships. SSHOCro builds on this foundation with SSH-specific extensions, adding constructs for linguistic resources, annotation workflows, and the research data lifecycle. Together, they provide a layered framework that combines cross-domain interoperability with the descriptive precision needed in the SSH domain.

4.1 The Mapping Strategy

The integration followed a combined top-down and bottom-up approach to align the CLARIN Metadata Framework with CIDOC CRM and SSHOCro. In the top-down phase, high-level CMDI metadata elements—such as Resource Type, Authors, and Funding—were mapped to CIDOC CRM classes selected for their semantic proximity to CMDI descriptors. Categories were chosen based on their ubiquity across profiles and their relevance for cross-infrastructure alignment.

The entry point was established at the level of *E1 CRM Entity*, the most abstract superclass in CIDOC CRM, ensuring that all mapped resources inherited a coherent semantic foundation. Subsequent mappings linked CMDI descriptors to entities such as *E21 Person* (authors), *E73 Information Object* (resources), and event-oriented classes like *E65 Creation* and *E7 Activity*.

To capture the specificity of the CLARIN domain, the mapping was refined through SSHOCro, which extends CRM with constructs for contributors' roles, linguistic resource types, and annotation workflows. Here, the catalogues already familiar to CLARIN users provided the natural entry point for ontology extensions:

Datasets. Under *SHE1_Dataset*, we introduced top-level nodes for *Corpus* and *LexicalResource*, further enriched with subclasses inspired by the CLARIN Resource Families (e.g., *ReferenceCorpus*, *ParallelCorpus*, *Lexicon*). This anchors the Resource Families taxonomy in a formal ontology, enabling precise alignment across infrastructures.

Tools and Services. Under *SHE10_Tool*, we defined the superclass *LinguisticTool*, extended with categories inspired by the Language Resource Switchboard (LRS). Examples include *CorpusQueryTool*, *Part-of-SpeechTagger*, and *NamedEntityRecognitionTool*. This gives formal semantic representation to classifications already used in practice, ensuring that Switchboard categories can be embedded within the H2IOSC framework.

This dual extension shows how community-driven catalogues—datasets from the Resource Families and services from the Switchboard—can be systematically formalized and aligned with SSHOCro and CIDOC CRM. In this way, the ontology both preserves community practices and enables rigorous integration.

This hierarchical modeling increases the semantic granularity of the ontology while maintaining alignment with SSHOCro. It facilitates more precise classification, discovery, and integration of CLARIN resources. All mappings and extensions were formalized in RDF and OWL to ensure compatibility with Semantic Web technologies. Collaborative modeling was carried out in Protégé,[5] with OntoGraf used for visual inspection and validation.

In the CLARIN Switchboard, matching between datasets and tools is pragmatic: metadata and formats are compared (e.g., a `.txt` resource can be sent to a tool that declares `.txt` compatibility). Effective

[5] Protégé official website: https://protege.stanford.edu/

for operations, this remains a technical compatibility check. In the ontology, by contrast, relations are explicit semantic assertions—e.g., *UDPipe ANALYSES ParlaMint* or *ParlaMint IS PROCESSED WITH NameTag*—embedded in the CSF and extensible with activities, agents, and outputs. This provides stable, reusable semantics across infrastructures.

4.2 Mapping example of CLARIN resources and tools

Datasets: Musisque Deoque. A representative case study is the Musisque Deoque (MQDQ) corpus, a multilingual digital library of Latin poetry enriched with philological and linguistic metadata (Boschetti et al., 2021). MQDQ was chosen for its structural complexity and interdisciplinary scope, sitting at the intersection of philology, literary studies, and computational linguistics. Metadata elements from its CMDI profile—such as Resource Type, Authors, and Funding—were mapped to SSHOCro and CIDOC CRM classes. The corpus itself was modeled as a *LiteraryCorpus* under *SHE1_Dataset*, contributors were aligned with *E21 Person*, and funding information with entities such as *nationalFunds*. These mappings, expressed as RDF triples, support Linked Open Data integration and links to external authority files (e.g., ORCID), funding registries, and cultural heritage datasets. MQDQ thus becomes discoverable both as a CLARIN corpus and as part of a wider semantic network, while CMDI structures remain intact.

Services: UDPipe and NameTag. The same methodology was applied to services exposed in the Language Resource Switchboard. Two illustrative examples are the parsing pipeline *UDPipe* and the named entity recognizer *NameTag*. *UDPipe* was modeled as an individual instance of the *PartOfSpeechTagger* and *DependencyParser* classes under *LinguisticTool*, while *NameTag* was aligned with the *NamedEntityRecognitionTool* class. The mappings, expressed as RDF triples, describe the tools in the same semantic framework as the datasets they are meant to process.

Although focused here on MQDQ and two selected services, the approach is generalizable. The same modeling principles can be applied across CLARIN resources and tools, enabling their gradual incorporation into an interoperable SSH knowledge graph. In this sense, this mapping example highlights the added value of ontology-based mediation in H2IOSC: datasets and services remain CLARIN assets, yet simultaneously become nodes in a larger, semantically coherent ecosystem.

5 Future Directions

H2IOSC promotes interoperability as a cornerstone of open science, with particular emphasis on the humanities and cultural heritage domains. By integrating data, tools, and services from E-RIHS.it, DARIAH-IT, CLARIN-IT, and OPERAS-IT into a shared environment, H2IOSC enables a harmonized infrastructure where resources will be semantically aligned and will be delivered through a unified Marketplace platform. This paper contributes to that goal by defining a reference framework for integrating CLARIN language resources into a collaborative and accessible ecosystem. A key achievement is the alignment between the CLARIN Metadata Framework and the H2IOSC Common Semantic Framework, which paves the way for cross-domain interoperability and future innovation.

Future work may focus on extending the mapping strategy to a broader range of CMDI profiles, promoting best practices in ontology-based metadata modeling within CLARIN. Publishing enriched CMDI metadata as RDF—aligned with CIDOC CRM and SSHOCro—could improve discoverability and facilitate interconnection with other SSH infrastructures on the Semantic Web.

To ensure broad adoption of this ontology-based approach, dedicated tools—such as guided interfaces and semi-automated mapping services—could be essential to support consistency, scalability, and long-term sustainability not only to CLARIN's ecosystem, but to the broader research landscape across the SSH domain and the EOSC ecosystem. Because the ontology-first strategy ensures compatibility with both ontology-based systems and pragmatic data models, CLARIN enriched metadata can be harvested directly into the H2IOSC Marketplace and, in parallel, interoperate with external platforms such as the SSHOC Marketplace. This dual alignment reinforces the purpose of this broader H2IOSC ambition.

Acknowledgments

This work is supported by the H2IOSC Project – Humanities and Cultural Heritage Italian Open Science Cloud, funded by the European Union – NextGenerationEU – NRRP M4C2 – Project code IR0000029 – https://www.h2iosc.cnr.it/.

References

Bekiari, C., Kritsotaki, A., Tsouloucha, E., & Theodoridou, M. (2022). D4.20 SSHOCro (final version) [Publisher: Zenodo]. Retrieved April 11, 2025, from https://zenodo.org/records/6771757

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.-E., Stead, S., & Velios, A. (2024, February). *Cidoc conceptual reference model, version 7.1.3* [Produced by the CIDOC CRM Special Interest Group. Volume A: Definition of the CIDOC Conceptual Reference Model.]. Retrieved April 11, 2025, from https://cidoc-crm.org/Version/version-7.1.3

Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021, December 14). La galassia *Musisque Deoque* : storia e prospettive. In *Antichistica* (Chapter_6411, Vol. 32). Fondazione Università Ca' Foscari. https://doi.org/10.30687/978-88-6969-557-5/026

de Jong, F., Maegaard, B., Smedt, K. D., Fiˇser, D., & Uytvanck, D. V. (2018). CLARIN: Towards FAIR and responsible data science using language resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Schuurman, I., & Windhouwer, M. (2015). CLARIN concept registry: The new semantic registry.

Sichera, P., Marras, C., & Pasini, E. (2024, November). Orchestrazione api per workflow applicativi nell'ambito delle digital humanities. https://doi.org/10.5281/zenodo.14187534

Sichera, P., Monachini, M., Quochi, V., Giampietro, N., Fabiani, V., Ottaviani, R., & Luzietti, R. B. (2025). Synergies between CLARIN-IT and OPERAS-IT within H2IOSC: Monitoring communities and orchestrating digital services [To appear]. *Proceedings of the CLARIN Annual Conference 2025*.

SSHOC Consortium. (2021a, January). *D7.1 system specification of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.1). SSHOC. https://zenodo.org/records/4558302

SSHOC Consortium. (2021b, November). *D7.2 release of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.2). SSHOC. https://zenodo.org/records/5749465

SSHOC Consortium. (2021c, December). *D7.3 final release of the ssh open marketplace* (tech. rep.) (SSHOC Deliverable D7.3). SSHOC. https://zenodo.org/records/5871651

Windhouwer, M., & Goosen, T. (2022, October 10). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *CLARIN* (pp. 191–222). De Gruyter. https://doi.org/10.1515/9783110767377-008

Windhouwer, M., Indarto, E., & Broeder, D. (2017). Cmd2rdf: Building a bridge from clarin to linked open data. In J. Odijk & A. Van Hessen (Eds.), *Clarin in the low countries* (pp. 95–103). Ubiquity Press. https://doi.org/10.5334/bbi.8

Zinn, C. (2018). Squib: The language resource switchboard. *Computational Linguistics*, *44*(4), 631–639. https://doi.org/10.1162/coli_a_00329