Computational Linguistics in the Netherlands Journal 15 (2026) 255-268 Submitted 01/2026; Published 06/2026

The CLARIN Knowledge Infrastructure for Linguistic and Language Technology Research

Vincent Vandeghinste[*] VINCENT.VANDEGHINSTE@IVDNT.ORG
Bente Maegaard[*,**] BMAEGAARD@HUM.KU.DK
Vesna Lušicky[***] VESNA.LUSICKY@UNIVIE.AC.AT

[*] *CLARIN Knowledge Centre for Dutch -K-Dutch , Instituut voor de Nederlandse Taal, Leiden & Centrum voor Computerlinguïstiek, KU Leuven*

[**] *Dept. of Nordic Studies and Linguistics, University of Copenhagen*

[***] *CLARIN Knowledge Centre for Terminology Resources and Translation Corpora -TRTC , Centre for Translation Studies, University of Vienna*

Abstract

While CLARIN is widely recognized for its technical infrastructure of language resources, tools, and services, its Knowledge Infrastructure is a vital and complementary pillar that supports researchers, educators, and developers in effectively using, sharing, and extending language data and technologies.

The CLARIN Knowledge Infrastructure promotes expertise exchange and best practices, and provides training and support across the research community. We present its main components and their relevance for the computational linguistics and digital humanities communities, with a focus on the Dutch and Flemish research landscape. We contextualize this by describing the knowledge infrastructure components in similar research infrastructures for other scientific domains.

1. Introduction

This paper describes the knowledge infrastructural part of the CLARIN infrastructure. It is therefore not a typical research paper, but nevertheless we think it is relevant to the CLIN Journal readers to be made aware that there is such a thing as the *Common Language Research Infrastructure* (CLARIN) and as a part thereof the Knowledge Infrastructure, which is a network of Centres of Expertise in specific subdomains of the social sciences and humanities, which often relate to the computational linguistics domain.

The notion of a *Knowledge Infrastructure* complements the provision of data and tools: it encompasses the organisational, technical, and social frameworks that enable the creation, sharing, and reuse of knowledge. Within the field of linguistic and language technology research, CLARIN exemplifies such an infrastructure. It connects distributed repositories, services, and expertise into a coherent ecosystem that supports evidence-based research on language in all its modalities and contexts.

Section 2 briefly introduces the CLARIN infrastructure. Section 3 contains the core content of this paper and describes the Knowledge Infrastructure and its different components. Section 4 explicitly links the CLARIN Knowledge Infrastructure to the computational linguistics community in the Netherlands and Flanders. Section 5 describes knowledge infrastructures in related research infrastructures, and Section 6 concludes.

2. Common Language Resource Infrastructure (CLARIN)

The Common Language Resources and Technology Infrastructure (CLARIN) is a European research infrastructure designed to make digital language data and language technology services easily and

sustainably available to researchers, in particular within the humanities and social sciences. Through CLARIN, scholars can discover, access, and combine a vast range of language resources, ranging from written and spoken corpora to lexical databases and processing tools, without needing to manage technical or legal complexities themselves.

From a user's perspective, CLARIN functions as a single, federated environment that connects certified national and thematic centres hosting data, tools, and expertise. Access is provided through a single sign-on mechanism, enabling researchers to move seamlessly across repositories and services. Thanks to shared standards, persistent identifiers, and common protocols, users can combine data and chain tools across centres in a reproducible manner (Hinrichs and Krauwer 2014).

Interoperability is a central design principle that directly benefits end users: it ensures that resources and tools can be found, understood, and reused across domains and languages. CLARIN addresses interoperability on multiple levels, ranging from governance and workflow orchestration to data curation and cross-community collaboration, positioning CLARIN as an Open Science platform that bridges the humanities, social sciences, and language technology (de Jong et al. 2020).

CLARIN's discovery layer offers researchers intuitive entry points to its distributed holdings. The *Virtual Language Observatory* (VLO)[1] provides a faceted browser that aggregates metadata for hundreds of thousands of resources, allowing users to search across formats, languages, and resource types (Van Uytvanck et al. 2012). Complementary services enable the creation of *virtual collections*,[2] allowing communities to curate and share subsets of relevant materials across repositories (Eskevich et al. 2020). Together, these services make it possible for end users to locate, combine, and reuse resources without deep technical knowledge of the underlying infrastructure.

An important organisational and user-facing concept within CLARIN is that of *Language Resource Families*.[3] These families group together resources of similar type, such as *Parallel Corpora*, *Learner Corpora*, *Speech Corpora*, *Treebanks*, *Lexical Resources*, or *Sign Language Resources*, collected across CLARIN centres (Fišer et al. 2018). For end users, resource families provide a structured, user-friendly overview that facilitates discovery, comparison, and reuse. Each family highlights representative datasets and tools, documentation standards, and metadata profiles tailored to that resource type. This family-based organisation also underpins the VLO's faceted search, ensuring that users can explore heterogeneous data sources through a consistent and interoperable interface.

In sum, CLARIN's federated structure, interoperability framework, and user-oriented discovery tools make it a central access point for language data and technologies in Europe. For computational linguists, corpus builders, and tool developers, it offers both a sustainable home for resources and a reliable gateway to multilingual, multimodal, and interdisciplinary research materials.

3. Knowledge Infrastructure within CLARIN

While CLARIN's technical infrastructure ensures sustainable access to language data and tools, its *Knowledge Infrastructure* focuses on enabling users to make effective use of these assets. It comprises the people, institutions, materials, and communication channels that facilitate knowledge creation, sharing, and reuse across the CLARIN community. In practice, the Knowledge Infrastructure translates the principles of FAIR data and Open Science into everyday research support: it helps researchers, educators, and developers to find the right resources, apply best practices, and acquire the necessary skills to work with language data in an interoperable way (CLARIN-ERIC 2024).

The Knowledge Infrastructure connects experts and users through certified Knowledge Centres (K-Centres) (Section 3.1), promotes standardised methodologies through *Best Practice* papers (Section 3.2), and disseminates knowledge via communication initiatives such as *Tour de CLARIN* (Section 3.3) and the *CLARIN Cafés* (Section 3.4). Together, these components form an ecosystem that

1. `https://vlo.clarin.eu`
2. `https://www.clarin.eu/content/virtual-collections`
3. `https://www.clarin.eu/resource-families`

complements the technical backbone of CLARIN. They ensure that expertise, guidance, and learning opportunities are as accessible and reusable as the data and tools themselves (CLARIN-ERIC 2024).

3.1 Knowledge Centres (K-Centres)

Knowledge Centres (K-Centres) form the backbone of the CLARIN Knowledge Infrastructure (CLARIN-ERIC 2024). They are certified centres of expertise that provide user support, consultancy, and training on specific languages, tools, or thematic areas of linguistic research. K-Centres are certified by CLARIN for a renewable three-year term. Certification guarantees that centres maintain active engagement, well-documented expertise, and up-to-date user support services.

Each K-Centre serves as an access point for researchers and developers seeking guidance on the use, creation, or integration of language resources and technologies. In doing so, they complement CLARIN's technical infrastructure with the human expertise needed to ensure that data and services are used effectively and sustainably. Currently, as shown in Table 1, there are thirty eight K-Centres distributed across the world (mainly in Europe), and they can mainly be divided into two categories:

Topic-specific K-Centres on a wide spectrum of topics, from speech resources to language learning, treebanking, and metadata curation.

Language-specific K-Centres about a single language (e.g. K-Dutch, see Section 4 for more info), or about a group of languages (e.g. CLASSLA for South-Slavic languages). Some K-Centres only treat a single variant of a language, such as German-AT, the K-Centre for German as spoken in Austria.

The K-Centres collaborate through an annual K-Centre workshop, where they exchange experiences, identify emerging user needs, and coordinate training and outreach activities (Vaičenonienė et al. 2024).

As discussed by Vaičenonienė et al. (2024), the K-Centre network embodies the transnational and collaborative character of CLARIN's Knowledge Infrastructure. It ensures that expertise is not confined within national boundaries but shared across Europe, enabling a sustainable exchange of knowledge between technology providers, researchers, and educators.

3.2 Best Practice Initiative: The K-Centre Library

Researchers can access a collection of best practice papers on various topics in a dedicated folder in the CLARIN Zotero library.[4]

The Best Practice papers consist of a collection of guidelines, workflows, and case studies that support researchers in adopting efficient, replicable practices when working with language data (CLARIN-ERIC 2025). From 2026, this collection will be called the *K-Centre Library*.

The Best Practice Initiative was started in 2023 with collecting best practice papers from K-Centres, and was later expanded to best practice papers from other sources. Frontini and Maegaard (2023) describe a definition of *best practice*, the selection criteria and the workflow for new submissions to the collection.

For the selection of papers, a set of criteria is applied that focuses on the clarity of the problem addressed, the usefulness of the proposed solution, and the justification of the chosen approach. First, the class of problems or the specific problem addressed by the paper should be clearly identified. If a broader class of problems is targeted, concrete guidelines or principles should be provided that can be followed to address problems of this type in general. Alternatively, if a single problem is addressed, the steps required to solve that problem should be described in detail, and it should be indicated how the proposed solution can be transferred or adapted to other domains, languages, datasets, or similar instances of the same problem. Finally, an explanation should be provided as to why the proposed solution is to be preferred over alternative approaches (Frontini and Maegaard 2023).

4. `https://www.zotero.org/groups/562080/clarin/collections/DNV2LZAN`.

Short name	Expertise
Topic Speficic	
ACE	Atypical Communication Expertise
CKCMC	Computer-Mediated Communication and Social Media Corpora
CKL2CORPORA	Learner Corpora
CKLD	Linguistic diversity and language documentation
CLARIN-APPLIED	Applied Comparative Discourse Analysis
CLARIN-ELEXIS	Lexicography
CLARIN-Learn	Language Learning Analysis
CLARIN-MULTISENS	Multimodal and Sensor-based Data
CLARIN-SPEECH	Speech Analysis
DiaRes	Diachronic Language Resources
DiPText-KC	Digital and Public Textual Scholarship
IMPACT-CKC	Digitisation
LLMs4SSH	Large Language Models in SS&H
K-OAr	Oral Archives in Italy
NLP:EL	Natural Language Processing in Greece
PhA-OeAW	Phonogrammarchiv, Austrian Academy of Sciences
SIKT- K-Centre	Data Management
Treebanking	Treebanking
TRTC	Terminology Resources and Translation Corpora
Language -s Specific	
CLARIN-SMS	Swedish in a Multilingual Setting
CLASSLA	South Slavic languages
CORLI-K	Corpora, Languages and Interaction -French
CorpLingCz	Czech Corpus Linguistics
CROATINA	Croatian Language
DANSK	DANish helpdeSK
K-Dutch	Dutch
GermanAT	German in Austria
HerLan	Heritage Languages in Europe
K-Icelandic	Icelandic
DR-LIB	Digital Resources for the Languages in Ireland and Britain
PolLinguaTec	Polish Language Technology
PORTULAN	Science and Technology of the Portuguese Language
RACAI4Ro	Romanian
RoNLP	Romanian Natural Language Processing
SAFMORIL	Systems and Frameworks for Morphologically Rich Languages
Spanish- K-Centre	Spanish
SWELANG	Languages of Sweden
UkrNLP-Corpora	Ukrainian NLP and Corpora

Table 1: List of Knowledge Centres

Some of the most recent additions to the collection are:

- Gomes et al. (2024), which advocates for a Research-Infrastructure-as-a-Service (RIaaS) model, unleashing accessibility to language processing services in as many web-based interface modalities as the current stage of technological development permits to support;

- Lee et al. (2024), which promotes best practices for sharing recordings of speech disorders within clinical phonetics and speech and language pathology.

- Draxler et al. (2024), which addresses the transcription portal and the webservices associated with speech processing at BAS, speech solutions developed at LINDAT, how to do it yourself with Whisper, remaining challenges, and future developments.

3.3 Tour de CLARIN

Tour de CLARIN is a communication initiative that showcases national consortia, use cases of CLARIN resources and services, and experiences from researchers across disciplines. Tour de CLARIN aims to raise awareness of available knowledge and tools over different user communities.

The latest publication in this series is Pahor de Maiti Tekavčič et al. (2025), which can be downloaded from https://zenodo.org/records/17406535 and which features articles on five CLARIN national consortia (Austria, Iceland, Lithuania, Poland and Switzerland), each featuring a tool, resource, event and interview with a user of the infrastructure in that country. It also features interviews with users of three K-Centres (Dansk, PhonogrammArchiv, CLARIN-SMS) and one CLARIN Technical Centre (a B-Centre).

3.4 CLARIN Cafés

CLARIN Cafés are an informal and interactive online space for discussion, held several times a year. Each Café has one or more speakers and lasts between one and two hours. For a list of topics: https://www.clarin.eu/content/clarin-cafe.

As examples of what these Cafés can contain, we briefly describe the three Cafés from 2025. The last one was organised by the Knowledge Infrastructure Committee and had the title *SSH Research with CLARIN K-Centres: Expertise for Multimodal Data, Large Language Models and Discourse Analysis*.[5] It was aimed at researchers in social sciences and humanities and had a focus on concrete use cases with multimodal data, with large language models and with multilingual comparative discourse analysis, explaining the role of the CLARIN K-Centres in this research. It had talks about speech, writing and interaction and the use of multimodal methods, provided by the K-Centre CLARIN-MULTISENS; the use of AI as a co-researcher showing creative and analytical use cases of large language models in social sciences and humanities, provided by LLMs4SSH and about how to do comparative discourse studies using corpus linguistics, provided by CLARIN-Applied.

In another Café, Alice Dijkstra was invited, which many CLIN researchers may know as she was involved in many language technology projects and programmes from the funder's side. She titled her talk *How to Talk to Your Funder - A Funder's Perspective*[6] and provided several useful insights and tips for project proposal writing in our field.

5. https://www.clarin.eu/event/2025/clarin-cafe-ssh-research-clarin-k-centres-expertise-multimodal-data-large-language
6. https://www.clarin.eu/event/2025/clarin-cafe-how-talk-your-funder-funders-perspective

The third Café was a joint organisation of CLARIN and DARIAH[7] and discussed *Project Management in Digital Humanities*.[8] Speakers from academia, industry and research infrastructures discussed their experience and approaches to project management and how some of the principles and methods can be incorporated into Digital Humanities research and teaching to improve the collaboration and efficiency of Digital Humanities project teams.

Every year we expect to organise at least three Cafés, and topics may vary from use cases for domain-specific topics to more generic topics, such as Legal and Ethical issues with the EU AI act.

3.5 The Learning Hub

The CLARIN Learning Hub offers a wide range of open educational resources, and provides full online training modules for learning new skills, as well as materials that can support the design of new university courses, training sessions, and workshops. In addition, the hub includes best practices and guidelines that have been developed in educational projects such as UPSKILLS or created in collaboration with other research infrastructures.

The Learning Hub consists of five different segments:

1. The *Tutorials* section features tutorials created by the CLARIN central hub for training workshops and other events, such as Grisot et al. (2025), as well as materials developed within EU-funded projects, such as UPSKILLS (Gledić et al. 2023).[9]

2. The new *Learning Resources Catalogue*[10] presents a curated collection of open educational materials produced within the CLARIN community. It highlights resources created for university teaching, training sessions, and workshop events, making it easier for educators and trainers to discover relevant teaching materials.

3. Another element is the *Digital Humanities Course Registry* (Wissik et al. 2020), developed in collaboration with DARIAH-ERIC, the research infrastructure for digital humanities. This registry provides an overview of current university courses and summer schools in Europe and beyond that focus on digital humanities. It serves as a valuable source for identifying up-to-date educational opportunities in the field.

4. The hub also includes a section on *Guidelines and Best Practices*, aimed at trainers and educators who want to develop learning content based on established practices within the CLARIN community and the broader social sciences and humanities network. This collection supports the creation of high-quality, methodologically sound educational resources. It consists of the following guidelines: Using CLARIN in Teaching and Learning, Research-Based Teaching, Guidelines for Students' Projects and Research Reporting Formats, Integrating Industry-Based Research Projects into Teaching, Learning and Training Materials Development and Sharing, Finding Open Learning and Training Resources within CLARIN & SSH, and Best Practices in Vocabularies Collection and Publication.

5. Finally, the *Trainers' Network* brings together a community of trainers with specialised expertise who deliver workshops and webinars across numerous disciplines. Their activities cover areas such as -computational linguistics, digital humanities, and social sciences, ensuring that learners have access to high-level training from specialists across these fields.

7. DARIAH stands for Digital Research Infrastructure for the Arts and Humanities. It is a European research infrastructure consortium (ERIC) that supports digitally enabled research and teaching in the arts and humanities by developing and maintaining tools, services, standards, and networks for managing, sharing, and analysing research data.
8. https://www.clarin.eu/event/2025/clarin-cafe-project-management-dh
9. https://upskillsproject.eu/
10. https://www.clarin.eu/learning-resources-catalogue

3.6 Funding instruments

To enhance the integration of core research infrastructure solutions at the level of human resources and to strengthen capacity in technical development, training, and adoption, CLARIN provides Mobility Grants. These grants support individual researchers, developers, and educators by funding short-term visits -e.g. one week between representatives of CLARIN centres in different countries, facilitating collaboration on the development, deployment, and use of the CLARIN infrastructure. In addition, Mobility Grants may support short-term exchanges between CLARIN and other research infrastructures.

CLARIN also provides support for workshops and collaborative initiatives, such as flagship projects. The new flagship project *PressMint: Interoperable Corpora of Historical Newspapers* has recently started.[11] This project aims to compile a multilingual, comparable, annotated, translated and interoperable set of corpora of European historical newspapers from around the start of the 20th century. The project follows a similar approach as the previous flagship project ParlaMint (Erjavec et al. 2023), in which parliamentary data from across Europe were all processed in a consistent manner. The most recent version of this dataset is available from Erjavec et al. (2025).

3.7 CLARIN Annual Conference

The CLARIN Annual Conference constitutes the primary forum for scholars and practitioners involved in the development and maintenance of the CLARIN infrastructure across Europe. It is designed to serve the broader humanities and social sciences communities by facilitating the exchange of ideas, best practices, and user experiences related to CLARIN. Discussions typically encompass the conceptualisation, construction, and operation of the infrastructure; the data, tools, and services it currently provides or requires; its uptake and use by researchers and educators; its connections to other infrastructures and collaborative projects; and the role of the CLARIN Knowledge Infrastructure.

The conference convenes authors of accepted extensive abstracts, members of national CLARIN consortia, representatives of CLARIN centres and partner organisations, as well as a wide range of participants interested in contributing to or becoming part of the CLARIN community.

The abstracts, edited by Grisot and Kontino (2025), are published as open access conference proceedings and can be downloaded from the CLARIN website. After the conference selected long papers are published in open access. Vandeghinste and Kontino (2025) edits the selected papers from the 2024 conference.

4. Relevance for the CLIN community

For the CLIN community, the CLARIN Knowledge Infrastructure is valuable because it consolidates expertise, documentation, training, and methodological guidance into a coherent environment that supports high-quality linguistic research.

Instead of merely providing tools or data, the knowledge infrastructure helps researchers understand how to use these resources effectively. For example, CLIN researchers working on corpus annotation, speech alignment, or NLP pipeline design can draw on CLARIN's expert forums, helpdesks, documentation portals, and training modules to resolve methodological questions, compare annotation standards, or identify best practices in metadata creation and resource sharing. This access to structured, community-maintained knowledge can greatly reduce the learning curve for both early-career researchers and experienced specialists exploring new technologies.

11. https://www.clarin.eu/pressmint

K-Dutch: Knowledge Centre for Dutch

An illustrative example for the CLINJournal audience is *K-Dutch*, the Knowledge Centre for Dutch, hosted by the Instituut voor de Nederlandse Taal. K-Dutch acts as a hub for expertise on Dutch language resources, tools, and best practices, offering access to corpora, lexica, NLP tools, and community support for researchers working with Dutch.[12] It also provides a service desk[13] to which researchers can direct their questions and can expect a human response within two working days.

In this paragraph we provide an anecdotal use case which contributed to linguistic research. A researcher wanted to know whether it was possible to get a distribution of adjective suffixes from a lexicon to get information on the productivity of these suffixes. The K-Centre pointed them to the e-Lex lexicon (*e-Lex (Version 1.1.1) [Data set]* 2014), which as its third data field provides the morphology of lemmas. As a service, the K-Centre created scripts that counted, per adjective lemma id the frequency of the last suffix. The most frequent suffixes are shown in Table 2, together with their frequencies. For cases with no morphology, the category '0' was assigned.

The K-Centre then provided the researcher with a spreadsheet that contained all the suffixed adjectives, organised per suffix, so that manual inspection and correction became possible. This information led to the publication of the chapter *Wat riekt het hier zwallig!*[14] (Cornips et al. 2023), a publication about the vocabulary related to smells.

Suffix	Frequency
0	11781
-ig	1781
-achtig	507
-baar	473
-isch	431
-elijk	392
-end	367
-en	292
-s	278
-lijk	237
-erig	229
-ief	168
-aal	155
-loos	138
-d	116
...	

Table 2: Distribution of adjective suffixes in the eLex Dutch lexicon

Without knowledge about the existence of e-Lex and the type of information it contains, and without scripting skills it would be very hard (or painstakingly much work) to obtain this type of information, demonstrating K-Dutch, and more general the K-Centres and the CLARIN Knowledge Infrastructure as an enabler or facilitator of certain types of research.

The knowledge infrastructure also serves as a dissemination channel for methodological innovation and collaboration. Through curated guidelines, case studies, and training materials based on real projects, it enables researchers to document their workflows, reflect on their approaches, and make their insights reusable by others. When CLIN researchers develop a new annotation manual for Dutch coreference, set up a reproducible training environment for an MT experiment, or evaluate

12. `https://kdutch.ivdnt.org`
13. servicedesk@ivdnt.org
14. A Dutch dialect expression translated by Google Translate as *How foul it smells in here!*

ASR performance on dialectal speech, the knowledge infrastructure provides a place to publish not just the outputs but also the know-how. This strengthens methodological transparency and fosters reuse and cross-project learning.

Moreover, the infrastructure actively links computational linguistics with relevant knowledge from adjacent fields, especially the digital humanities and social sciences. By integrating educational materials from projects such as UPSKILLS, aligning with standards initiatives, and collaborating with other research infrastructures, it enables the CLIN journal audience to situate their methodologies within broader scholarly practices. This is particularly helpful when working on interdisciplinary topics, for instance, when a computational linguist needs guidelines for ethically collecting user-generated content, or when developing annotation schemes that must remain compatible with humanities-oriented standards such as Text Encoding Initiative -TEI or Component Metadata Infrastructure -CMDI .

Ultimately, the strength of the CLARIN Knowledge Infrastructure lies in how it combines technical support with a living, evolving knowledge base shaped by its research community. It helps CLIN researchers access language resources and technologies, and equips them with the expertise, methodological clarity, and shared understanding required to use those resources correctly and effectively. In doing so, it enhances the quality, impact, and longevity of computational linguistic research across the Netherlands, Flanders and beyond.

5. Related Work: Knowledge Infrastructure in Related Infrastructures

Beyond CLARIN, several European research infrastructures have developed mature knowledge-infrastructure components that support researcher skills, methodological guidance, discovery, and reuse across domains. While these infrastructures have developed independently of each other, there is cooperation at different levels, such as the ERIC forum and ESFRI -European Strategy Forum on Research Infrastructures .

DARIAH serves the digital humanities and arts community, supporting scholars who work with cultural, historical, linguistic, and artistic data. Its knowledge layer centres on DARIAH-Campus, a discovery and hosting platform for open, reusable DH training materials developed during the H2020 DESIR project (DARIAH-ERIC 2017). By curating lessons, tutorials, and course packs and embedding them in typical DH workflows, it lowers barriers for lecturers and researchers to adopt best practices and assemble tailored training. For end users, it provides a single, quality-assured entry point to DH pedagogy aligned with the services of the infrastructure (DARIAH-ERIC 2019).

CESSDA serves the social science community, particularly users of empirical, survey-based, and longitudinal data. It maintains a widely used Data Management Expert Guide and complementary Data Citation Guide, which translate FAIR, PID, and archiving principles into concrete guidance for researchers, repositories, and journals (CESSDA-ERIC 2021, CESSDA-ERIC 2020). In practice, these resources function as policy-aware handbooks that can be directly applied in proposals and research workflows.

SSHOC built on the collaboration of DARIAH, CLARIN, and CESSDA to develop components that integrate the broader SSH ecosystem. Its SSH Open Marketplace is a curated catalogue of tools, services, datasets, workflows, and training materials (SSHOC-Consortium 2022). It offers researchers a contextualised discovery environment in which software, data, methods, and documentation appear together within typical SSH research scenarios, exemplifying cross-infrastructure integration. The SSH Open Marketplace aggregates resources from infrastructures such as CLARIN, DARIAH, and CESSDA, but this is done through specific ingestion workflows rather than a blanket automatic harvest (Barbot et al. 2022).

OPERAS supports the scholarly communication landscape in the social sciences and humanities, with a strong focus on open-access publishing, multilingualism, and community-led open science. Instead of a centralised knowledge hub, OPERAS organises its methodological and training activities through a network of Special Interest Groups -SIGs . These SIGs—covering areas such as metrics,

multilingualism, innovation, scholarly publishing workflows, and open peer review—bring together experts to co-develop guidelines, best practices, and training resources (OPERAS 2024). In this way, OPERAS provides a distributed but structured knowledge-infrastructure layer that supports researchers, publishers, and service providers across the SSH publishing ecosystem.

E-RIHS supports the heritage science community, spanning conservation science, archaeological science, and technical art history. While it does not maintain a centralised knowledge hub, it offers a distributed set of methodological and training resources through its access services and project activities. Much of this material is produced within IPERION HS, which provides documentation on analytical methods, laboratory and in situ workflows, and training schools and workshops for heritage scientists (IPERION HS Consortium 2023). Together, these materials constitute a dispersed but robust knowledge layer that supports methodological transparency, reproducible research, and capacity building in the heritage science domain.

ARIADNEplus supports archaeological research through a combined Training Hub and Knowledge Base that expose enriched, interoperable archaeological metadata via CIDOC-CRM and Linked Open Data (ARIADNEplus-Consortium 2020). Its integration of training with semantically harmonised data services provides domain specialists with a highly actionable, technically advanced knowledge environment.

ELIXIR, although situated in a distant scientific domain, provides a mature model for knowledge-infrastructure design in the life sciences. Its RDMkit offers continually updated FAIR-aligned guidance for data stewards and researchers (ELIXIR-Europe 2021), while TeSS -Training e-Support System aggregates training events and materials across nodes and providers (ELIXIR-Europe 2022). Together, these function as a life-cycle skills and methods playbook for both wet-lab and computational research.

EUDAT provides cross-disciplinary, domain-agnostic data services supporting FAIR data management across scientific fields. While it does not operate a centralised "knowledge hub," the infrastructure offers extensive user guidance through its Service Catalogue, which documents workflows, interfaces, and best practices for services such as B2SHARE, B2FIND, and B2SAFE (EUDAT Collaborative Data Infrastructure 2024). Complementary training materials, webinars, and hands-on guides are hosted across its support pages and project outputs, providing practical assistance for researchers and data stewards implementing interoperable and sustainable data management workflows. Together, this distributed body of documentation and training fulfils a knowledge-infrastructure role by enabling users to adopt EUDAT services effectively and in alignment with FAIR and EOSC practices.

OpenAIRE underpins the European Open Science Cloud with guidelines, training, and community support for open science and scholarly communication. Its documentation on repository interoperability, metadata standards, and open-access workflows helps institutions and researchers implement open-science practices in their daily work (OpenAIRE 2021).

At the policy level, ESFRI frames research infrastructures as cross-cutting "knowledge and innovation hubs," emphasising that training, methodological guidance, and capacity building are intrinsic missions rather than optional additions (ESFRI 2021). This perspective motivates the consolidation of discovery, guidance, and skills under a unified knowledge-infrastructure umbrella across scientific domains.

Across research communities, successful knowledge infrastructures combine -i curated guidance - best practices, policies, how-tos , -ii training registries and reusable materials, and -iii contextualised discovery portals. CLARIN's Knowledge Infrastructure follows this model—through K-Centres, best-practice papers, Tour de CLARIN, Cafés, and the Learning Hub—while remaining integrated with discovery services such as the VLO and resource families. The resulting environment enables faster onboarding, more reproducible workflows, and easier reuse across centres and countries.

6. Conclusions

This paper has outlined the structure and purpose of the CLARIN Knowledge Infrastructure and explained how it supports research in linguistics and language technology. CLARIN is often associated with the provision of corpora, tools, and services, but its Knowledge Infrastructure is equally significant. It provides an organised framework through which expertise, documentation, and training become accessible to researchers, teachers, and developers. The K-Centres, the best-practice initiative, the Tour de CLARIN series, the CLARIN Cafés, and the Learning Hub all contribute to this framework and help the community to work in consistent and well-documented ways.

For the CLIN community, the Knowledge Infrastructure is useful because it reduces the effort required to locate guidance, compare methods, and find specialised expertise. It offers practical support for setting up reproducible workflows, selecting appropriate standards, and becoming familiar with existing tools and resources. These features help both newcomers and experienced researchers who wish to extend their methodological repertoire or explore new types of data.

The comparison with other European research infrastructures shows that CLARIN is part of a broader landscape in which community support, training, and methodological clarity have become essential components. Infrastructures such as DARIAH, CESSDA, SSHOC, OPERAS, ARIADNE, ELIXIR, E-RIHS, OpenAIRE, and EUDAT address similar needs in their respective domains. Each of them provides examples of how guidance, documentation, and training can be organised in a sustainable way. CLARIN contributes to this ecosystem by focusing on language resources and language-related research practices.

Further development of the Knowledge Infrastructure will remain important. The rapid growth of multimodal datasets, new annotation practices, and large language models calls for continued attention to documentation, standards, and training. Collaboration with neighbouring infrastructures can help to maintain coherence across disciplines, especially in areas where research communities face similar challenges.

In conclusion, the CLARIN Knowledge Infrastructure provides the organisational and intellectual support that researchers need to work effectively with language data and tools. It brings together expertise that would otherwise remain dispersed across institutions and projects, and it offers a stable environment for the exchange of methods and experiences. This contributes to more transparent, coherent, and durable research practices within the CLIN community and beyond.

Acknowledgements

This paper was written by members of the CLARIN Knowledge Infrastructure Committee (KIC). Bente Maegaard is the chair and Vesna Lušicky is the vice-chair. She is also the coordinator of the CLARIN Knowledge Centre for Terminology Resources and Translation Corpora (TRTC).[15] Vincent Vandeghinste is a member of the KIC and of the CLARIN Board of Directors. He is the coordinator of the CLARIN Knowledge Centre for Dutch (K-Dutch),[16] and national coordinator for CLARIN-Belgium (CLARIN-BE).[17]

References

ARIADNEplus-Consortium (2020), ARIADNEplus Portal and Knowledge Base. https://portal.ariadne-infrastructure.eu.

15. `https://trtc.univie.ac.at/`
16. https://kdutch.ivdnt.org/
17. `https://clarin-be.ivdnt.org/`

Barbot, Laure, Edward Gray, Frank Fischer, Matej Ďurčo, Alexander König, Marie Puren, Stefan Buddenbohm, Cesare Concordia, and Klaus Illmayer (2022), The SSH Open Marketplace: a multi-voiced story. https://doi.org/10.5281/zenodo.6580303.

CESSDA-ERIC (2020), CESSDA Data Citation Guide. https://zenodo.org/record/4062386.

CESSDA-ERIC (2021), CESSDA Data Management Expert Guide. https://dmeguide.cessda.eu.

CLARIN-ERIC (2024), CLARIN Knowledge Infrastructure. https://www.clarin.eu/content/knowledge-infrastructure.

CLARIN-ERIC (2025), Collection of Best-Practice Papers in Zotero Library. https://www.clarin.eu/content/collection-best-practice-papers-clarin.

Cornips, Leonie, Jeroen van Craenenbroeck, Nicoline van der Sijs, and Jos Swanenberg (2023), Wat riekt het hier zwallig! Geurwoorden in het Nederlands en zijn dialecten, *in* Leemans, Inger and Caro Verbeek, editors, *NeusWijzer. Geuratlas van de Lage Landen*, Boom, Amsterdam, p. 163–174.

DARIAH-ERIC (2017), DESIR (DARIAH-ERIC Sustainability Refined) Project. https://cordis.europa.eu/project/id/731081.

DARIAH-ERIC (2019), DARIAH-Campus: Discovery Framework and Hosting Platform for Learning Resources. https://campus.dariah.eu.

de Jong, Franciska, Bente Maegaard, Darja Fišer, Dieter van Uytvanck, and Andreas Witt (2020), Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN, *in* Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 3406–3413. https://aclanthology.org/2020.lrec-1.417/.

Draxler, Christoph, Henk van den Heuvel, Arjan van Hessen, Pavel Ircing, and Jan Lehečka (2024), Speech technology services for oral history research, *in* Anuradha, Isuri, Martin Wynne, Francesca Frontini, and Alistair Plum, editors, *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, pp. 38–43. https://aclanthology.org/2024.htres-1.6/.

e-Lex (Version 1.1.1) [Data set] (2014). http://hdl.handle.net/10032/tm-a2-h2.

ELIXIR-Europe (2021), RDMkit - Research Data Management Toolkit for Life Sciences. https://rdmkit.elixir-europe.org.

ELIXIR-Europe (2022), TeSS - Training eSupport System. https://tess.elixir-europe.org.

Erjavec, Tomaž, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavrilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Giancarlo Luxardo, Carmen Magariños, Måns

Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Andriana Rii, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer (2025), Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/2004.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinthór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Marx Maarten, and Darja Fišer (2023), The parlamint corpora of parliamentary proceedings, *Language Resources and Evaluation* **57**, pp. 415–448. https://doi.org/10.1007/s10579-021-09574-0.

ESFRI (2021), ESFRI Roadmap 2021. https://roadmap2021.esfri.eu.

Eskevich, Maria, Franciska de Jong, Alexander König, Darja Fišer, Dieter Van Uytvanck, Tero Aalto, Lars Borin, Olga Gerassimenko, Jan Hajic, Henk van den Heuvel, Neeme Kahusk, Krista Liin, Martin Matthiesen, Stelios Piperidis, and Kadri Vider (2020), CLARIN: Distributed language resources and technology in a European infrastructure, *in* Rehm, Georg, Kalina Bontcheva, Khalid Choukri, Jan Hajič, Stelios Piperidis, and Andrejs Vasiļjevs, editors, *Proceedings of the 1st International Workshop on Language Technology Platforms*, European Language Resources Association, Marseille, France, pp. 28–34. https://aclanthology.org/2020.iwltp-1.5/.

EUDAT Collaborative Data Infrastructure (2024), Eudat service catalogue. https://eudat.eu/catalogue.

Fišer, Darja, Jakob Lenardič, and Tomaž Erjavec (2018), CLARIN's key resource families, *in* Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1210/.

Frontini, Franceska and Bente Maegaard (2023), Guidelines Best Practice Papers, *Technical report*, CLARIN ERIC. https://office.clarin.eu/v/CE-2023-2205-Guidelines-Best-Practice_v1-final.pdf.

Gledić, Jelena, Maja Dukanović, Jelena Budimirović, Nada Soldatić, Maja Miličević Petrović, Silvia Bernardini, Adriano Ferraresi, Novella Tedesco, Iulianna van der Lek, Darja Fišer, Genoveva Puskas, Margherita Pallottino, Marie Berthouzoz, Tihana Kraš, Martina Podboj, Marko Simonović, Tanja Samardžić, Lonneke van der Plas, Marc Tanti, Louis ten Bosch, Henk van den Heuvel, Stavros Assimakopoulos, Albert Gatt, and Michela Vella (2023), UPSKILLS teaching and learning content. http://hdl.handle.net/11356/1865.

Gomes, Luís., Antonio Branco, João Silva, and Ruben Branco (2024), From greatest simplicity to full power, *Language Resources & Evaluation*, Springer. https://doi.org/10.1007/s10579-024-09772-6.

Grisot, Cristina and Thalassia Kontino, editors (2025), *CLARIN Annual Conference Proceedings*, CLARIN ERIC, Vienna, Austria. Online edition. https://www.clarin.eu/sites/default/files/CLARIN2025_ConferenceProceedings.pdf.

Grisot, Cristina, Erik Körner, Thomas Eckart, Petya Osenova, Maciej Piasecki, Mietta Lennes, and Iulianna van der Lek (2025), CLARIN101 - Introduction to CLARIN. https://doi.org/10.5281/zenodo.17379981.

Hinrichs, Erhard and Steven Krauwer (2014), The CLARIN research infrastructure: Resources and tools for eHumanities scholars, *in* Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 1525–1531. https://aclanthology.org/L14-1356/.

IPERION HS Consortium (2023), IPERION HS Training and Capacity Building Activities. https://www.iperionhs.eu/training/.

Lee, Alice, Nicola Bessell, Henk van den Heuvel, Katarzyna Klessa, and Satu Saalasti (2024), The DELAD initiative for sharing language resources on speech disorders, *Language Resources & Evaluation*, Springer. https://doi.org/10.1007/s10579-023-09655-2.

OpenAIRE (2021), OpenAIRE Guidelines for Open Science and Repository Interoperability. https://guidelines.openaire.eu/.

OPERAS (2024), Operas special interest groups: Community-led development of standards, workflows and best practices. https://operas-eu.org/special-interest-groups/.

Pahor de Maiti Tekavčič, Kristina, Jakob Lenardič, and Karina Berger, editors (2025), *Tour de CLARIN*, Vol. 5, CLARIN. https://doi.org/10.5281/zenodo.17406535.

SSHOC-Consortium (2022), SSH Open Marketplace. https://marketplace.sshopencloud.eu.

Vaičenonienė, Jurgita, Michal Kren, Vesna Lušicky, and Vincent Vandeghinste (2024), Transnational Research Infrastructure: A Journey Through CLARIN Knowledge Centres, *Digital Humanities in the Nordic and Baltic Countries Publications*, Oslo, Norway. https://journals.uio.no/dhnbpub/article/view/11520.

Van Uytvanck, Dieter, Herman Stehouwer, and Lari Lampen (2012), Semantic metadata mapping in practice: the virtual language observatory, *in* Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 1029–1034. https://aclanthology.org/L12-1227/.

Vandeghinste, Vincent and Thalassia Kontino, editors (2025), *Selected Papers from the CLARIN Annual Conference 2024*, Vol. 216 of *Linköping Electronic Conference Proceedings*, Linköping University Electronic Press, Barcelona, Spain. https://doi.org/10.3384/ecp216.

Wissik, Tanja, Jennifer Edmond, Frank Fischer, Franciska de Jong, Stefania Scagliola, Andrea Scharnhorst, Hendrik Schmeer, Walter Scholger, and Leon Wessels (2020), Teaching Digital Humanities Around the World: An Infrastructural Approach to a Community-Driven DH Course Registry. https://hal.science/hal-02500871.