Originally published in: Fišer, Darja/Witt, Andreas (Eds.): CLARIN. The infrastructure for language resources. – Berlin/Boston: de Gruyter, 2022. Pp. 31-57. (Digital Linguistics 1)
DOI: *https://doi.org/10.1515/9783110767377*

Franciska de Jong, Dieter Van Uytvanck, Francesca Frontini, Antal van den Bosch, Darja Fišer, and Andreas Witt

Language Matters

The European Research Infrastructure CLARIN, Today and Tomorrow

Abstract: CLARIN stands for "Common Language Resources and Technology Infrastructure". In 2012 CLARIN ERIC was established as a legal entity with the mission to create and maintain a digital infrastructure to support the sharing, use, and sustainability of language data (in written, spoken, or multimodal form) available through repositories from all over Europe, in support of research in the humanities and social sciences and beyond. Since 2016 CLARIN has had the status of Landmark research infrastructure and currently it provides easy and sustainable access to digital language data and also offers advanced tools to discover, explore, exploit, annotate, analyse, or combine such datasets, wherever they are located. This is enabled through a networked federation of centres: language data repositories, service centres, and knowledge centres with single sign-on access for all members of the academic community in all participating countries. In addition, CLARIN offers open access facilities for other interested communities of use, both inside and outside of academia. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources can be chained to perform operations at different levels of complexity. The strategic agenda adopted by CLARIN and the activities

Acknowledgements: The authors of this chapter are grateful for the input provided by Karina Berger, John Picard, and Leon Wessels. Funding for some of the work underlying the content of this chapter was made available through the grants that CLARIN has received throughout the years. (See for an overview: https://www.clarin.eu/content/clarin-eu-projects.)

Franciska de Jong, CLARIN ERIC's Executive Director, e-mail: f.m.g.dejong@uu.nl
Dieter Van Uytvanck, CLARIN ERIC's Vice Executive Director and Technical Director, e-mail: dieter@clarin.eu
Francesca Frontini, Institute for Computational Linguistics "A. Zampolli", Pisa, Italy, member of the CLARIN ERIC's BoD, e-mail: francesca.frontini@ilc.cnr.it
Antal van den Bosch, Meertens Instituut, Amsterdam, the Netherlands, member of the CLARIN ERIC's BoD, e-mail: antal.van.den.bosch@meertens.knaw.nl
Darja Fišer, Institute of Contemporary History, Ljubljana, Slovenia, member of the CLARIN ERIC's BoD from 2016 to 2020, e-mail: darja.fiser@ff.uni-lj.si
Andreas Witt, Leibniz Institute for the German Language, Mannheim, Germany, member of the CLARIN ERIC's BoD from 2019 to 2022, e-mail: witt@ids-mannheim.de

https://doi.org/10.1515/9783110767377-002

undertaken are rooted in a strong commitment to the Open Science paradigm and the FAIR data principles. This also enables CLARIN to express its added value for the European Research Area and to act as a key driver of innovation and contributor to the increasing number of industry programmes running on data-driven processes and the digitalization of society at large.

Keywords: research infrastructure, language resources, language technology, open science, service interoperability, innovation, SSH

1 Introduction

In this chapter, the CLARIN research infrastructure will be presented from a strategic and organizational perspective. It is authored by some of the current and previous members of the CLARIN Board of Directors (BoD). Krauwer and Maegaard (2022) describe the rationale behind the choice to implement the original ideas for the sharing of language resources in the way that CLARIN is set up – that is, a distributed infrastructure covering a multitude of languages and disciplinary needs – and the provision of a range of tools for the processing of language materials, in alignment with the Open Science agenda. The same chapter also outlines the European interest in structural support for research infrastructures that paved the way for the establishment of the CLARIN consortium as an ERIC[1] in 2012. This chapter will focus on what the intellectual and monetary investments of the past 10 years have produced. The impact of the dynamics in the European ecosystem on the modes of collaboration and the strategic agenda will also be outlined. Additionally, the various types of impact and the sustainability of the uptake, the models of collaboration, the overall service provision and the innovation ambition will be reflected upon. But to start with, the *raison d'être* for CLARIN will be addressed from a philosophical angle.

1.1 The neo-Babylonian paradox

According to a well-known passage from the Hebrew Bible, thousands of years ago, every person on earth spoke the same language. One day, man decided to build a city with a tower that would reach into heaven. But while constructing this tower, the people began to speak different languages. Confused by this sudden emergence

1 ERIC stands for European Infrastructure Consortium, a governance model for cross-country collaboration on research infrastructure.

of multilinguality, the construction of the city with its impressive tower – which was called Babel or Babylon, from the Hebrew word for 'confusion' – was stopped. The story of the Tower of Babel teaches us a contradictory lesson. Language allows humans to communicate. Through language we can tell stories, make agreements, write poetry, plan the construction of skyscrapers, or discuss how to fight global warming. But language also leads to confusion and misunderstanding. Some decades ago, work began on a second Tower of Babel: the internet. Since then, the World Wide Web has connected billions of people across the world. Any device connected to the internet gives access to a wealth of information, ranging from ancient philosophy to tomorrow's weather forecast, and from wildlife documentaries to the quickest route from Vienna to Bangalore. Online discourses affect the outcome of elections and the way people respond to restrictions meant to reduce the impact of pandemics or other global crises. Data has become valuable capital for governments, commercial enterprises, and science. But the internet is not just a goldmine; it is also a junkyard. It is estimated that 80% of all data is unstructured and text-heavy (Sumathy and Chidambaram 2013). It can be written in any of over 7,000 known, actively spoken languages, and may contain fake news, hate speech, and spam. How do we deal with this neo-Babylonian paradox? The CLARIN infrastructure is rooted in the belief that understanding the dynamics of language is key to addressing the challenges of our time. Enabling the use of language materials in scholarly contexts through the sharing of language resources and tools, and strengthening digital literacy, the ability to use and understand language data of any type, are commonly seen by the various communities of researchers and developers involved in CLARIN as key vehicles for the increased understanding of human language in all its forms and facets. Empowering citizens in becoming more versatile and digitally literate in a multilingual world in turn empowers society at large to be more democratic and to more effectively pursue humankind's intellectual and cultural ambitions.

We live in yesterday's future and tomorrow's past. Language has brought humans a great deal. The digital turn in communication as well as the pervasive access to information resources and Artificial Intelligence can help boost the potential impact of language-based service provision, and disentangle the neo-Babylonian paradox. With proper attention for language diversity and by advocating responsible use of the technology on offer, we increase the potential of language as a vehicle that not only allows humans to write history, but also to contribute to development goals for a better future.

1.2 Why language matters

Language is a carrier of socio-cultural content and information. Language also plays a role as the reflection of scientific and societal knowledge, as an instrument for human communication and persuasion, as one of the central aspects of the identity of individuals, groups, cultures, and nations, as an instrument for human cognition and creative expression, and as a formal system. Moreover, language materials form a considerable part of the historical records that are seen as cultural heritage. The faceted nature of language is reinforced by its internal dynamics, which has both synchronic and diachronic dimensions. Recognition of the value of understanding language in all its various facets and the importance of incorporating language data in the spectrum of data types that capture the full range of cultural and social dynamics has inspired the vision underlying the CLARIN initiative.

The CLARIN vision reads: "All digital language resources and tools from all over Europe and beyond are accessible through a single sign-on on-line environment for the support of researchers in the humanities and social sciences". In line with this vision, CLARIN was established as a research infrastructure with the following mission: "Create and maintain an infrastructure to support the sharing, use, and sustainability of language data and tools for research in the humanities and social sciences". The CLARIN infrastructure is thus rooted in the wide acknowledgement of the role of language as social and cultural data and the increased potential for comparative research on cultural and social phenomena across the boundaries of languages.

1.3 For whom CLARIN matters

With its richly faceted nature and its role in determining identity, context, origin, and use, language is a leading data source for researchers in the humanities and social sciences. At the same time, language data has also been recognized as relevant from the perspective of information science, data science, and Artificial Intelligence. CLARIN's aim thus has become to make language resources and tools available and reusable for all disciplines that work with language resources. And while the roots of the CLARIN research infrastructure were mainly in linguistics and language technology, the scholarly communities for which the infrastructure is operated also include fields such as Literary Studies, History, Journalism and Media Studies, Communication Studies, Ethnography and Anthropology, Migration Studies, Political Studies, Culture Studies, Mental Health Studies, Sociology, and Psychology. All in all, the activities taken up, the services developed, and the collaborative links with other RIs have led to a value proposition that, in princi-

ple, facilitates researchers working with language materials irrespective of the domain they are rooted in.

To reach out to its diverse potential user base and to stimulate the uptake of the services on offer in the relevant communities of use, in addition to the technical service provision for data sharing and processing through a distributed technical infrastructure, CLARIN has also developed an ecosystem for the exchange of knowledge and information and is investing in a network of experts on topics related to standards (Bański and Hedeland 2022), training (Wissik, Wessels, and Fischer 2022; Hennelly et al. 2022), and legal and ethical issues (Kamocki, Kelli, and Lindén 2022). The value proposition of CLARIN is also addressing the needs of non-academic parties, for example as embodied in the structural cooperation with the GLAM sector (GLAM = Galleries, Libraries, Archives, Museums) and the EU programmes promoting digital cultural heritage. CLARIN acts also as a driver of innovation in the European Research Area (ERA),[2] and the experts in the network provide advice and support on all aspects of the application of language technologies to European industry, both to SMEs developing Artificial Intelligence and Machine Learning applications, as well as in innovation projects set up in the context of the EU Digital Transformation and Recovery Plan across a wide range of industrial sectors.[3]

1.4 Key values: Open access and interoperability

The design, construction, and operation of CLARIN has been strongly inspired by the aim of facilitating the sharing of resources, providing a platform for open access, and stimulating the interoperability of data and services at all levels. The value attributed to open access has been operationalized by working towards a network of certified service centres distributed over all participating countries. The resources hosted by the centre repositories constitute the in-kind contribution from the members of the CLARIN consortium. Via the central services for metadata harvesting and the identity federation that enables login for associated researchers to the central services, access can be granted to the shared resources, irrespective of the centre in which they have been deposited. A crucial precondition for the effectiveness of this model for the sharing of language resources is the interoperability of the services. The harmonization of metadata is a prominent feature of the approach taken by CLARIN, but in addition to this kind of technical

2 See also action 8 in the ERA Policy Agenda: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/era-en.

3 The CLARIN Value Proposition can be accessed here: https://www.clarin.eu/content/clarin-value-proposition.

interoperability, CLARIN also promotes interoperability along other dimensions, in line with the demands of the Open Science agenda that are addressed in subsection 2.2. (See also de Jong et al. 2020.)

2 CLARIN as part of the European ecosystem of research infrastructures

CLARIN is positioned in the European Strategy Forum on Research Infrastructures (ESFRI) cluster "Social and Cultural Innovation", which largely overlaps with what is commonly referred to as the domain of Social Sciences and Humanities (SSH). Over the past decade, numerous cross-national initiatives supported by the participating countries and the European Commission have contributed to the ecosystem of European Research Infrastructures. The communities that initiated them have taken on the responsibility for enabling the production of new knowledge and innovation in order to help understand and tackle the societal, environmental, and economic challenges facing Europe and the world in the 21st century. Collaboration between the various research strands is often argued to be essential for the promise of advancing the level of excellence in foundational fields of study and the progress towards realizing the potential for impact, especially in research carried out in the context of agendas driven by societal missions. In addition, a crucial role is attributed to the availability of research data and infrastructural services that provide access to data and analysis tools.

2.1 The policy landscape

Partly under the umbrella of the European Strategy Forum on Research Infrastructures (ESFRI), a rich landscape of research infrastructures has emerged. CLARIN is one of the more than twenty ERICs that have been established. It is positioned in the ESFRI cluster "Social and Cultural Innovation", which largely overlaps with what is commonly referred to as the domain of Social Sciences and Humanities (SSH).

The Open Science agenda and in particular open access to data are at the heart of CLARIN's values. The objective of interoperability of data and services has paved the way for large-scale data sharing and growing reuse of language resources, but interoperability has also proven a crucial precondition for the increased support of multidisciplinary collaboration and comparative research agendas. In combination with the inherent multilinguality of Europe and the

growing attention paid to language equality, the Open Science agenda is bringing strong incentives for investigations into cultural and societal phenomena across countries and regions. It is CLARIN's ambition to consolidate its role in supporting the emerging research agendas for the SSH domain and to contribute to the innovation potential of the advanced models for interaction between people, data, and machinery (or tools) for data processing. This is facilitated by the strong embedding of the developers of tools and data collections in their local, culturally specific context, and the interoperability paradigm for the model of collaboration between the centres involved.

CLARIN ERIC is one of the infrastructures that have been established under the umbrella of ESFRI. The increasingly rich ESFRI landscape, with a growing recognition of the potential for collaboration for the thematic clusters,[4] collaboration among the established ERICs united in the ERIC Forum,[5] and the emerging European Open Science Cloud (EOSC[6]) are likely to offer interesting opportunities for rearticulating CLARIN's position and the activities aimed at the exchange of knowledge and best practices among research organizations, and to establish CLARIN's profile as a spoke in the more generic knowledge hub for Research Infrastructures (RIs) that is currently being developed.[7]

2.2 Response to the demands of Open Science

The advance of data-driven methods in academia and the promotion of paradigms for open access to research data has increased the need for data registries and data management services to adhere to the guiding principles that make data FAIR: Findable, Accessible, Interoperable, Reusable.[8] In principle, the size of CLARIN's potential user base in Europe could be as big as the entire community of professional SSH researchers, which in Europe alone is estimated to be around 500,000 scholars (=30% of the researchers from all domains).

Since the early days of CLARIN, the values of what has become known as the Open Science agenda have inspired the conception and development of the infra-

4 See the 2020 position paper of the five cluster projects: https://zenodo.org/record/3675081#.Yt71MexBzlw.

5 ERIC Forum aims at advancing the position of the ERICs in the RI landscape. For details, see https://www.eric-forum.eu/.

6 The way in which CLARIN participates in the process of realizing the EOSC is described here: https://www.clarin.eu/eosc.

7 Making Science Happen: ESFRI White Paper 2020, see https://www.esfri.eu/esfri-white-paper.

8 The FAIR Data Principles, Force11, https://www.force11.org/group/fairgroup/fairprinciples.

structure. Providing data in open access and the sharing of language resources in order to allow reuse have been central to the approach adopted. Furthermore, providing open data, open source code, and open standards can help ensure studies based on these open resources are reproducible and replicable, as well as allowing for proper recognition and citation of resources, in alignment with the fundamental principles of academic research. FAIRness of data as a concept did not exist at the time CLARIN was set up, but the CLARIN approach to data curation and integration was FAIR *avant la lettre* (de Jong et al. 2020). Interoperability guidelines have affected integration and collaboration at a range of levels, most prominently in the adoption of a common metadata standard (Monachini et al. 2011; Soria et al. 2014). This has paved the way for the development of a number of technical services that derive their added value in part from the distributed and multilingual nature of the CLARIN data offering: the Virtual Language Observatory (VLO; Windhouwer and Goosen 2022), the Federated Content Search (FCS), and the Language Resource Switchboard (Zinn and Dima 2022). This approach has also enabled the interoperability of data and services across the boundaries of regions, languages, and disciplines, which helped position CLARIN as an initiative that stimulates multidisciplinarity, especially among the various SSH domains.

Putting the principles of Open Science into practice can be an arduous endeavour, as it depends on an interlocked chain of responsibilities and practices. For Open Sciences to succeed, data collectors, curators, data stewards, providers, and researchers need to commit to the adoption of open standards, open data, open source code, and open access. Making language data openly available is particularly challenging. Firstly, for the most part, contemporary language data fall within the ambit of copyright protection, as most linguistic expressions qualify as their author's own intellectual creations. Apart from some rare cases, copyright law grants authors exclusive rights to reproduce their work and communicate them to the public. Secondly, a significant portion of language data relate to identified or identifiable natural persons, and therefore constitute personal data. Providing and processing personal data is restricted by the General Data Protection Regulation (GDPR). Despite these complications, CLARIN is striving to make its data as open and accessible as possible, and only as closed as necessary. This is achieved in part by negotiating contracts with rights-holders which grant as many rights as possible to end users via standardized licenses. Furthermore, a dedicated CLARIN Committee on Legal and Ethical Issues (Kamocki, Kelli, and Lindén 2022) keeps the community informed on new developments in data protection law and practice, with particular attention to solutions that allow sharing of relevant datasets in open access conditions (or as close to these conditions as possible). Finally, alternative approaches are explored, to communicate the results of certain operations on data to the end user, without sharing the underlying data.

2.3 Collaboration with other RIs and platforms

The vision of borderless and seamless interoperability between data and services has recently provided a fertile ground for initiatives such as EOSC and the SSH Open Cluster a model for collaboration between RIs in the SSH domain aimed at sustaining and expanding the results of the cluster project SSHOC (2018-2022). The CLARIN infrastructure has been and will remain closely connected to these upcoming cloud platforms. Similarly, CLARIN has forged active collaborations with consortia and portals that promote language equality and easy access to digital resources, including language resources, such as the European Language Grid (Rehm et al. 2020), Europeana,[9] and the European Open Science marketplaces – the EOSC Portal[10] and the recently launched SSH Open Marketplace.[11] With the reduction of the traditional obstacles for (re)using data from other domains and the sharing of results, it has become clear that the interest in language material as an object of study is shared by quite a range of disciplines. The adoption of the interoperability paradigm has enabled CLARIN to take full advantage of the potential for comparative research based on data from multiple periods, regions, and languages. This insight has led to a number of investments in improved metadata curation and harmonization, carried out in the initiative known as CLARIN Resource Families (Lenardič and Fišer 2022). For a growing number of data types and tools, a continuous and structured effort has been made to increase the diversity of those families in terms of languages and regional background.

The need to foster and encourage an even greater interoperability level within the Resource Families has led CLARIN to launch its flagship project ParlaMint, dedicated to the creation of comparable and uniformly annotated multilingual corpora of parliamentary sessions. ParlaMint is currently available in about 20 languages, and new data and languages are being added for parliaments in Europe and beyond (Erjavec et al. 2021, 2022). The adoption of a common encoding format – TEI ParlaMint – will enable comparative research on topics such as Covid-19 legislations, gender studies, and green transition, among others. The ParlaMint example shows how an infrastructure such as CLARIN can go beyond supporting open data practices and become an actor for the creation of resources that are FAIR by design, and the promotion of agendas for comparable research.

9 See https://pro.europeana.eu/page/clarin
10 See https://eosc-portal.eu/
11 See https://marketplace.sshopencloud.eu/

The increased interoperability of the overall service offering and the growing coverage of the Resource Families is beneficial for a number of the research agendas for which CLARIN aims to provide infrastructural support, in particular in the domains that aim at innovation roadmaps through multidisciplinary collaboration and data-driven methodologies, such as Digital Humanities, Artificial Intelligence (including variants such as human-centered AI), computational social sciences, and political studies.

3 Organizational structure of CLARIN ERIC

A robust and efficient organizational structure is a *conditio sine qua non* for the action lines undertaken to lead to sustainable outcomes. Moreover, a faceted sustainability strategy is crucial for any organization that is dependent on stakeholder support, and trust is necessary for establishing a community within and around the infrastructure. This holds true not only for CLARIN ERIC, but also more generally for any infrastructure or long-term research project. In this section, the model of organization and the rationale behind it will be outlined. The implementation of this model may inspire other infrastructural initiatives and the lessons learned may enable them to benefit from the experience gained during the 10 years of CLARIN's existence.

3.1 CLARIN as ERIC

The organizational structure adopted in CLARIN is, to a large extent, guided by the kind of legal entity that underlies the CLARIN organization. CLARIN is a so-called ERIC: a European Research Infrastructure Consortium. The ERIC model was introduced in 2009 by the European Commission (EC), which defines research infrastructures as facilities that provide resources and services for research communities to conduct research and foster innovation.[12] ERIC status can be granted to research infrastructures that comply with the conditions specified in the ERIC Regulation.[13]

12 See https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures_en.

13 Council Regulation (EC) No 723/2009 of 25 June 2009 on the Community legal framework for a European Research Infrastructure Consortium (ERIC). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009R0723.

3.2 CLARIN ERIC and national consortia

The ERIC model comes with a crucial role for the membership of countries that form the basis of the *C* in the term ERIC: the consortium. Together, the countries form the highest decision-taking body in CLARIN ERIC: the General Assembly. As CLARIN is a distributed digital research infrastructure, which depends heavily on the decentralized service offering and the coordination between these developments, the role of the national consortia is a critical aspect at all other levels in the organizational model, as reflected in the representation of countries in the higher-level committees. The CLARIN website contains a section on its governance structure with an overview of the various bodies and their relationship.[14] The following paragraph describes their role and composition in more detail.

All member and observer countries create their own national consortia, which contribute to the construction and operation of the CLARIN infrastructure, as well as to the outreach to communities of use. For such a joint effort to be successful, coordination is required and, more importantly, collaboration. Each country is represented by a National Coordinator, who acts as the main liaison between CLARIN ERIC and the national consortium. To ensure effective collaboration between CLARIN ERIC's Board of Directors (BoD) and the national consortia, four committees are in place. All National Coordinators participate in the National Coordinators' Forum (NCF), the main tasks of which are to coordinate national activities, exchange ideas and experiences, and advise the BoD. In the monthly NCF meetings, the BoD reports about newly adopted policies and recent activities and solicits input from National Coordinators. The Strategy and Management Board (SAMBA), a subcommittee of the NCF, consists of a balanced delegation of National Coordinators. The SAMBA convenes at least every quarter to discuss matters related to strategic planning, budgeting, and financing with the BoD and to prepare decisions to be taken by the NCF. The CLARIN centre network offers sustainable access to resources, services, and knowledge. The Standing Committee for CLARIN Technical Centres (SCCTC) is responsible for the coordination of the activities of the technical centre network. Each member or observer country has a representative on this committee. The User Involvement Committee (UIC) coordinates the activities aimed at outreach to the relevant communities of use in the national context and to the visibility of their efforts in order to demonstrate the added value of CLARIN. By combining the diversified nature of a distributed infrastructure with a cooperative governance model, CLARIN can work towards its objectives in a truly collaborative manner.

14 Overview of CLARIN governance structure: https://www.clarin.eu/content/governance.

3.3 Central operations

A model has been implemented for collaboration and sharing of responsibilities among the Office team members, who work from a service-oriented mindset that contributes to the overall trust-building among the various national nodes and the central organization. The Office capacity covers topics such as training and education coordination, communication, event organization, technology watch, and collaboration with experts on web design and development. The responsibility for the day-to-day management of the central organization lies with the Board of Directors. On some aspects, the BoD is advised by thematic committees (see Chapters 3.2 and 4.1). The BoD is responsible for the development of multi-annual strategies, annual budget proposals, communication with the Scientific Advisory Board, the acquisition of externally funded projects, the communication with the EC, ESFRI, and other relevant policy bodies and international alliances, the approval of new centres, the models for funding (based on calls for expressions of interest) and grant approval, and as indicated above, collaborating with the various thematic committees and their governance.

3.4 CLARIN and ESFRI

As mentioned already, and as described in detail in this book's chapter on the history of CLARIN and how it all started (Krauwer and Maegaard 2022), CLARIN is one of the infrastructures that have been established under the umbrella of the ESFRI. CLARIN was included in the first ESFRI Roadmap and as of 2016 it was listed by ESFRI as one of its Landmark RIs. In many countries, the national consortia are eligible for infrastructure funding under the condition of ESFRI recognition. Therefore, many of the national CLARIN consortia are dependent on ESFRI recognition. In some countries, the national consortia for CLARIN apply for national funding together with the national DARIAH consortium.[15]

15 In many cases the collaboration has led to the adoption of "CLARIAH" as the common name.

4 Knowledge Infrastructure and Technical Infrastructure: The key pillars

In this section the two main pillars of CLARIN's activities will be introduced and discussed: the Knowledge Infrastructure and the Technical Infrastructure. While the two aspects are presented separately, they are highly intertwined; together they fulfil the overarching objective of bringing language resources and technologies to researchers, students, lecturers, and other users, and enhancing competences for those using them and the potential for impact along a range of dimensions.

4.1 Knowledge Infrastructure

An infrastructure such as CLARIN is built upon the sharing of knowledge, be it factual knowledge (where to find data or tools) or procedural knowledge (workflows, best practices, standards that are used to create, curate, and use language resources). While the technical infrastructure is built to facilitate the discoverability of tools and resources, the CLARIN Knowledge Infrastructure has been developed as the "glue" for the various communities engaged with CLARIN, and as the structure that aims at securing a continuous transfer of knowledge between diverse parties involved in the construction, operation, and use of the infrastructure. The first gateway to the CLARIN Knowledge Infrastructure is the CLARIN website, a channel for disseminating high-quality information aimed at the exchange of knowledge, explaining the organization of the infrastructure and the activities undertaken, and illustrating the function and use of the services. Via the website, researchers and scholars can also access a rich catalogue of video recordings of CLARIN events, many of which originate from the Annual CLARIN conference, which is another pillar of the CLARIN knowledge sharing strategy.

Another crucial element is the network of CLARIN knowledge centres (K-centres) which bring together expertise on specific domains, topics, data modalities, and so on. Currently the K-centres, which can be operated by a single institute/group or arranged as a distributed structure, already cover a large number of research topics, languages, and resource types. However, CLARIN's strategy aims at broadening the range of topics covered by K-centres, incentivizing closer cooperation between them, and promoting their geographic distribution across CLARIN member countries. Knowledge offered by K-centres, the certified techni-

cal centres and the national consortia is also promoted by the Tour de CLARIN,[16] an annual publication showcasing resources and competences from CLARIN's distributed network.

The CLARIN Knowledge Infrastructure, together with CLARIN national nodes, is an important source of support and information for researchers who need to comply with the requirements of FAIR and open data in their projects and activities. In particular, the Legal and Ethical Issues Committee (CLIC) offers guidance and expertise on matters of Intellectual Property Rights and licenses, data protection, and privacy, as well as ethical and scientific integrity and responsible data science, while the Standards Committee offers advice on the standards to be supported and adopted within the infrastructure.

The Knowledge Infrastructure also aims to play an important role in training the next generation of scholars in specialized competences and skills, while supporting teachers and trainers throughout the network. The DH Course Registry (Wissik, Wessels, and Fischer 2022) is a joint initiative with DARIAH ERIC, which offers students an overview of the Digital Humanities programmes offered in Europe and beyond; in addition to this, a Teaching with CLARIN[17] section has been added to the website, hosting a selection of training materials shared by members of CLARIN's communities. The recognition of the importance of students, teachers, lecturers, and trainers as users of CLARIN has also led to dedicated support actions, both in terms of funding for the creation of training materials and of dedicated initiatives (such as the Teaching with CLARIN Award).

Finally, CLARIN's Knowledge Infrastructure has recently been strengthened by a network of Ambassadors, that is, recognized researchers in various disciplines, appointed by the central office to reach out to new communities of use. In spring 2020, during the COVID-19 pandemic, the CLARIN ambassadors were instrumental in initiating a series of CLARIN cafés, virtual events which are currently being held on a monthly basis, providing a platform for informal discussion on topics relevant for the infrastructure. The organization of cafés and other virtual events (including two virtual annual conferences) has provided us with a new way to engage with new research communities and to broaden CLARIN's user base, and will become a new element of the Knowledge Infrastructure in the post-Covid era.

16 See https://www.clarin.eu/Tour-de-CLARIN

17 See https://www.clarin.eu/content/teaching-clarin

4.2 Technical Infrastructure

Over the past few years, CLARIN has constructed a sound and robust technical basis to enable the sharing and reuse of language data and tools across institutional, disciplinary, and international borders. By its very nature, technology used for language processing is heterogeneous and country-specific: countries develop technologies that best cater to the needs of their official language. CLARIN's mission is to unify this heterogeneous landscape by building interoperable interfaces and a federated offer of thematic services (i.e., services addressing discipline-specific needs, in contrast to services with domain-independent functionality).

In contrast to many other research infrastructures, especially the single-sited ones operated in the domain of physics, CLARIN was never conceived as an RI that was to be built up from scratch. When CLARIN ERIC was founded in 2012, several of its centres had a long history of archiving, developing, and sharing language resources. Having this experience at hand was beneficial for newcomers to better understand what the result of investing in building up a new centre could look like. A stable repository, well-curated metadata descriptions, persistent identifiers, federated login, interoperable web services: seeing these in action elsewhere is often a better motivator than reading about their merits in a technical report, and having the capability to demonstrate parts of the Technical Infrastructure has always been crucial for reaching out to researchers and policymakers.

In the subsections to follow, the implementation steps and the building blocks of the Technical Infrastructure pillar will be outlined.

4.2.1 From founding principles to centre assessments

With the large interest in establishing technical CLARIN centres, the so-called B-centres, the need to formalize and assess the associated requirements quickly arose. This was a stepwise process, largely inspired by the founding principles that had already been defined in 2009.[18]

- **Principle of Independence**: Every participating centre is independent in its choices of internal organization and set-up as long as it adheres to the agreements that are defined for a smooth interaction within the network.
- **Principle of Service**: Every participating centre needs to make an explicit statement about the services it wants to offer and about the quality characteristics of these services.

18 See D2R-1a, Centres Network Formation, http://hdl.handle.net/11372/DOC-27.

- **Principle of Consistency**: Every participating centre needs to guarantee that the content it provides, when a unique and persistent identifier is used to refer to the content, will not change over time.
- **Principle of Interoperation**: Every participating centre needs to adhere to the set of interaction protocols and agreements defined within CLARIN.
- **Principle of Responsibility**: Every participating centre takes over a responsibility for the coverage of the services it offers.

These principles, balancing the freedom of technical and organizational choices with interoperability and standardization, reflect the philosophy behind CLARIN's infrastructure.

Throughout the preparatory phase of CLARIN that preceded the establishment of the ERIC and ended in 2011 (Krauwer and Maegaard 2022), the operationalization of the principles led to the first versions of the requirements for technical centres.[19] Afterwards this evolved into the B-centre checklist, with some incremental updates.[20] Just like CLARIN ERIC itself, the centre requirements are now 10 years old. Overall, they have not changed drastically: some centre types were scrapped, slightly controversial labels to measure the compliancy (gold, silver, etc.) eventually never saw the light of day. Still, the following interesting evolutions can be spotted, which also apply to other aspects of CLARIN's Technical Infrastructure.

More centres lead to more rules

In the early days, most centres that wanted to achieve B-centre status were actively involved in the drafting of the requirements and fully subscribed to the founding principles. While complying with the rules, later candidates introduced new boundary cases, leading to the introduction of new rules that from that point on applied to all centres, also when applying for re-certification (every three years).

Growth requires more predictability

With more centres queuing for an assessment, it is important that the rules are clear and predictable. Establishing a centre requires careful planning. While the overall construction period differs between individual cases, sudden changes in the rules should not interfere with this process.

---

19 See D2R-1b, Centres Network Formation – Centre types, http://hdl.handle.net/11372/DOC-28.
20 See http://hdl.handle.net/11372/DOC-78

The growing importance of multi-channel communication

To reach more ears at more locations, updates on the assessment procedure need to be broadcast more widely. To achieve this, regular bundled updates on the role of centres in the Technical Infrastructure are distributed under the heading "Centre News".

Overall, the evolution of the centre assessments has been continuously based on the founding principles mentioned above. These have helped to maintain a model that respects the diversity among the centres while maintaining technical compatibility, with changes where needed (e.g., moving from a two-year to a three-year period of validity for a centre's certification to maintain a time window that is in sync with the CoreTrustSeal procedure[21]) and stability where possible.

One principle that was not listed explicitly above was that of mutual trust between CLARIN and its centres. Nevertheless, this has played an important role over time. The proverbial carrot – in the form of recommendations, documentation, and best practices – has been used much more frequently than the stick. This in turn helped to keep up a positive and supportive atmosphere, which is probably at least as crucial as a sound technological framework for a research infrastructure.

4.2.2 Architectural approaches

Now that the technical centre model, and even more importantly the principles behind this model, have been introduced, we can take a look into CLARIN's infrastructural architecture. In this section, after introducing the technical building blocks, an overview of the related balancing acts will be given, concluding with some observations on the role of the people and the teams behind the Technical Infrastructure.

Technical Architecture: The building blocks

Without claiming to be complete, the following subsections will introduce some of the important parts of CLARIN's technical architecture.

21 See also https://www.coretrustseal.org/.

Repositories

The repository is the centrepiece of CLARIN's data infrastructure: it is the place that allows access to language resources via the web (HTTP) protocol, gives access to the associated metadata and persistent identifiers, and takes care of authentication and authorization. The repository is the primary access point for machine–machine communication (e.g., metadata harvesting), and most often also for human–machine communication (e.g., manual inspection of a deposited data set).

Each technical centre has a repository, which is subject to assessment. Internally, the assessment committee checks if all technical and CLARIN-internal requirements are fulfilled. Externally, the CoreTrustSeal assessment ensures that the repository is stable, well-maintained, and sustainable. Popular options for repository software are Fedora Commons and DSpace. For the latter, the LINDAT-CLARIAH/CZ team even created a CLARIN-specific version (Hajič et al. 2022), which has proven to be very popular.

An interesting development in the field of CLARIN repositories looks somewhat contradictory. First, there seems to be a growing interest in the adoption of large third-party open source repositories, such as DataVerse and the Zenodo-based InvenioRDM. An important point to note here is that these systems are not fully CLARIN-compliant off the shelf. Here, the need for one or more plug-ins providing this functionality seems obvious. On the other hand, many of the larger CLARIN centres have chosen to implement their repository system themselves, often based on home-made components brought together with a PHP-based frontend.

As always, it is impossible to predict reliably how the future of CLARIN repositories will look. Given the variation in the set-up of centres, however, it might very well be that both models will co-exist.

Metadata

Since the early conception of CLARIN, metadata has always played a key role in the architecture. This is illustrated by the fact that this book contains a full chapter on this subject (Windhouwer and Goosen 2022).

Persistent identifiers

The founding principle of consistency already demands the use of persistent identifiers to ensure reliable references to language resources. This principle was technically translated into the requirement to use the Handle system for persistent identification, based on its proven stability, scalability, and wide adoption. As of 2019, the Handle-based Digital Object Identifier (DOI) scheme is also recognized as valid technology for persistent identifiers. This important step – since

DOIs are an increasingly popular way of citing digital resources – was made possible when it became clear that some key requirements for the technical Centre assessment (the use of content negotiation for CMDI metadata) could be fulfilled by the DOI ecosystem.

Today, CLARIN ERIC is a member of both ePIC[22] (provider of handles) and DataCite[23] (provider of DOIs) and can thus provide access to both persistent identifiers to its centres.

Federated Identity

Language resources sometimes cannot be made openly accessible, due to copyright and privacy-related reasons, while agreements exist with the rights holder that allow the materials to be used for research purposes. In such cases it is important to allow for low-threshold access for researchers who can be granted permission. The use of Federated Identity, sometimes called Single Sign-On or Authentication and Authorization Infrastructure, ensures that a person can reuse institutional credentials (username and password) to access resources that are hosted elsewhere.

More details about CLARIN's implementation of Federated Identity, and some options for future steps in this realm, are described in a report on this topic.[24]

Interoperable web services and applications

Achieving interoperability between different language processing tools has always been an important goal in CLARIN's existence. At the same time, it is also a very ambitious goal that comes with many practical issues that need to be solved. Broadly speaking, there are two levels of interoperability we can distinguish.

Firstly, there is interoperability within the technology stack of a single centre. This level occurs most frequently, since interoperability is a matter of sticking to self-defined standards and the enforcement of these standards is quite easy. The typical case is an NLP pipeline for a single language hosted at one location. Many of these are described in the Tour de CLARIN.

Secondly, there are frameworks to interconnect services that are located at different centres, bringing the potential for a broader palette of tools but requiring more infrastructural efforts to orchestrate the whole. A noteworthy example is WebLicht (Hinrichs, Hinrichs, and Zastrow 2010; Dima et al. 2012), because it has

22 See https://www.pidconsortium.net/.

23 See https://datacite.org.

24 D2.7, SPF full extension, https://office.clarin.eu/v/CE-2017-1014-CLARINPLUS-D2_7.pdf.

also been maintained and developed over a long period and it includes services from many different CLARIN centres.

A simpler level of interoperability can be achieved by passing on a reference to a file and having it processed by the frameworks within a browser. Although limited in functionality and best suited for demonstration purposes, this is the approach chosen for the Language Resource Switchboard.

Finally, it is also worth mentioning that the rise of easy-to-use development libraries for Natural Language Processing (such as NLTK and spaCy) in combination with the popularity of Python and related frameworks (such as Jupyter notebooks) is enabling interoperability in many directions by combining a variety of APIs, including some based on RESTful web services. While requiring more technical skills from the user, these approaches allow by far the most flexibility. This insight is also the reason why CLARIN ERIC has included the topic "CLARIN for programmers" in its multi-year strategy.[25]

Federated Content Search

While it would technically be attractive to apply central indexation to all the corpora available in CLARIN, this is not possible – mostly for legal reasons: centres are not allowed to redistribute resources that are under copyright. Therefore the concept of Federated Content Search was conceived: queries are sent to the centres that host the corpora and the resulting hits are presented in a web application suitably titled "the FCS Aggregator".

This approach requires an enhanced level of infrastructural compatibility, just as it does for the interoperable web services. The initial "low-hanging fruit" approach, based on a simple text search, has been extended with a more powerful multi-layer search protocol,[26] which naturally requires more effort on the side of the implementing endpoints that do the translation for the central aggregator.

The tension between improved functionality and more stringent requirements on the part of the centres is a very apt illustration of some of the recurring infrastructural balancing acts that will be described in the next section.

25 See https://www.clarin.eu/content/vision-and-strategy

26 D2.9, Federated Content Search Engine v2 (software), https://office.clarin.eu/v/CE-2017-1035-CLARINPLUS-D2_9.pdf

4.2.3 Infrastructural balancing acts

In any infrastructure, but especially in a distributed one such as CLARIN, choices need to be made continuously between different organizational and evolutionary models. The options typically do not represent absolute dichotomies, nor do the choices have to be implemented in an absolute manner. Still, it is important to be aware of these options and the consequences of any choices made, as they tend to surface in many of the technological development tracks.

Shop window *versus* deep integration

Showing what CLARIN, as a growing distributed infrastructure, has to offer can be done in many ways. The simplest option is to create a virtual shop window (e.g., a portal or web page) with manually maintained descriptions about and links to the language resources at the centres. This is cost-effective and fast, but not so easy to maintain in the longer term. The other extreme is to create a deeply connected framework in which the resources can be accessed and used together (e.g., via a Virtual Research Environment). While this approach allows for better demonstration of the added value of the research infrastructure, it costs significantly more and requires strict protocols, standards, and policies on all sides to ensure a reasonable service level.

Central *versus* centres

Many parts of the Technical Infrastructure could be implemented and maintained centrally or decentrally. Originally, when CLARIN was initiated, all services were provided by the centres. Some of these offered many technical components and therefore played a crucial role as strongholds of the technology. Later, when the status of some of these centres changed over time, and the ERIC built up a central development team, several services were transferred to the central level.

It is mainly in relation to the technical services that fall outside the scope of language resources that the discussion about where to optimally position a component is raised. Transferring all of these to the central node sounds appealing in terms of efficiency, but misses the importance of decentralized know-how and scalability.

Similar discussions exist regarding the subject of running services in computing centres or networks of computing centres (organized as part of the European Open Science Cloud). Related debates also exist on the usage of commercially provided cloud services (e.g., for helpdesks or monitoring) versus self-hosting of such services.

Stability *versus* flexibility
An infrastructure needs to be stable. A static infrastructure provides optimal stability. On the other hand, staying up to date with upcoming requirements and technology stacks is a prerequisite to avoid obsolescence, and only regular updates provide a shield against huge migration operations with a high failure rate.

Related questions are when to apply the changes, and who can take the risk of being a first mover. CLARIN's history shows that it often makes sense if either the larger centres or the central node can take up these risks and share their experience with the rest of the centre network.

5 Strategy towards impact and sustainability

5.1 Human know-how: The real capital of CLARIN

Notwithstanding all the relevant considerations in the sections above, we should not forget to spotlight the single most important factor behind a successful technical infrastructure: the human know-how. While this aspect was already recognized during the preparatory phase, and has always played an important role up till today, ensuring that the built-up know-how reaches all centres remains challenging, if only because of CLARIN's growth. That is also why the Knowledge Infrastructure (see Section 4.1) is of such paramount importance.

A good example of successful knowledge maintenance and dissemination are the several cases where people who built up experience in designing and implementing the infrastructure passed on their knowledge to another centre as a result of changing jobs. Such scenarios are clearly a mark of success in the effort to maintain and distribute the infrastructural know-how, as is the informal and constructive atmosphere at the expert meetings. After all, it is often during informal discussions and brainstorming sessions that some of the key parts of the infrastructure first emerged.

5.2 The power of the distributed nature of the CLARIN service offering

For a research infrastructure such as CLARIN to offer a sustainable context for the various communities engaged in the development and uptake of the distributed and faceted thematic service provision, a balanced combination of stability and

progression is mandatory (Broeder and Odijk 2022). Capitalizing on the federated nature of the infrastructure has proven a critical precondition for remaining at the forefront of technology. Recognition of the contribution from over 170 local nodes that together form the basis for the access to language resources and the exchange of knowledge and expertise is another critical condition for a sustainable service offering.

5.3 Impact

In line with CLARIN's primary mission to enable scientific excellence, over the years a wide range of high-quality and innovative research projects have been realized that were supported by CLARIN tools and resources. A dedicated section on the CLARIN website presents a selection of impact stories that illustrate the variety of disciplines for which the CLARIN infrastructure has proven to be of added value.[27] In view of the number of professional researchers working on SSH agendas it is to be expected that with adequate instruments for enhancing awareness and visibility of the value proposition the scientific and societal impact realized thus far can easily be increased.

The potential for impact that CLARIN and the social sciences and humanities have on societal issues is also illustrated by several of the impact stories; and in addition, this potential is underlined by the next stage of the ParlaMint project, in which the harmonized parliamentary corpora that will have been prepared in around 20 languages will form the basis for studies aiming to capture the public debate on the COVID-19 pandemic from a comparative perspective. Similar investigations of public debate and the corresponding traces of information and opinions on social media channels are vital for studying and developing solutions for the major societal challenges of our time, including worldwide inequality, migration, and climate change.

The aim of fostering the sustainable development of our world is expressed in the Agenda for Sustainable Development adopted by the General Assembly of the United Nations (UN) in 2015. The UN identified 17 Sustainable Development Goals (SDGs). As an international research infrastructure, CLARIN shares these goals and aims to make a contribution towards achieving them. A living web page summarizes these activities.[28]

27 See https://www.clarin.eu/content/clarin-impact-stories.
28 See https://www.clarin.eu/sustainable-development-goals

The CLARIN strategy also specifies action lines aimed at realizing the potential for collaboration with non-academic parties. This is illustrated by the fact that in many countries, institutes from the GLAM-sector, often national libraries and archives, contribute to the work of the national consortia as partners, as they are increasingly adapting to FAIR principles for their language-heavy collections and archives as well.

The existing collaborative links with industrial parties in many regional contexts, for example, for machine translation and speech processing, function as stepping stones for a more systematic innovation strategy that positions CLARIN as a key driver of the digital transformation in society at large. Evidently many CLARIN tools and resources are desirable building blocks in commercial software development; language is an integral part of many AI systems (e.g., chatbots, recommender systems, sentiment mining) and the growing market for AI-powered innovations is likely to lead to a surge in the interest in CLARIN technologies and data.

To ensure that the potential for impact is realized and that the role of CLARIN in the RI ecosystem is sustainable, the uptake of the CLARIN service offering in the various communities of use is a crucial precondition. CLARIN will continue to seek collaboration with other research infrastructures, national infrastructural initiatives, and communities involved in the articulation of disciplinary research agendas that could benefit from the research enabling services offered by CLARIN. Language matters in some way or other in all disciplines and societal domains, but the value proposition will come across only with clear promotion, branding, instruction, illustration, and demonstration.

6 The next decade

Where could CLARIN be in ten years from now? Our future plans focus on:

- reinforced support for multidisciplinary agendas, within and beyond SSH;
- models supporting the use of heterogeneous data/AI;
- responsible use of technology;
- training/capacity development;
- collaboration beyond academia;
- collaboration beyond Europe.

Robustness has been and will continue to be a distinctive quality of CLARIN. In the coming years, CLARIN will sustain, improve, and consolidate both infrastructural pillars, that is, the Knowledge Infrastructure and the Technical Infrastruc-

ture. Researchers and developers will be stimulated to integrate (multi)disciplinary research agendas and domain-specific quality requirements in the thematic service offer. Education, training, and capacity-building will be offered and facilitated to enhance the skills of the developers involved, increase the level of data literacy among researchers and citizens, and contribute to the education of new generations of data professionals for whom language data will increasingly demand advanced methods and tools.

Bibliography

Bański, Piotr & Hanna Hedeland. 2022. Standards in CLARIN. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Broeder, Daan & Jan Odijk. 2022. Sustainability and genericity of CLARIN services in the Netherlands. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Dima, Emanuel, Erhard Hinrichs, Marie Hinrichs, Alexander Kislev, Thorsten Trippel & Thomas Zastrow. 2012. Integration of WebLicht into the CLARIN infrastructure. In *Proceedings of the joint CLARIN-D/DARIAH workshop "Service-oriented architectures (SOAs) for the humanities: Solutions and impacts" at Digital Humanities Conference 2012*, 17–23.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkar-son, Steinhór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darǧis, Andrius Utka, Mindaugas Petkevičius, Monika Briediene˙, Tomas Krilavičius, Vaidas Morkevičius, Sascha Diwersy, Giancarlo Luxardo & Paul Rayson. 2021. Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. http://hdl.handle.net/11356/1432 (accessed June 14, 2022), Slovenian language resource repository CLARIN.SI.

Erjavec, T., Maciej Ogrodniczuk, Petya N. Osenova, Nikola Ljubesic, Kiril Ivanov Simov, Andrej Pancur, Michal Rudolf, Matyás Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çagri Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevicius, Tomas Krilavicius, Roberts Dargis, Orsolya Ring, R. van Heusden, Maarten Marx & Darja Fiser. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* 1–34. https://doi.org/https://doi.org/10.1007/s10579-021-09574-0.

Hajič, Jan, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko & Pavel Straňák. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Hennelly, Martin, Langa Khumalo, Juan Steyn & Menno van Zaanen. 2022. Training of digital language resources skills in South Africa. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In Sandra Kübler (ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: System demonstrations*, 25–29. Stroudsberg, PA: Association for Computational Linguistics.

Jong, Franciska de, Bente Maegaard, Darja Fišer, Dieter Van Uytvanck & Andreas Witt. 2020. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. *International Conference on Language Resources and Evaluation (LREC) 12*, 3406–3413.

Kamocki, Paweł, Aleksei Kelli & Krister Lindén. 2022. The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Krauwer, Steven & Bente Maegaard. 2022. CLARIN – how it started. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Lenardič, Jakob & Darja Fišer. 2022. The CLARIN Resource and Tool Families. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Ljubešić, Nikola, Tomaž Erjavec, Maja Miličević Petrović & Tanja Samardžić. 2022. Together we are stronger: Bootstrapping language technology infrastructure for South Slavic languages with CLARIN.SI. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Monachini, Monica, Valeria Quochi, Nicoletta Calzolari, Núria Bel, Gerhard Budin, P. Caselli, Khalid Choukri, Gil Francopoulou, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Joseph Mariani, Jan Odijk, Stelios Piperidis, Adam Przepiorkowski, Laurent Romary, Helmut Schmidt, Hans Uszkoreit & Peter Wittenburg. 2011. The standards' landscape towards an interoperability framework: The FLaReNet proposal building on the CLARIN standardisation action plan. http://dspace.library.uu.nl/handle/1874/ 285299.

Rehm, Georg, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasil,jevs, Orians Anvari, Andis Lagzdin, š, Jūlija Mel,n, ika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals & Ondrej Klejch. 2020. European language grid: An overview. *International Conference on Language Resources and Evaluation (LREC)* 12, 3366–3380.

Soria, Claudia, Nicoletta Calzolari, Monica Monachini, Valeria Quochi, Núria Bel, Khalid Choukri, Joseph Mariani, Jan Odijk & Stelios Piperidis. 2014. The language resource strategic agenda: the FLaReNet synthesis of community recommendations. *Language Resources and Evaluation* 48 (4), 753–775. https://doi.org/10.1007/s10579-014-9279-y

Sumathy, K. L. & M. Chidambaram. 2013. Text mining: Concepts, applications, tools and issues – an overview. *International Journal of Computer Applications* 80 (4), 29–32.

Windhouwer, Menzo & Twan Goosen. 2022. Component Metadata Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Wissik, Tanja, Leon Wessels & Frank Fischer. 2022. The DH Course Registry: A piece of the puzzle in CLARIN's Technical and Knowledge Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Zinn, Claus & Emanuel Dima. 2022. The CLARIN Language Resource Switchboard: Current state, impact, and future roadmap. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.