CLARIN

Common Language Resources and Technology Infrastructure

VALUE PROPOSITION 2021

Table of contents

1 Value proposition at a glance

CLARIN is a networked federation of language data repositories, service centres and centres of expertise. CLARIN makes digital language resources available to scholars, researchers, students, and citzen-scientists from all disciplines, especially in the Social Sciences and Humanities (SSH), through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analysing and sustaining digital language data and tools. CLARIN supports scholars who want to engage in cutting-edge data-driven research, contributing to a truly multilingual European Research Area.

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.

Vision

All digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of research in the humanities and social sciences.

Disciplines

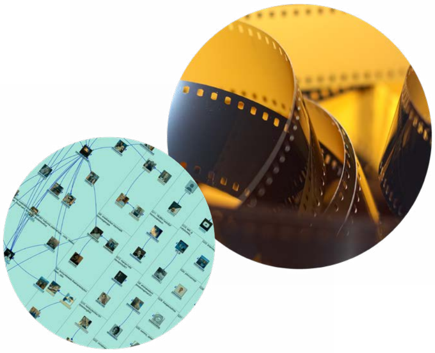
CLARIN stimulates the reuse and repurposing of available language data, thereby enabling scholars in the (digital) humanities and social sciences to open up new research avenues within and across disciplines that address one or more of the multiple societal roles of language. Why is language such an important type of data? It is a carrier of cultural content and information, both synchronically and diachronically, but it also plays a role as the reflection of scientific, cultural and societal dynamics, as an instrument for human communication, as one of the central components of the identity of individuals, groups, cultures or nations, as an instrument for human cognition and expression, as a training source for data-driven analytics, and as an object of study or preservation. Because of this multitude of roles, language is a key object of interest and study for a wide range of disciplines.

Open Science

CLARIN offers an infrastructural facility, but it is also a player in the ecosystem that is working towards the vision underlying the national and European Open Science policies: interconnecting researchers across national and disciplinary borders by offering seamless access to data and services in line with the FAIR data principles.

Stakeholders

The construction and operation of the CLARIN research infrastructure (RI) has involved many different stakeholders, each with their own interests and expectations, ranging from individual researchers, academic organisations, research data archives, infrastructure service providers, funding bodies and governments, the European Strategy Forum on Research Infrastructures (ESFRI) to sectors that are not primarily rooted in academia, such as the data industry as well as galleries, libraries, archives, and museums.



> " CLARIN infrastructure has a big potential, not only as a network of language technologies, but also of invaluable expertise. "
>
> *Beatrice Nava, University of Bologna*

2 Value proposition for researchers, lecturers and students

2.1 Support for discovering and depositing resources

Researchers can search for language resources via metadata in the CLARIN catalogue called the Virtual Language Observatory (VLO), or search in the data itself, using the CLARIN Content Search service. The VLO contains references to more than 700,000 resources, the majority of which are hosted at CLARIN centres, but it also contains references to relevant resource collections maintained by other organisations.

With 24 CLARIN member and observer countries, the VLO covers many languages, both national and regional. It enables fast identification of relevant resources, allowing researchers, lecturers and students to reuse resources that already exist, rather than having to produce their own from scratch. The VLO enables creators of datasets to make their resources visible to others through the publication of the metadata. Researchers, lecturers and students from all around Europe and beyond can benefit from the advantages of the cross-border coordination and data interoperability which form the basis for the VLO. Easily accessible, user-friendly and manually curated overviews of the resources and tools available in CLARIN are also provided to researchers, lecturers and students from the digital humanities, social sciences and human language technologies through the CLARIN Resource Families initiative, which organises corpora, lexica and tools according to their type and includes listings sorted by language. The listings include the most important metadata and brief descriptions, such as resource size, text sources, time periods, annotations and licenses, as well as links to download pages and concordancers, whenever available. Hyperlinks to other relevant materials are also provided, such as the thematic CLARIN workshops and tutorials and their accompanying video lectures, as well as a list of key publications on the resources included in the overviews.

Long-term preservation

One of the fundamental services of the CLARIN infrastructure is making sure that language resources can be archived and made available to the community in a safe and sustainable manner. To help researchers store their resources (e.g. corpora, lexica, audio and video recordings, annotations, grammars, etc.) in a sustainable way, at least one CLARIN data centre in each country offers a depositing service. These centres have agreed to store resources in their repositories in accordance with commonly adopted standards, and assist with the technical and organisational details. This has a wide range of advantages:

- Guaranteed long-term archiving
- The resources can be cited easily and reliably as they have a persistent identifier, and
- All resources and their metadata are equally accessible and searchable throughout the CLARIN infrastructure, irrespective of their physical location.

Researchers, Lecturers and Students as Content Providers

Researchers are not just users of data and tools, but also providers in that they are encouraged to share the resources they created, if necessary in a protected way, so that others can build further on their results. In addition to synergic effects, sharing also has consolidation effects for research. Sharing is supported by the availability of repositories (see above) and stimulated by facilitating data citation and licensing (see below).

Data Citation

It is a major undertaking to produce a research dataset, and the academic world increasingly recognises the value of such contributions, and mechanisms have been developed for data citation that encourage creators of corpora or other data collections to publish their data. CLARIN offers a platform for data publication and subsequent citation, which contributes to resource visibility because of the search facility (VLO) and because of the use of persistent identifiers for referring to resources instead of notoriously unstable URLs. Additionally, CLARIN provides know-how on optimal ways of data citation, insights into and influence on the latest citation technology and standards, and access to data citation services (e.g. ePIC, DataCite and the Virtual Collection Registry).

Licensing

CLARIN centres make data available through licensing and clear conditions for use. This involves CLARIN centres making deals with rights owners, signing Deposition License Agreements which include End User License Agreements, categorising licenses in clearly marked license categories, and writing Terms of Service.

Guidance is offered to creators of data in order for them to select the most appropriate licensing conditions when publishing their data.

2.2 Advanced tools and computing facilities

CLARIN offers state-of-the-art tools and online services for many languages which support researchers as they work to annotate, analyse and publish their language data. Automatic annotation and analysis perform best on large amounts of data, and CLARIN makes it easy to combine data and tools from different repositories and centres within the Language Resource Switchboard.

Examples of the functionality offered:

Advanced analysis and visualisations for large data sets that may help gaining deeper insights:

- WebSty: is a state-of-the-art tool for stylometric analysis. It has been used to analyse language use in the Polish Parliament, but also to study the writing style of Hungarian literary authors.

Fast automated analysis of text and speech, leading to more time for the actual research:

- Automatic Speech Recognition (e.g. WebMAUS) can drastically speed up the transcription process of spoken language recordings, such as interviews, for various languages. CLARIN supports ready-to-use speech recognition tools via the Oral History Portal together with online tutorials by speech recognition experts.

Reproducible scientific analysis flows, leading to more data sharing and better replicability of research results, e.g.:

- CLARIN provides guidelines about creating reproducible NLP workflows and provides suitable computing infrastructure, in collaboration with the H2020 project EOSC-hub, for the actual replication. The CLARIN website provides an overview of workshops, shared tasks and reports addressing reproducibility.

- CLARIN is facilitating the development and wide adoption for standard formats to encode specific corpus families which enables interoperability and reproducibility. A pioneering showcase of this activity is the Parla-CLARIN TEI standard for parliamentary data which has been developed at the initiative of the CLARIN Interoperability Committee and has been successfully implemented in parliamentary corpora for several languages in the project ParlaMint.

Access to popular corpora through specialised query interfaces, e.g.:

- Parliamentary data sets available through a wide range of concordancers
- International Computer Archive of Modern and Medieval English (ICAME corpora)

Access to HPC facilities:

- For large and computation-intensive tasks, e.g. the training of deep learning models with GPGPUs, CLARIN can connect scientists to highly ranked High Performance Computing (HPC) centres. Depending on the amount of computing resources needed, the researcher might need to enrol in a competitive call to be granted access to a computing facility. In any case, the use of these HPC facilities can be offered free of charge.

2.3 Federated login: easier access to more resources

CLARIN has established a Service Provider Federation, i.e. a trusted network of identity providers that offer ‘single sign-on’. This means that researchers, lecturers and students can login with their institutional credentials to get access to protected language resources and applications in other countries.

One advantage for the researchers is that they gain time and have the benefit of having to use only one access code. Another advantage is that they also get access to protected resources in other countries. Without this single sign-on, researchers would either have no access to otherwise valuable resources, or they would have to apply for accounts for each repository.

The volume of protected online resources to which CLARIN gives easy access is still growing. Statistics show that over 90 visitors per day use the federated login to access CLARIN resources [1].

[1] Measured using Matomo during the first half of 2020 at the CLARIN discovery service. As not all service providers are using this service and CLARIN respects users that opt out of tracking, this figure is an underestimation of the real number. These remarks also apply for all other usage statistics in this document. See CE-2015-0528 for details.

2.4 Access to specialised expertise

Complementary to the access to data and tools CLARIN offers researchers, lecturers and students can access CLARIN's expertise through its knowledge infrastructure. All CLARIN centres offering access to data and tools operate a help desk (in English as well as in their local languages) where users can get information about the data and services offered, get help in using the services, and report any issues they encounter.

Certified CLARIN Knowledge Centres (K-centres) are institutions whose main mission is to share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure. K-centres serve researchers, lecturers and students from any discipline where language plays one of its many roles, ranging from object of study, a means of communication or expression, a means to store information, object of learning or teaching activities, a training source for data-driven analytics, and many others. K-centres all have their own specific areas of expertise, which can belong to many different categories, such as:

- Individual languages (e.g. Danish, Czech, Portuguese), language families (e.g. South Slavic) or groups of languages (e.g. morphologically rich languages, the languages of Sweden)
- Written text and modalities other than written text (e.g. spoken language, sign language)
- Linguistic topics (e.g. language diversity, language learning, diachronic studies)
- Language processing topics (e.g. speech analysis, building treebanks, machine translation)
- Data types other than corpora (e.g. lexical data, word nets, terminology banks)
- Using or processing families of language data that will exist for most languages (e.g. newspapers, parliamentary records, oral histories)
- Generic methods and issues (e.g. data management, ethics, IPR, OCR)

Services offered by K-centres can take different shapes. Apart from the help desk service, some offer online courses, some offer best-practice documents, some offer guidance in getting access to and using data and tools, some act as hosts for people traveling on CLARIN mobility grants, and there are many more models in which the expertise is offered and shared.

For those not sure which CLARIN centre to get in touch with, the CLARIN ERIC Office also operates a central help desk that replies directly to requests for help or information, or channels the requests to the most appropriate experts.

A wide-ranging overview of the work and expertise available in the CLARIN network is offered through the Tour de CLARIN initiative, which presents national consortia and K-centres, highlights their prominent resources and tools, and features interviews with the researchers, lecturers and students who have benefitted from the CLARIN infrastructure.

2.5 Online tutorials and teaching materials

The knowledge infrastructure also offers online tutorials in order to promote and stimulate the uptake of new techniques by researchers, such as Voices of the Parliament, CLARIN in EOSC, and use cases, such as the ones from a boot camp for librarians: Gender in Parliamentary Discourse, Annotated Corpus. These materials are provided by experts from national CLARIN consortia, and are available in English or with English subtitles. The use cases serve to demonstrate the successful application of digital methods to specific research questions, and to inspire researchers with similar or related questions.

2.6 Event organisation

CLARIN organises a broad range of workshops and webinars focused on specific topics where members of the CLARIN community get together and work on topics of common interest. (An overview of events is maintained on the CLARIN website.)

At the CLARIN Annual Conference infrastructure providers and users from all CLARIN countries exchange expertise and ideas. The event features a scientific programme with invited keynotes and contributions that are selected based on peer review, and presented as paper or poster. The Bazaar session offers an informal space where conference participants can meet people from other centres and countries, find out about work in progress, exchange ideas, and discuss ongoing and potential collaboration projects. In addition there are special sessions for early-stage researchers, the CLARIN Clinic provides room for individual consultation, and lecturers can share their experience and seek advice at the CLARIN in the Classroom session. Each year a post-conference proceedings volume is published with a selection of extended papers.

CLARIN Office is collecting best practices regarding the organisation of virtual meetings and events that have been published on the website. In addition to this, permanent support capacity is available at CLARIN Office for guidance and advice on preparing and promoting a virtual event.

2.7 Funding for workshops, seminars, mobility and development tracks

Through open calls, members of the CLARIN community are invited to submit proposals for funding of outreach events and workshops on strategic priorities for CLARIN.
Financial support is also available for smaller development projects or for the preparation of teaching materials. A cross-border mobility scheme for short stays at CLARIN centres stimulates the exchange of knowledge and expertise between staff or teachers of different centres, or between staff from centres and researchers or teachers.

" CLARIN already provides amazing resources and opportunities. I find the Federated Content Search – the ability to search across a range of CLARIN resources at the same time – particularly fascinating! "

Michaela Mahlberg, University of Birmingham

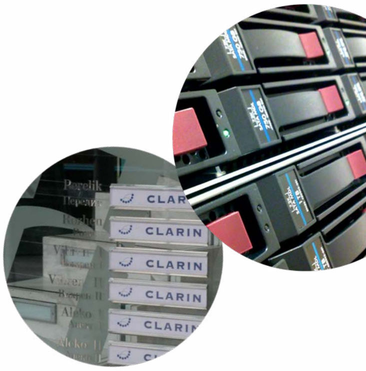
" I think educational activities should be a major priority for CLARIN at this stage, as this is the most efficient way for experienced and novice researchers to learn how to integrate the CLARIN infrastructure in their own work."

Sanita Reinsone, University of Latvia

"The fact that CLARIN exists has made me realise more that we as researchers do not work in isolation, but are part of a larger network. As such, I especially enjoy the events organised by CLARIN. For instance, the Twin Talks events in which people present how they work together with researchers from different fields provide wonderful ideas on how to tackle the interdisciplinary communication problems. "

Menno van Zaanen, North West University

3 Value proposition for academic organisations



There are a number of technical, organisational and social benefits that institutions and institutes can draw from their country's membership in the CLARIN infrastructure.

3.1 Guaranteed access to the CLARIN infrastructure and its services

The infrastructure offers and incorporates a wide range of services from which all users and all centres in CLARIN ERIC member countries and third parties can benefit. These services are developed and provided in close collaboration between the national teams and CLARIN ERIC. They include:

- Services for creating, sharing and reusing rich metadata descriptions: Component Registry (creation and editing of metadata schemas), Concept Registry (registration and reuse of semantic definitions), OAI harvester (automated distribution of metadata files).
- The CLARIN centre registry, a central dashboard to register services and to connect them to automated checking and publication processes (monitoring, harvesting, content search, etc.).
- Access to the advanced data services offered by EUDAT: e.g. B2DROP (workspaces) and B2SAFE (safe replication).
- Advanced distributed user statistics, using Matomo.
- Free CoreTrustSeal certification for each centre (normal fee: 1,000 EUR per centre).
- First-class access to Persistent Identifier services, including a prefix for each centre, thanks to CLARIN's membership of ePIC (for Handles) and DataCite (for DOIs).

Institutions and institutes are not only the homes of users of data and services, but they are also providers thereof. Research institutes can share the results and spin-offs of their research through the CLARIN infrastructure, thereby leveraging their visibility, whereas institutions with an archiving function can use CLARIN to make their digital collections, often created with public funding, more widely visible, accessible and (re)usable.

3.2 Access to a network of experts

The CLARIN community brings together a wide range of institutions and institutes dealing with various aspects of language resources. For institutions planning to establish themselves as CLARIN technical centres or aiming to keep themselves up-to-date with recent developments in research and education, various types of support are available, including funding for collaboration.

- Guidelines and documentation on the creation and maintenance of infrastructure services.
- Thematic workshops or tutorial sessions on specific infrastructure topics are organised to bring together staff from new and established centres.
- CLARIN is a forum for exchange of experience and collaboration, and supports the harmonisation of the regulations of the diverse academic institutions involved.
- A mobility scheme for short (typically up to one week) exchanges between individual representatives of new and established centres is running successfully.
- CLARIN hosts several committees as a venue for experts to discuss a variety of issues related to the development and offer of language resources (e.g. legal and ethical topics, standards for language data), and it supports network activities, such as the network of trainers.

CLARIN offers access to expertise and experts to all organisations looking for opportunities for collaboration across borders with relevant parties, and thereby facilitates the creation of new projects and activities across the borders of disciplines, regions and countries.

3.3 Participation in European projects

CLARIN ERIC actively participates in EU-funded projects. Participation in European consortia is considered whenever the envisaged project workplan can be aligned with CLARIN's general strategy and/or development. In many cases CLARIN has the opportunity to suggest involvement of one or more CLARIN nodes for a part of the workload. Joining CLARIN ERIC thus opens enhanced opportunities for participation in European funded projects, as staff of national CLARIN consortia may work on such projects together with or on behalf of CLARIN ERIC. This applies to CLARIN consortia in EU member and associated countries, as well as in all other countries that are eligible to participate in and receive funding from EU programmes.

By the end of 2020, CLARIN ERIC and staff from more than a dozen CLARIN ERIC member or observer countries are participating in projects such as SSH Open Cloud (SSHOC), a cluster project aimed at the collaboration among research infrastructures in the social sciences and humanities in the context of the emerging European Open Science Cloud, and EOSC-hub, a development project aimed at bringing together multiple service providers to create a single contact point for European researchers and innovators to discover, access, use and reuse a broad spectrum of resources for advanced data-driven research. More information about ongoing EU projects with the involvement of CLARIN ERIC and other CLARIN nodes can be found on the EU project section of the website.

3.4 Funding opportunities

Participating organisations and the developers, researchers and lecturers they employ can apply for a number of funding instruments that can facilitate outreach events and workshops on strategic priorities for CLARIN, travel for knowledge exchange, the preparation of teaching materials or smaller development projects.

```
dir = os.
suffix = 'a.xml'

def read_replace(filename):
    source_file = open(path + '/' + f
    text = source_file.read()
    source_file.close()
    new_text = text.replace('<td>', '
    return(new_text)

def new_file(filename):
    newfile = open(path + '/' + fil
    newfile.write(new_text)
    newfile.close()

filename in dir:
    filename.endswith
```

> “ Research infrastructures such as CLARIN represent an invaluable source of easily accessible resources, services and support for early-stage researchers, who are usually restricted to very limited funding and need help navigating the complex landscape of digital language resources. ”

Kaja Dobrovoljc,
University of Ljubljana

4 Value Proposition for non-academic organisations

Although the CLARIN infrastructure is primarily aimed at supporting the research community, some of its tools, resources and services are of high interest for non-academic organisations as well, particularly in the sector of galleries, libraries, archives, and museums (GLAM) and in industry.

4.1 Collaboration with the GLAM sector

In the GLAM sector, libraries and archives are likely to benefit the most from CLARIN services.

Libraries and archives host a wealth of documents. Today, an ever-increasing amount of these documents are digitised or 'born digital', which makes these institutions major players in the field of digital objects.

For centuries, documents have been collected by these institutions for specific purposes like archiving and distribution of the information. With the current state-of-the-art automatic processing of texts, the informational value of these documents can be maximised via the use of language technologies, such as automatic information retrieval, semantic information assessment, text summarisation, or linking of documents and knowledge bases. CLARIN is able to support the GLAM institutions in increasing the informational value of digital textual documents, e.g. through tools and workflows for enriching text with additional metadata and consultation regarding enhancing the various levels of data interoperability. For both libraries and archives, CLARIN can serve as an intermediary and contact point in establishing collaborations and preparing joint projects with academia.

4.2 Services for industry

The European economy is in transition, spurred by digital data-intensive technologies and the ubiquity of digital communication in modern society. CLARIN is at the forefront of several of the technologies that play a role in this trend, such as deep learning, automatic speech recognition, machine translation and AI (artificial intelligence). In line with the vision of ESFRI and the UN's Sustainable Development Goals [2], CLARIN aims to contribute to the steering and acceleration of the transition and to an inclusive, knowledge-based economy with potential for secure and sustainable economic growth. CLARIN is a non-profit organisation, so at present contacts with industry are primarily focussed on the exchange of information, but CLARIN is open to cooperation with the private sector that may have a positive impact on European industry, as well as research and society.

The importance of digital data for the economy of the European Union is steadily increasing. In 2018, the data economy in EU27 was worth 301 billion EUR (2.4% of EU's GDP), and the European Commission estimates that it will grow to 829 billion EUR [3]. Today, the technology companies are acquiring vast amounts of data from users and from information available on the web. Since the biggest companies in this field are not based in Europe, they adhere to non-European legal standards. CLARIN tools and services provide for a suitable alternative for the European information industry, emphasising the quality of data over quantity, and maximising its informational value through metadata, annotations and links.

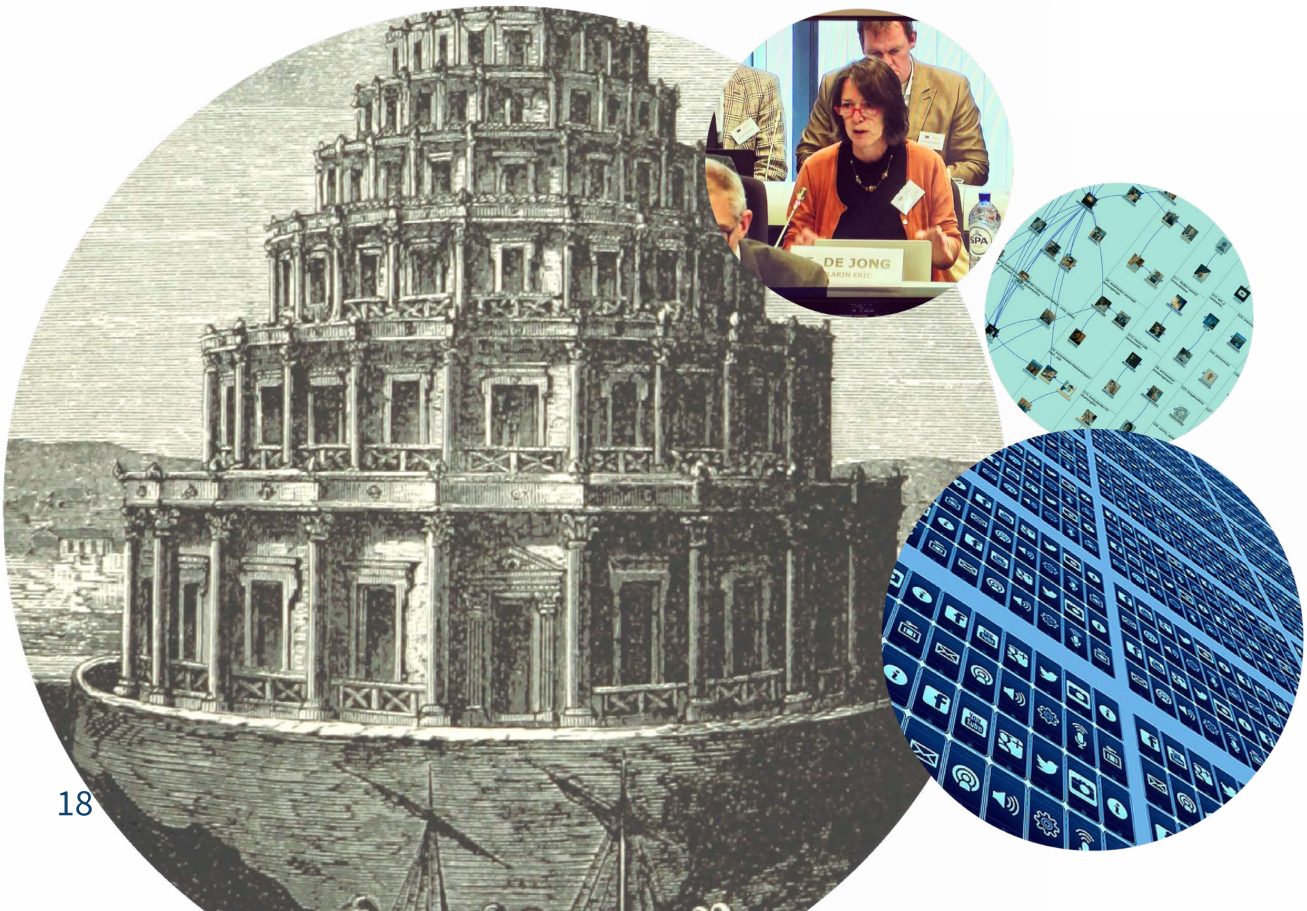
The EU is a multicultural and multilingual economic area, in which language-data-related services, e.g. human and machine translation, text analytics and language resources (e.g. dictionaries and terminologies) play a special role. For this sector CLARIN can offer access to an invaluable wealth of tools and data, part of which is available under Open Access conditions (i.e. not limited to research in affiliated institutions), and expertise and training for data-processing tasks. Moreover, in the local context many national CLARIN consortia serve as a liaison between industry and academia for those industry actors who seek academic partners for collaboration and joint projects.

[2]See goal number 8 out of the 17 goals listed here: https://www.un.org/sustainabledevelopment/sustainable-development-goals/ .

[3]https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en
This tendency for growth is even more visible in the USA, where the information industry currently contributes 6.5% of the national GDP. The US information industry experienced average annual growth of 9% since 2015, and its share in the national GDP increased 1.5 times in the past decade. https://www2.deloitte.com/us/en/insights/economy/spotlight/economics-insights-analysis-07-2019.html

Another field in which CLARIN can make a valuable contribution to European industry is the training of skilled data professionals. According to the European Strategy for Data published in May 2020, Europe faces a 'critical skills shortage' in the field of big data analytics with approximately half a million unfilled positions. By providing a digital research infrastructure to researchers in the field of humanities, CLARIN contributes to increasing digital literacy among graduates and young professionals, many of whom join the private sector at some point in their careers.

> "Making information sources accessible, analysing languages in them, linking objects, data and documents, enriching texts with metadata, and making accessible or referenceable the created virtual corpus through decentralised collections will enable a new way to interact with our past, understand the present and plan for the future."
>
> *Mikel Iruskieta, University of the Basque Country*



5 Value proposition for countries

CLARIN ERIC's value proposition for countries is to offer homogeneous approaches in a heterogeneous landscape. Europe is not a uniform area and is not destined to become one. Instead, it is a rich mosaic of languages, cultures and traditions to which European nations are deeply attached. This cultural wealth needs to be protected and promoted. In the research ecosystem this diversity is exemplified by a variety of national frameworks for the funding of research, as well as country-specific agendas and organisational models. This is particularly visible in the domain of social sciences and humanities, especially in the fields of linguistics and language technology, where, for obvious reasons, priority is given to research on the national language over English or other foreign languages.

At the same time, academic research naturally goes beyond the limits of national borders and – with the growing digitisation of data – is an endeavour with a strong international orientation. CLARIN develops approaches that can help generate even more synergy between European states in the research and development area.

5.1 A bridge between local and European practices and standards

In the realm of technological support for academia, unified connectivity through eduroam (which started in Europe) has proven to be a good example of a solution overcoming this tension between the national and international, but in other respects uniformity is neither convenient nor called for. Often harmonisation has to be built as an additional layer on top of national frameworks, preserving their uniqueness while establishing a common ground for collaboration, data sharing and knowledge transfer. This is what the CLARIN infrastructure aims for.

Since CLARIN services can be used across borders, members of the CLARIN infrastructure can benefit from reduced costs of development and operation of these services. Generic services can be mutualised, and some language-specific services can be ported between languages, rather than developed from scratch.

5.2 Increased visibility of national assets

In order to achieve its goals, the CLARIN infrastructure is developed in close collaboration with the national centres and consortia involved. CLARIN can therefore interconnect domain and language-specific approaches and provide individual solutions for local languages, their resources and digital history.

By participating in CLARIN ERIC, members increase the visibility of their language(s) and cultures, gain access to a growing catalogue of resources (data, tools and methods), and benefit from lower and shared costs for the development of new resources. Moreover, the CLARIN national consortia, the CLARIN centres, and all individual researchers from a participating country, can benefit from the various funding instruments, such as the ones listed in section 2.7.

For emerging national consortia in countries that have just joined or are preparing themselves for joining CLARIN ERIC, workshops are offered on how to set up the CLARIN infrastructure at the national level. Additionally, online information is available on a range of subjects such as: how to build a national consortium, how to use CLARIN metadata standards, how to prepare for centre certification, etc. The website contains an information section for (potential) new members.

5.3 A channel for staying up-to-date with RI policies

CLARIN member countries have a direct influence on the decision-making about all aspects of the infrastructure, ranging from construction and operation, to longer-term evolution and strategic priorities, through representation in the General Assembly, the National Coordinators' Forum, etc., as well as through participation in the various committees. Through membership of CLARIN, member countries have an extra channel for staying up-to-date with emerging policies related to the European RI landscape, including the dynamics around the European Open Science Cloud. More details about the governance of CLARIN ERIC can be found in the CLARIN Statutes.

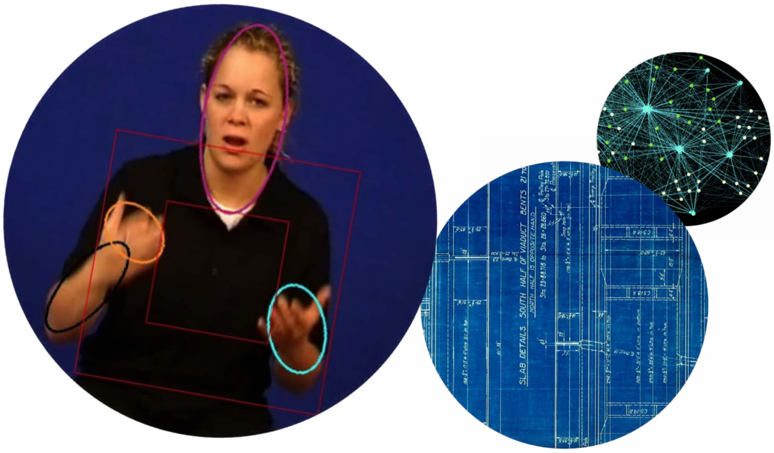
"The collaboration of the Phonogrammarchiv, as a K-centre, with researchers like us on how to handle materials on the cultural heritage of minority communities is an excellent example of putting the idea of CLARIN's research infrastructure into practice"

Beate Eder Jordan, University of Innsbruck

6 Values for Europe

6.1 Reinforcement of the European Research Area

CLARIN's primary role is to provide infrastructural support to all who want to engage in cutting edge data-driven research and to contribute to a truly multilingual European Research Area. CLARIN has developed a unique offer of multilingual language resources that can enable cutting-edge data-driven research and support scholarly excellence in a broad variety of disciplines. CLARIN is excellently positioned in the ESFRI cluster of Social and Cultural Innovation/Social Sciences and Humanities, and in particular in those disciplines where language is a significant data type, such as linguistics, language teaching, literary research, history, political science, cultural heritage, social sciences, language technology, artificial intelligence, data science, and so on. As increasing volumes of cultural heritage language data are becoming available in digital form, the potential for supporting digital humanities agendas is continuously increasing. There is also a vast potential for supporting multidisciplinary research that addresses Europe's societal challenges. As language is a data type that captures and reflects linguistic, cultural and social phenomena, CLARIN services can considerably accelerate progress in collaboration with researchers working with language data to address the missions of Horizon Europe. In the emerging European Open Science Cloud, CLARIN acts as one of the disciplinary spokes in the more generic knowledge hub for RIs that is currently being developed.



6.2 Enabling services for cross-border, cross-language research

CLARIN offers access to data and services that support researchers involved in language-specific projects. At the same time, CLARIN also facilitates and promotes cross-language research by offering overviews of data collections and tools particularly suited for cross-border, cross-lingual and cross-disciplinary research by the humanities and social sciences research community at large. If available for multiple languages, examples of such data types (e.g. newspaper collections, corpora of parliamentary records, social media data sets, and learner corpora) and tools (e.g. sentiment analysis, named entity recognition) reflect European culture and society. In addition to offering harmonised access to these collections, CLARIN has taken up the coordinating role of creating overviews of such comparable ‘Families of Resources’ and of bringing together researchers from various disciplines to explore the options for joint cross-border and cross-language research, and to exchange expertise.

Various instruments have been developed to address this ambition, including:

- Workshops on cross-border, cross-language or cross-discipline topics in order to ensure a continuous exchange of knowledge and expertise between different communities, and
- Initiatives for or participation in EU projects with a strong language dimension that bring together researchers from national CLARIN consortia, thereby providing excellent input and expertise for these projects.

6.3 Contribution to the Open Science agenda

CLARIN is not just a disciplinary facility, but also an infrastructural initiative that is well-embedded in the European research infrastructure landscape at large, and fully committed to the European agenda towards Open Science. By offering resources in open access and promoting the curation and depositing of data in alignment with the requirements for the interoperability of data and services, CLARIN has paved the way for large-scale data sharing and increased reuse of resources. In combination with the inherent multilinguality of Europe and the growing attention directed towards support measures for language equality, this Open Science agenda is another incentive for investigations into cultural and societal phenomena across countries and regions, and it reinforces CLARIN in its ambition to support the emerging research agendas for the SSH domain at large and to contribute to the innovation potential of the advanced models for interactions among people, machines and data.

> "CLARIN in three words: "Internationality, openness and user-friendliness!"
>
> *Tommi Jantunen, University of Jyväskylä*

Annex 1 CLARIN strategy 2021-2023

CLARIN at a glance

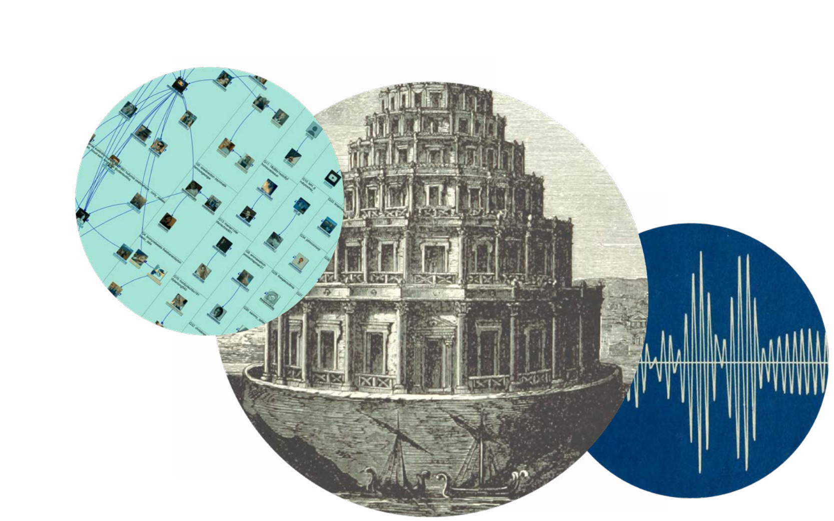
The Common Language Resources and Technology Infrastructure (CLARIN) is a virtual platform for everyone interested in language. **CLARIN offers access to language resources, technology, and knowledge, and enables cross-country collaboration among academia, industry, policy-makers, cultural institutions, and the general public.**

Researchers, students, and citizens are offered access to digital language resources and technology services to deploy, connect, analyse and sustain such resources. In line with the Open Science agenda, CLARIN enables scholars from the Social Sciences and Humanities (SSH) and beyond to engage in and contribute to cutting-edge, data-driven research. Furthermore, CLARIN facilitates researchers' participation in the European Research Area and the mission-driven programmes of Horizon Europe and the wider global context. The GLAM sector (Galleries, Libraries, Archives, and Museums) can use CLARIN as a hub for services to access and enrich digital cultural heritage objects and to establish collaborations with academic partners. For governmental institutions and the private sector, CLARIN offers tools and services to maximize the informational value of the growing amount of digital data. CLARIN is at the forefront of technological innovations, such as deep learning, automatic speech recognition, machine translation, and artificial intelligence. By organising the training of new generations of skilled data professionals, CLARIN contributes to reinforcing digital literacy and reducing the shortage of critical skills in the field of data analytics.

CLARIN is a distributed data infrastructure, governed and coordinated by CLARIN ERIC, a European Research Infrastructure Consortium, currently joined by 24 member and observer countries. CLARIN ERIC is recognised as a Landmark by the European Strategy Forum on Research Infrastructures (ESFRI). As a distributed research infrastructure, CLARIN consists of a federation of centres that offer resources, technology, and knowledge.
To promote the interoperability of resources and technology, CLARIN encourages the use of common metadata standards.

Impact

Language technology has become one of the most influential technological innovations of the data science era. Whether it is researchers using resources and tools to address new research questions, governments and industry applying text-mining algorithms to find valuable patterns in large amounts of language data and discriminating valid information from misinformation, or citizens using applications such as automatic speech recognition, machine translation or autocomplete, language technology is omnipresent. Unsurprisingly, experts have high hopes of applying smart language technologies to develop appropriate measures for today's societal challenges, such as climate change, inequality, pandemics, and illnesses. Each of these leave trails of language data, allowing researchers to study their causes and effects as well as the social and cultural dynamics stirred by them.



Priority areas 2021-2023

In the strategy period 2021-2023, CLARIN will focus on four priority areas to increase the potential for uptake and impact:

Sustainability

The stage of maturity now reached by CLARIN calls for greater attention to be paid to technical, financial, and organisational sustainability. CLARIN has had a considerable impact on a variety of fields, including digital history, media studies, language variation, election studies, social signal processing, translation studies, speech pathology, and migration studies. In addition, CLARIN helps in shaping the development of the European research infrastructure landscape, by actively contributing to advancing the models for collaboration among different research infrastructures. In the strategy period 2021-2023, further collaboration will be sought with non-European parties. To ensure CLARIN's future, the membership base will be consolidated and extended. Policies will be developed to protect valuable resources that are in danger of becoming inaccessible. Finally, the financial portfolio will be diversified and viable models for collaboration with industry will be articulated and promoted.

Technical infrastructure

The technical infrastructure is the foundational layer of CLARIN. Through a single sign-on environment, people from all over the world can access the resources and technology offered by CLARIN. Over the next three years, CLARIN will further invest in the three components that make up the architecture of the technical infrastructure: robustness, interoperability, and innovation. The visibility and interoperability of resources and tools will be enhanced by synchronising various sources of information and preparing for further integration into the European Open Science Cloud, the SSH Open Marketplace and other relevant platforms. Adherence to the FAIR Data Principles (i.e., making digital objects Findable, Accessible, Interoperable, and Re-usable) will be advocated both within the CLARIN community and to third parties. Relevant technological trends are closely watched and examined for their potential to be incorporated in the CLARIN infrastructure.

Knowledge infrastructure

In addition to resources and technology, CLARIN offers a broad range of expertise related to language data and tools. CLARIN's knowledge infrastructure is not only supported by thematic knowledge centres and committees, but also by large numbers of lecturers and teachers who have integrated CLARIN in their courses. Numerous initiatives, such as an annual conference, various workshops and events, an ambassadors programme, and Tour de CLARIN – highlighting contributions from national CLARIN consortia and CLARIN centres – help the exchange of knowledge both within the existing CLARIN community and with potential new communities of use and stakeholders. **By deploying this knowledge infrastructure and maintaining partnerships, it is ensured that CLARIN is aligned with the research agendas of the digital humanities, computational social sciences, and data science at large.** In the next strategy period, efforts will be made to reach out to policy-makers, library networks, science journalists, and the wider audience interested in the dynamics of data science. The visibility of services offered by CLARIN will be increased by improving the information structure of the CLARIN ERIC website, by offering materials like video tutorials and showcases, and setting up a sustainable network of CLARIN trainers.

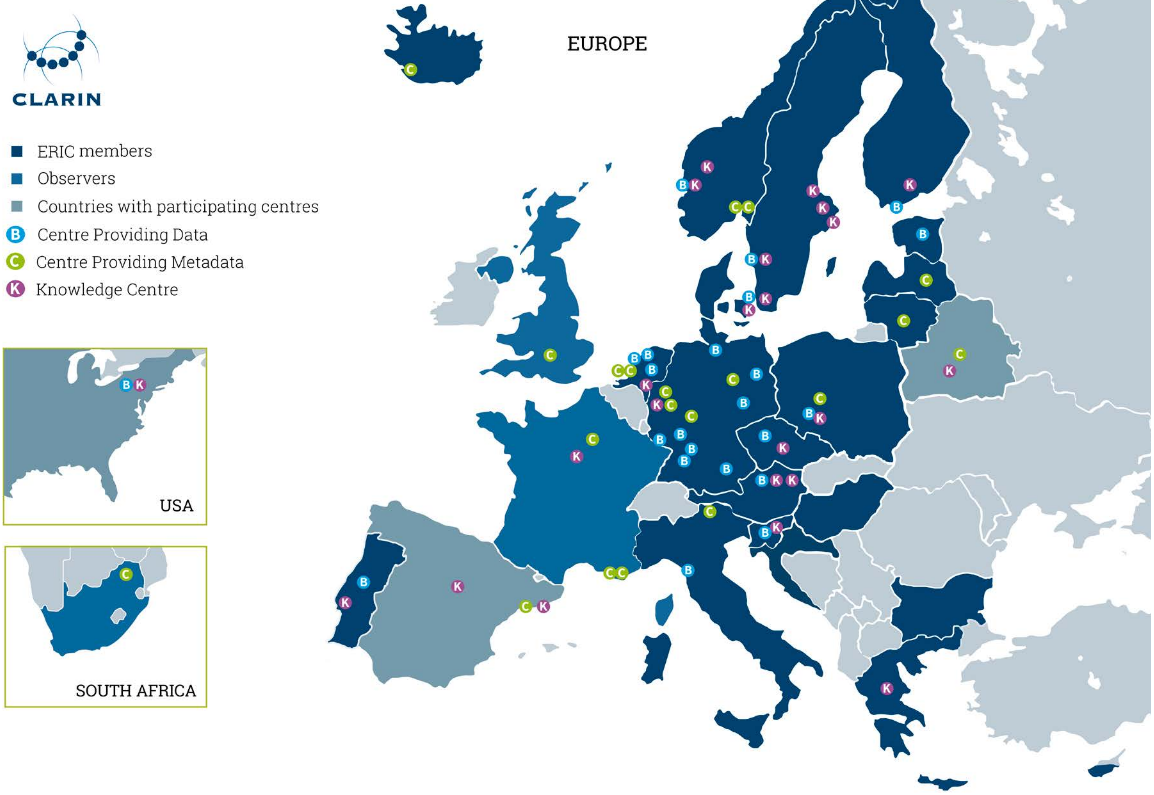
Organisational development

In recent years, CLARIN ERIC has gradually evolved from a relatively small project organisation into a permanent organisation with a professional governance model and competent support office. To optimise the use of human resources, efforts will be made to evaluate the current organisation and explore opportunities for reinforcement of collaboration among various CLARIN bodies and staying connected with existing and new networks. In alignment with other research infrastructures, arrangements will be made to ensure that the right instruments are in place to develop the capacities of the central staff. These endeavours will increase the expertise and skills of the central staff needed to build, maintain, and innovate the CLARIN infrastructure, while simultaneously reinforcing the career perspectives of the central staff. ERIC Forum, an alliance of European research infrastructures, has proven to be a particular fruitful consultative body in this respect.

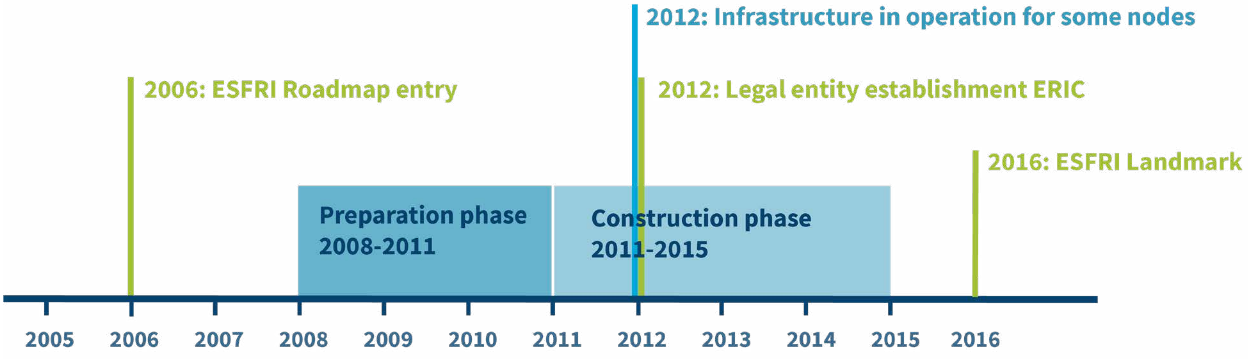
A more elaborate version of the strategy 2021-2023 can be found on the CLARIN ERIC website.

Annex 2

CLARIN map



Timeline


2012: Infrastructure in operation for some nodes
2006: ESFRI Roadmap entry
2012: Legal entity establishment ERIC
2016: ESFRI Landmark
Preparation phase 2008-2011
Construction phase 2011-2015
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

Colophon

This document is an updated version of the CLARIN Value Proposition (2017). The content is partly based on one of the deliverables of the H2020 project CLARIN-PLUS (D5.4 CLARIN Value Proposition, CE-2016-0847).

Editing & coordination:

CLARIN ERIC

Design:

Karolina Badzmierowska, Elisa Gorgaini

Proofreading:

Paul Steed

Contact:

CLARIN ERIC
c/o Utrecht University Drift 10,
3512 BS Utrecht
The Netherlands
clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure