Common Language Resources and Technology Infrastructure

CLARIN ERIC TECHNICAL and SCIENTIFIC DESCRIPTION

Document for final submission

July 2011

www.clarin.eu

Table of contents

CLARIN ERIC Technical and Scientific Description

1. Overall Objectives and Vision

The ultimate objective of CLARIN ERIC is to advance research in humanities and social sciences by giving researchers unified single sign-on access to a platform which integrates language-based resources and advanced tools at a European level. This shall be implemented by the construction and operation of a shared distributed infrastructure that aims at making language resources, technology and expertise available to the humanities and social sciences (henceforth abbreviated HSS) research communities at large.

In many countries as well as at the EU level much effort has been and is being put into the creation of digital collections of language based data (language resources, which could be in the form of textual, audio, visual or multimodal data) and the development of technologies and tools to explore, retrieve, exploit, study or enhance these data. This has led to a wealth of resources and tools, but geographically and technically fragmented and not easily found or accessed by HSS scholars without technological background. The CLARIN vision is to create a sustainable infrastructure that will provide the HSS research community with easy and lasting access to existing and future language resources and state-of-the-art tools, wherever they are located, thus enabling world class, innovative HSS research capable of crossing national, linguistic and discipline boundaries. CLARIN will empower HSS scholars working with language material to meet new research opportunities being introduced by the ever growing and accessible data collections and novel combinations of existing services.

Language is at the heart of many disciplines in the Humanities and Social Sciences (HSS), be it as an object of study, an instrument for communication and expression, or a means to record information and knowledge. Thus language resources and the tools that can extract knowledge from them are at the basis of scientific discovery in many disciplines. The multilingual nature of Europe constitutes a special challenge here, but over the last decades Europe has become a world leader in multilingual language processing, from which the CLARIN user community can greatly benefit.

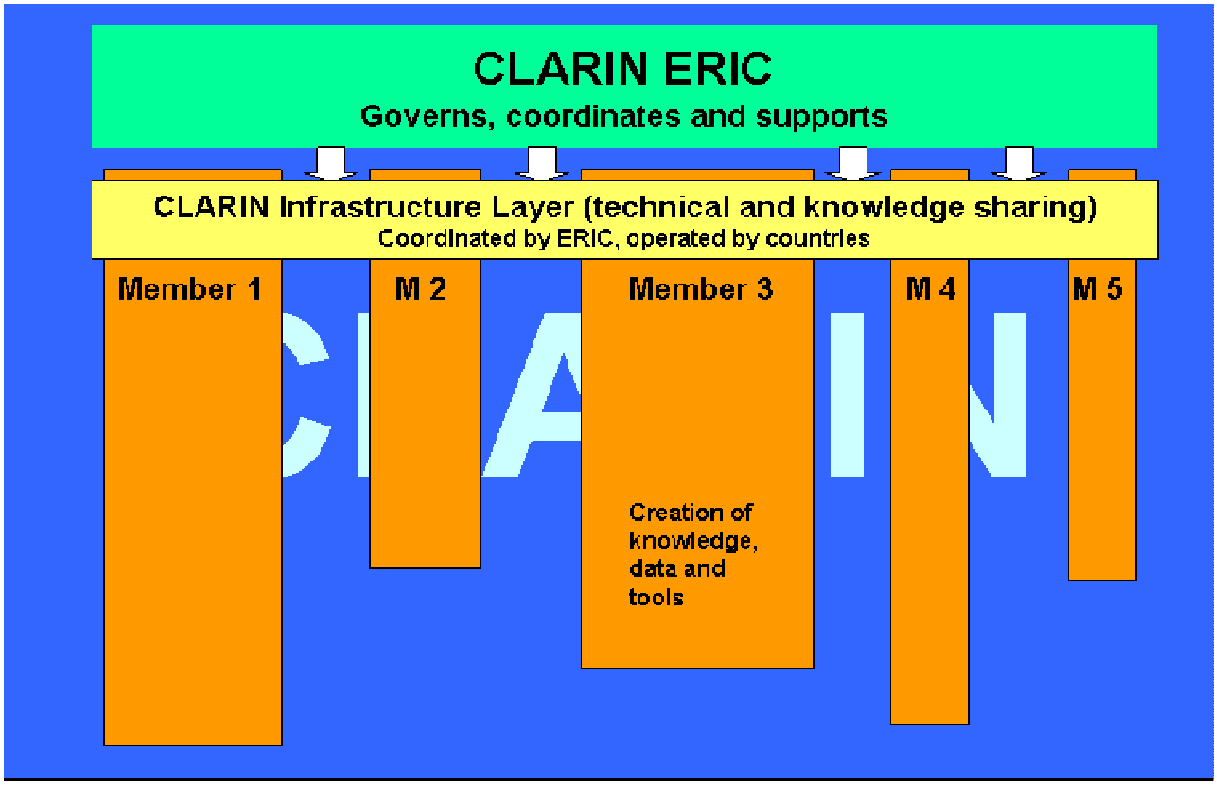
A number of features make CLARIN different from large scale document retrieval facilities such as Google. One is that CLARIN does not only give access to documents that can be found on public archives, freely accessible for everybody but also to archives with restricted access (e.g. for research purposes only), and that users can build virtual collections consisting of data found in different places. Another is that users have access to advanced language processing tools that enable them to perform operations on such collections, which allows them to ask questions such as „*Find all documents in CLARIN archives from the 18th century that speak negatively about slavery*“, or „*Find all recorded German TV news programmes from 2003 where German is spoken with a French accent*“.

To achieve its objectives CLARIN ERIC will undertake a variety of activities along various dimensions, the most important of which we will describe briefly here. First we will give a global overview of how CLARIN is shaped in terms of scope, organisation and responsibilities and the specific role of CLARIN ERIC.

2. Structure, scope and responsibilities

2.1. Structure

The picture below sketches the most important components of CLARIN and the way they are interrelated.



The vertical orange rectangles labelled „Member“ or „M“ represent the CLARIN-related activities in each of the member countries (or intergovernmental organisations). They typically comprise (i) the creation of expert knowledge, digital content, data, and tools (the lower part of the rectangle, also referred to as the Content Component), and (ii) the operation of technical facilities and organisational measures at the national level aimed at preserving, sharing and distributing the content (represented by the upper part of the boxes, referred to as the Infrastructure Component at the national level).

The horizontal yellow rectangle covering and interconnecting the Infrastructure Components of the members' activities is the core of CLARIN, and will be referred to as the CLARIN Infrastructure, where content, tools and knowledge are seamlessly shared and interconnected across member countries.

The role of CLARIN ERIC, which is represented by the horizontal green box labelled CLARIN ERIC, is to act as the governance and coordination body of all activities of the CLARIN Infrastructure layer, ranging from establishment, construction, and operation to further evolution following the advancement of research and technology. This governance and coordination role is reflected by the white downward arrows.
CLARIN ERIC is the element of CLARIN that turns a collection of unconnected or loosely connected infrastructural activities and initiatives at the national level into a truly European

Infrastructure, that operates on a European scale, at the same time respecting the subsidiarity principle by leaving the responsibility for the technical implementation of the construction and the operation with the members.

2.2 Scope and responsibilities

The scope of **CLARIN** (as the name of a pan-European endeavour) is wide and involves all aspects of creating, providing access to, sharing, maintaining and preserving content, comprising human knowledge and expertise, digital data, digital tools and services related to the use of language in HSS research.

The scope of the **CLARIN Infrastructure** is much narrower and covers the technical and organisational measures and facilities that make access, sharing, maintaining and preserving content possible. The creation of all forms of content, as well as carrying out research other than what is needed for the operation or instrumentation of the infrastructure does not fall within its scope, but it goes without saying that without the content the infrastructure would be of little practical value for the research community.

The focus of **CLARIN ERIC** is on the establishment and operation of the CLARIN infrastructure, with a view to providing access to the content, and not on the creation of the content itself. The responsibility of CLARIN ERIC is to provide the governance and coordination that is required to make the joint operation of the CLARIN Infrastructure by the members possible.

CLARIN ERIC Members have the following main responsibilities:
They contribute to CLARIN ERIC by an annual fee in cash, and a contribution in kind by participation in the governance and coordination activities of CLARIN ERIC.
They contribute to the CLARIN Infrastructure by providing infrastructure services, to be coordinated by CLARIN ERIC in order to ensure interoperability on a European scale.
They contribute to the CLARIN Content Component by providing (and funding) the creation and maintenance of expertise, data, and tools to be shared through CLARIN. Volume and nature of this part of the contribution is left to the discretion of the member, but in order to count as a contribution to CLARIN compliance with CLARIN standards is required. Coordination of such activities under the auspices of CLARIN ERIC is possible, but not required.

The statutes specify to which of CLARIN's activities members have to contribute, and for each member a detailed description will be contained in the CLARIN Agreement that has to be signed between CLARIN ERIC and the consortium created by the member for the execution of the tasks.

3 Main Areas of Activity within CLARIN

In this section we present the main areas of activity in CLARIN:
- The technical infrastructure
- The knowledge sharing infrastructure
- Content creation

3.1 The Technical Infrastructure

Language resources are expensive and time-consuming to create or collect, and only sharing them will allow the research community to build on the achievements of others and thus to advance HSS research rather than to reinvent wheels. The creation of language resources is the joint responsibility

of all CLARIN members. The selection of material to be digitized or created, tools to be built will normally depend on priorities following from national research programmes. The added value of CLARIN at the European level is that by supporting a set of standards for representation, interoperability and quality, resources can be exchanged and shared, and combined with other resources. The technical infrastructure will make this sharing possible in practice. The involvement of the ministries in CLARIN ERIC will help promoting the adoption of CLARIN standards in national programmes in a coordinated way at the European level.

Even if CLARIN standards will be widely adopted many digital archives and repositories still contain legacy material or provide services that don't comply with any external standards. A considerable effort will be required at the national level to bring them up to current standards, and CLARIN will offer a platform to exchange both expertise and tools to support this work at the European level.

The Technical Infrastructure is the core of CLARIN since it will provide access to the integrated and interoperable domain of language resources that can be seamlessly used by HSS researchers. From the user point of view there will be one single collection of data and tools, covering the offerings from tens or eventually even hundreds of centres all across Europe, which he (or she) can all find and access from behind the desk working with an easy to use web application. This should be possible for the user by using one identity, one single sign-on and by creating virtual collections he/she wants to work on, and virtual workflows[1] he/she wants to execute. The definition and maintenance of a collection of formal and de facto standards (emerging or existing), and the provision of tools and databases to support structural and semantic mapping will facilitate integration and interoperability between data and services, both within and across centres. This will allow researchers to ask old questions to new (and larger) collections of data, to ask new questions (based on new tools) to old data, and new questions to new data – not to speak of hitherto unthought-of combinations of data and research questions.

Given the current fragmentation and the differences between languages and their usage turning this vision into reality is a huge challenge which can only be approximated by a stepwise process.

Agreements and Standards

Most important is the establishment of an integrated and interoperable domain of language resources and services optimized for seamless access and use. Many standards and de facto standards already exist for the domain, both for resources and tools, and others still need to be developed.

It is important to distinguish between the data object level where interoperability is based on the external object characteristics, and the content level where interoperability is based on the structure and semantic encoding. With respect to the data object level there is a community wide agreement (and beyond) on the use of PIDs with associated information and component based metadata. Structure of metadata and vocabularies are subject of ISO standardization and schema/concept registries (ISO 12620) are already fully operational.

With respect to content level interoperability much more work has to be done. CLARIN is working on generic data models (ISO processes) for various data types and on registering and defining

[1] Workflows are sequences of operations that are executed to achieve a certain result which could be the creation of an annotation or the extraction of named entities to just mention two examples.

linguistic concepts (ISO 12620) for various linguistic annotation levels. The progress is widely dependent on the maturity of encoding systems, thus it will be a stepwise progress. To map between encoding systems CLARIN is working on flexible frameworks for the definition and registration of relations allowing users to easily manipulate them.

It should be mentioned here that Europeans, including CLARIN partners, have participated actively to the creation of standards for the field. As the adherence to common standards is a prerequisite to the success of CLARIN all CLARIN national consortia and other participants agree to contribute to developing standards and agreements, adapt them for their own language if necessary and adhere to them. Where necessary CLARIN ERIC will create working groups in close collaboration with the national teams and appoint chairpersons who are knowledgeable in their domain to foster harmonization actively.

Centres

It is widely agreed in many research infrastructures that centres form the backbone of the set of stable and persistent services that will enable e-Science. Language resources and tools are created in a highly distributed fashion and that will remain the major source for scientific progress. Turning such fragmented resources and tools into a stable, integrated and interoperable domain can only be done by specialized service centres that support the research process. Such centres can appear in various forms and can offer different service levels. According to the CLARIN classification there will be institutions that offer services, but cannot guarantee stability and CLARIN compliance. These are very useful, but they alone cannot create CLARIN. It is the centres that fulfil the CLARIN requirements that will actively contribute to offering the required services at the level needed:

- They will host language resources and their metadata and offer resource related services such as deposit, long-term archiving, metadata, resource access and resource utilization services.
- They will offer operational services by developing, deploying and maintaining standalone tools, web applications, virtual research environments and web services that can be used by researchers.
- They will offer expertise of various sorts such as giving support and help to users, maintaining high level knowledge, participating in education and training programs, developing and maintaining converters and taking care of agreed formats and standards.
- They will offer a variety of infrastructure services part of which will be provided in collaboration with CLARIN-external service centres such as large computer centres. A few examples can be given here, but this list may be subject to changes due to technological innovation and the emergence of the eco-system of infrastructures (see section 4):
    - Open registries will host important information about centres (harvesting addresses, etc.), persistent identifiers of resources and tools[2], categories that are used to describe linguistic phenomena and thus foster semantic interoperability and schemas that are used to specify the structure of resources and messages being exchanged.
    - For a smoothly operating distributed authentication and authorization domain, centres will be assigned to maintain information about the service provider federation, the state of agreements with the national identity federations and eduGain, the virtual organization database storing additional user information and other topics that may arise.

[2] As an example we can refer to EPIC (European PID Consortium) that will allow registering and resolving persistent identifiers.

- Portals giving access to the rich domain of language resources and tools will be established and maintained. They will include virtual research environments of various sorts, offer searching and browsing interfaces to find suitable resources and tools, carry out harvesting activities and the required semantic mapping, provide filters and workflow creation and execution environments and perhaps also maintain information about projects, persons and other contextual information.
- Of great importance in the future will be services that implement a web-based processing scenario, i.e. services that offer CPU time, storage capacity, workspaces for temporary results, etc.

The way centres will be organized may vary, some may want to focus on a single type of service, and others may offer several of them. This will also change in the future, dependent on technological developments. The services provided by each member will be described in the membership agreement.

Obviously, certain services will be offered simultaneously by a number of centres to achieve a high availability by redundancy. Also the development of certain services will be shared amongst a number of countries[3]. In particular certain areas of expertise will not only be hosted by one institution. CLARIN will set up virtual competence centres where necessary to bring the experts together.

As the CLARIN technical infrastructure centres are such important components in the CLARIN ERIC organisation, there will be a standing committee for CLARIN technical centres. This standing committee consists of the centre directors (or representatives designated by the directors) for all CLARIN technical centres. The role of the committee shall be to take implementation decisions, to coordinate the implementation, to give advice and make requests and proposals to the CLARIN ERIC and to the National Coordinators in order to ensure consistency, coherence and stability of services across CLARIN member countries and centres. The Annex to this document describes in more detail the types of centres that will be part of CLARIN and the way they will be assessed and selected.

Quality Assessment

One of the most important aspects in an increasingly anonymous domain of distributed services is centrally organized quality assessment. CLARIN ERIC will maintain a reference document that specifies the requirements at all levels. Each activity that is claiming being a CLARIN activity, needs to be evaluated according to the requirements. A close synchronization with quality assessment initiatives such as MOIMS-RAC, DRAMBORA and DSA is envisaged, in particular to assess the quality of the centres that are participating in the backbone of the infrastructure. Adherence of activities to agreements and standards will be assessed and associated with the CLARIN Compliance Seal (CCS). Activities which are compliant are allowed to use CCS.

Access

All researchers from all countries will have access to material for which no authentication is required by the owner.

[3] As an example we can refer to the development of the tools supporting CMDI (Component based MetaData Infrastructure) which was shared amongst The Netherlands, Germany, Austria and Sweden.

Researchers from CLARIN ERIC member countries who are part of an authentication system through their home institutions will have full, single sign-on, access to all facilities. No fees will be applicable except for services offered to CLARIN by third parties against a fee.
Researchers from CLARIN ERIC member countries who are not part of an authentication system and who can provide proper credentials can apply for guest status, and will have the same rights.
Researchers from countries that are not members of CLARIN ERIC will have full access if their home institution has signed an agreement with CLARIN ERIC and is part of an authentication system. The agreement will normally include an annual fee to be paid by the institution.
Special arrangements can be made with researchers from non-member countries who are not part of an authentication system, if they can provide proper credentials. A fee will be applicable.

3.2 Knowledge sharing infrastructure

The possibility alone to get access to more data and tools will not be sufficient to advance research and to integrate research efforts on a European scale. First of all the use of digital methods in the humanities and social sciences is not yet as wide-spread and well-developed as in other research areas, which means that a major education and awareness effort is needed to equip a whole new generation of researchers with the skills and methods to integrate digital methods in their day-to-day research activities. Secondly the vast amount of experience and expertise that exist in many different places in Europe can only be mobilized and exploited on a European scale through coordinated efforts. This means that in order to have a real impact CLARIN cannot rely on just providing and coordinating a technical infrastructure, but it needs to be accompanied by a knowledge sharing infrastructure, covering the whole spectrum from basic training and education to the creation of real and virtual centres of expertise where cutting edge research can be conducted and expertise and results can be shared.

An infrastructure where the individual researchers deposit data and deploy tools so that many can use them will only work, when there are experts who can help other researchers with advice about them. They will be the provide help-desk functions and consultancy, assisting with all kinds of questions.

Currently the number of experts that have a complete overview of the relevant collection of standards for example is relatively small, and mostly self-taught, also there are only few experts that combine fair knowledge about the application domain and the set of information technologies that need to be applied and extended. Thus one essential part is to set up systematic training facilities to educate our own experts who can then carry out the necessary acceleration in infrastructure building.

Expert Network

The content will be created and maintained on a broad scale by the national centres and experts, since primarily it is the task of each country to take care of their languages and the language material the corresponding centres are hosting for the benefits of the research. However, the knowledge about these resources, certainly the tools/services and the secondary resources, will be shared at European level and even beyond, e.g. the creation and maintenance of the data categories registered in ISOcat is a task shared by experts worldwide. Certainly in the start phase much support and help will be required to allow researchers working with the new facilities seamlessly.

Ethical and legal Issues

Another important component of expert knowledge has to do with the complex situation of rights, which cannot be ignored, and which can create real obstacles for progress. As many of the data repositories in CLARIN are distributors rather than owners of data, the rights of the original owners

and the constraints they have imposed on the use of their data, as well as the rights of the people described by the data have to be protected and respected. In CLARIN this will be done through the creation and maintenance of a licensing, access and authentication framework that on the one hand ensures light and easy access and at the same time protects the legitimate and reasonable rights of owners of data and tools and privacy of individuals. Experts are needed who deeply understand this complex matter in an increasingly international scenario, and who can translate this into suggestions guaranteeing usability. Experts give advice to depositors, centres and users of how to set up regulations that will help to overcome current boundaries hampering the optimal usage of an integrated and interoperable resource and tool domain. To this end, CLARIN will create a network of collaborating experts from a number of countries.

Education and Training

As was already successfully started at the national level, much education and training effort will be required to help the current generation of scholars, to create a new generation of researchers and to build up a network of young technologists to maintain the infrastructure and to support the researchers. A whole set of activities will be planned addressing the various levels of expertise and intentions such as working out university curricula and creating e-Learning objects, giving rotating lectures, organizing seminars, workshops and tutorials on specific topics, organizing summer schools. To make this feasible CLARIN will define modules of expertise that can be integrated easily into courses and have experts that can share the job of presenting these modules. As experienced in 2010, education and training are very time consuming activities and need to be spread on many shoulders. The fact that several countries have already started this and have built up some expertise, is a very good point of departure for CLARIN ERIC. Obviously there will be language-specific topics, but most issues can be dealt with by experts at the European level.

3.3 Creation of content

Above we have described the technical infrastructure as the backbone of CLARIN, but of course the technical infrastructure needs to be populated with language resources and tools, and we need expertise in all the disciplines that are contributing to CLARIN. This is what we call the Content Component. Like the technical infrastructure, this component is the responsibility of the national consortia. This means that the Content Component is based on the availability and usability of language related content stored in databases of various sorts (texts, recordings, annotations, lexica, grammars, time series, ontologies, etc.) and on the availability of human resources that give assistance in all kinds of ways and that make use of all types of useful information channels.

National consortia

It is the responsibility of each member to create a national consortium and to appoint a national coordinator. The national consortium will comprise those institutions (universities, research centres, libraries and the like) which are judged to have an interest and expertise in building up CLARIN resources and tools. All institutions in the country will of course be able to ***use*** the infrastructure, through the authentication and authorisation system. The list of the partners in each national consortium will appear as an annex to the membership agreement. The national coordinator will act as a main liaison between CLARIN ERIC and the national partners. At the same time, the national coordinator will liaise with the other national coordinators in order to ensure coherence and consistency across member countries.

Language resources and tools

Language resources and tools are the ***content*** that will populate the technical infrastructure. Many resources and tools exist already and these have to be integrated into the repositories at the CLARIN

centres. This involves a good deal of expertise, for creating metadata, for converting to other standards if relevant etc. So the curation of existing data and tools will be a major effort to undertake during the first years. The benefits of this endeavour are obvious: the fragmented legacy resources and tools become available in a uniform way, in larger collections with other resources, and to a much larger audience. Apart from integrating existing resources, it is a grand challenge for the national groups to create new resources and tools; this will be done in accordance with national programmes and priorities. The membership agreement has an annex describing these efforts.

It is important to keep in mind that CLARIN ERIC will not be responsible for the planning and coordination of the creation of data and tools at the national level. CLARIN members will be completely autonomous in the organisation of the creation of content according to their own national priorities. Their main commitment in connection with the creation of content is that, wherever possible, they will insist on sharing data, tools and expertise with the research community at large, and on complying with CLARIN standards.

CLARIN ERIC will offer a platform for its members to initiate and stimulate a variety of joint activities which may be joint projects based on national funds, or joint EC projects where ERICs have the right to participate as a full-fledged consortium. Such joint actions could very well be dedicated to the creation of content, but participation will be on a voluntary basis.

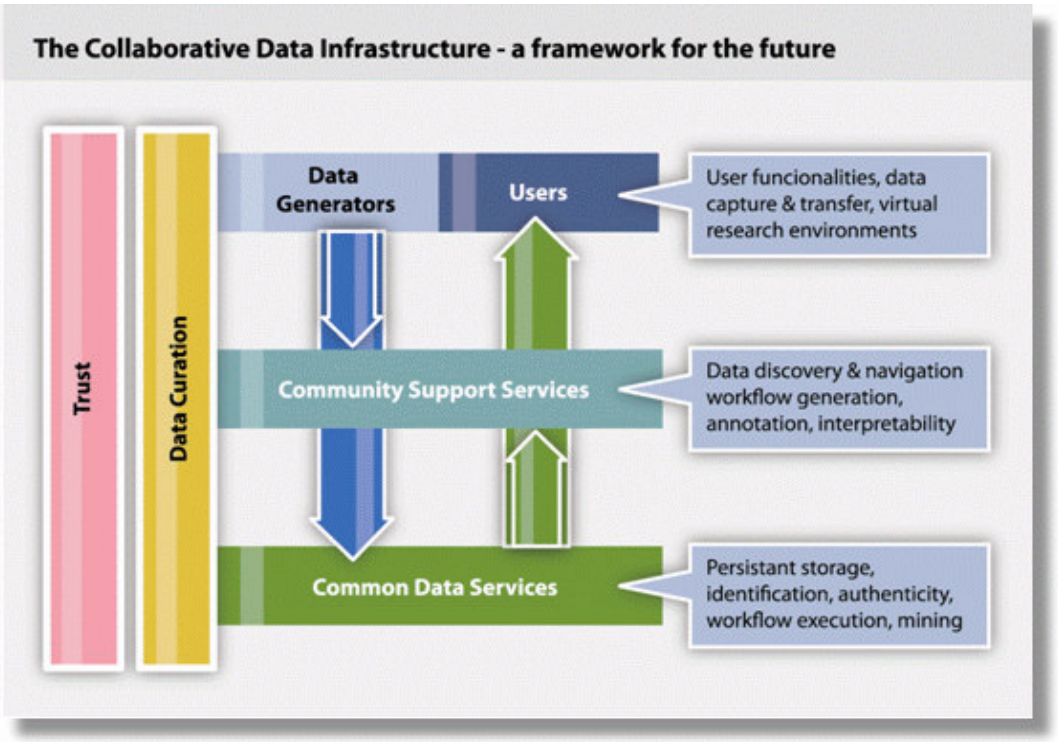
4 Embedding

The eco-system of infrastructures

The ESFRI process and the e-Infrastructure activities have clearly indicated that there is a broad awareness of the need to revolutionize the infrastructures to enable e-Science, thus CLARIN is embedded in a whole spectrum of activities in the HSS domain (e.g. DARIAH, CESSDA, ESS, SHARE, EHRI, centerNet, ADHO), in other research fields and by approaches delivering common services (GEANT, EGI, PRACE, ...). For the management of data, this collaborative scenario has been perfectly described by the High Level Expert Group on Scientific Data as denoted in the figure. Research infrastructures operating close to the researchers will make use of common services offered by data/computing centres for example. Along the discipline dimension CLARIN has a strong commitment to close collaboration with existing and emerging infrastructures in related areas and will actively explore collaboration possibilities with other disciplines where language plays a role.

To improve mutual cross-fertilization CLARIN is closely collaborating with DARIAH and also CESSDA in the DASISH project where two related topics have been defined: 1) create a portal for cross-discipline usable tools and methods and 2) carry out joint training courses etc. Also in some countries such as in Germany, Denmark, Netherlands and Austria a close collaboration even at workgroup level is planned. However, it will be a challenge to understand the various requirements, since in particular DARIAH is addressing a wide variety of disciplines**.**

At this moment CLARIN is participating through its members in a large variety of initiatives. Once CLARIN ERIC has been established, these collaborations have to be given an official status and should be channelled partly by CLARIN ERIC.



International embedding
Along the geographical dimension it is a high priority for CLARIN to liaise with related initiatives in further countries in Europe and other continents. Collaborations on world-scale are already in progress in the area of Digital Humanities (e.g. CHAIN, centerNet, Bamboo), in natural language and speech processing (e.g. Brazil), in the development and harvesting of metadata (e.g. OLAC, LDC, ISOcat), and in the documentation of endangered languages (e.g. INET). This collaboration is continuously expanding and will significantly increase the volume of data and services our research community will have access to, and open up for new collaborations.

5. The ERIC requirements

5.1 Necessity

HSS research is European: In Europe neither the history nor the present can be studied on a country by country basis. The present fragmentation of data collections across Europe (and even within countries) creates a significant stumbling block for researchers who want to get access to existing data collections to understand the broader perspective. Only a Europe-wide initiative can offer principled and sustainable solutions to this problem.

Language processing tools and services should be shared: From the perspective of technology and tools available to the researcher it should be noted that many EU and national programmes have led to the creation of advanced language processing tools and facilities that are now available in many countries and for many languages. By sharing these tools and facilities between research communities across Europe they can be more widely used and enhanced, thus giving a significantly better return on the investment made and enabling countries with smaller economies to benefit from

the results of European and national language and speech technology programmes. CLARIN offers the technical facilities, the agreements on standards and the organisational framework that are necessary to make this sharing possible at a European scale.

The ERIC as the governance body: During the execution of the CLARIN Preparatory Phase Project various potential organisation models for the coordination between tens, or even hundreds of data, service and knowledge centres have been investigated. Based of the consortium's own and (in parallel) the EC's explorations of possible models the ERIC has been selected as the best possible candidate, because it is firmly anchored in governments (as opposed to research institutions), which gives it the following important advantages:

- it is more sustainable than e.g. a project based approach
- it has more authority than e.g. an association or foundation set up by research institutes
- through the funding agencies in the participating countries it is in a better position to enforce sharing of research results and data, and the adoption of standards by the research community and by the operators of infrastructures at the national level
- it allows for bundling activities at the national level through national consortia and collaborating between consortia
- an ERIC can apply for funding from EC programmes as a full consortium

5.2 Strengthening the ERA

Joint access to data: Joint research across Europe requires joint access to the same data collections, across national frontiers. By interconnecting these collections and providing single sign-on access CLARIN ERIC constitutes a key facilitator for transnational research in the humanities and social sciences, which will give the ERA new impulses.

Broad dissemination of results: Using CLARIN as an instrument to disseminate and share on a European scale results obtained by national and international research activities, CLARIN facilitates building new research on the achievements of earlier efforts.

Crossing national and discipline borders: In general CLARIN will allow for interconnecting researchers, their data and their results across national and discipline borders. CLARIN will promote this collaboration and its new results, which may also be new methods and new paradigms for research. This way, CLARIN is one of the necessary building blocks for HSS research and HSS e-Science in the ERA.

5.3 Effective access

Access to all: As a principle, CLARIN aims to give access to all users, restricted only by the necessary licenses and privacy clauses.

Sharing limited capacity: CLARIN is a data and service infrastructure. This means that normally users will not experience problems cause by limited capacity. In those cases where capacity is a problem (e.g. in connection with access to specific computing facilities) CLARIN will make use of peer reviews to select high quality proposals, but at the same time reserve part of the limited capacity to provide new, promising players with access and support to help them to reach maturity.

Open access: CLARIN will actively promote open access, but it can not overrule restrictions following from legal or ethical considerations or access restrictions imposed by the rightful owner of

resources. Information about such restrictions and procedures to obtain permission to access restricted materials will be clearly indicated.

Taking away technical obstacles: As the target audience includes scholars without any technical background the most prominent obstacle for effective access may be the lack of expertise in making effective use of the data and tools which are available. The CLARIN knowledge sharing infrastructure will help educating future (generations of) researchers and supporting existing users.

Ease of access: Especially for its members, CLARIN will offer easy single sign-on access to those data and services that require authentication or identification. Non-members will still have access to the same data collections, but they may have to make their own arrangements with owners or centres to obtain permission, set up accounts and sign licenses.

5.4 Mobility

Virtual mobility: Data infrastructures such as CLARIN do not require or necessarily lead to mobility of researchers in the physical space. At the same time it generates an unprecedented virtual mobility in that researchers can access from behind their desk and even within one single workflow data and services residing in many different parts of Europe or, eventually, even the world. It will facilitate collaboration with colleagues, working with exactly the same data sets and tools, but sitting in different countries. Results that take the shape of new data or tools will, if possible, be fed back into CLARIN, so that the research community at large can benefit from them and build on them. By promoting and facilitating the use of Virtual Collections, and of citable metadata references to resources that have been used for a specific research goal, CLARIN will enable researchers to find out which other scholars are using specific resources.

Physical mobility: Virtual mobility cannot replace physical mobility entirely, and in CLARIN's knowledge sharing infrastructure member countries are expected to provide support for their own researchers for short visits to centres of expertise abroad or to receive researchers in their own centres of expertise. In the CLARIN ERIC budget an item 'travel grants' is foreseen under dissemination and training. Furthermore CLARIN ERIC can apply for funding from EC programmes to support this.

5.5 Dissemination

Dissemination in CLARIN is the main purpose of the knowledge sharing and technical infrastructure, along various dimensions, addressing different audiences and with different objectives.

For all researchers: The main purpose of CLARIN is to provide a facility to share (and thus disseminate) digital data and language processing tools on a European scale. Metadata services will help the user finding the data and tools he needs. Persistent identifier services will ensure that data can be referred to in a sustainable way. Centres of excellence, dedicated to advanced and specialized topics will help disseminating advanced knowledge and expertise through the research community.

For non-technical and future researchers: Broad adoption and integration of digital methods in the HSS research practice requires a massive training, education and awareness effort, as well as well-organised help-desk and support facilities. Such activities will be at the heart of the CLARIN knowledge sharing infrastructure.

For experts needed to support the further development of the CLARIN infrastructure: Special training initiatives will be deployed to create the next generation of experts to work on the infrastructure.

6 Concluding remarks

In this document we have presented the CLARIN vision and objectives, the way it is structured and the division of roles between the various components of CLARIN that can be identified at the European and at the national level.

We have highlighted the most important features and challenges of the technical infrastructure, presented the knowledge sharing infrastructure that goes hand in hand with the technical infrastructure, and we have emphasized the importance of the creation of content to populate the infrastructure with data, tools and knowledge.

We have shown that CLARIN is not an isolated incident but that it should be seen as part of an evolving ecosystem of research infrastructures, both at the European and at the global level.

In the last section we have explained how CLARIN satisfies the criteria for the creation of an ERIC as its governing and coordinating body.

The capability to conduct cutting edge research in the humanities and social sciences does not depend on the size of the language or the country, but is determined by the quality of the researchers and the access to relevant data and tools, which is why CLARIN ERIC has the ultimate ambition to cover all EU, candidate and associated countries. This is necessarily an evolutionary process, as not all countries find themselves in the same starting position that would allow them to join CLARIN ERIC from the very beginning. Typical obstacles can originate from many factors external to CLARIN, such as the timing of the national roadmap process, existing financial commitments and funding cycles, VAT issues, recognition of the ERIC as a legal entity, or the state of development of the infrastructure facilities at the national level.

The fact that not all countries that signed the CLARIN ERIC MoU may also be capable of becoming founding members does not mean that their will be no CLARIN activities in those countries. The CLARIN research infrastructure will reach beyond the borders of the (founding) member states because - as is foreseen in the statutes - CLARIN ERIC will make collaboration agreements with institutes from non-CLARIN ERIC member states that are valuable for setting up and operating the CLARIN research infrastructure.

ANNEX: CLARIN CENTRE TYPES

The purpose of this annex is to establish the procedure for selecting the centres that will participate in the Standing Committee for CLARIN Centres.

CLARIN distinguishes a number of different centre types that have different impact for the emerging language resources and tools landscape, and for completeness we list them all here.

1. Centre Types

At the core of the CLARIN infrastructure backbone are

1.1 Infrastructure Centres (Type A)

Task: Type A centres offer services that are relevant for the infrastructure as a whole and that need to be offered at a high level of commitment (stability, availability, persistence); in contrast to Type B they offer services that are used by other centres as well;
Examples: *joint metadata portal, data category registration service, schema registration services, etc.*
Requirements: Type A centres need to fulfil the requirements mentioned in chapter 2 where they do apply.
Agreement: CLARIN ERIC will sign a Service Level Agreement to specify type and characteristics of the offered services.

1.2 Service Providing Centres (Type B)

Task: Type B centres offer services that include the access to the resources stored by them and tools deployed at the centre via specified and CLARIN compliant interfaces in a stable and persistent way;
Examples: *the corpora stored at the centre, the language tools being developed by that centre, etc.*
Requirements: Type B centres need to fulfil the requirements mentioned in chapter 2 where they do apply.
Agreement: CLARIN ERIC will sign a Service Level Agreement to specify type and characteristics of the offered services.

1.3 Knowledge Centres (Type K)

Task: Type K centres offer expertise and advice about various matters that are relevant for the researchers to easily make use of the CLARIN services and that are not covered by the other centres;
Examples*: how to do the digitization, OCR and integration of book material, how to find taggers and parsers for medieval documents, etc.*
Requirements: Type K centres need to fulfil requirements which need to be specified in an agreement.
Agreement: CLARIN ERIC will sign a Service Level Agreement to specify type and characteristics of the offered expertise.

1.4 External Centres (Type E)

Task: Type E centres offer CLARIN relevant services, but these services are not offered by members of CLARIN; in general these will be common infrastructure services, i.e. external centres will often be type A centres[4];
Examples: *persistent identifier service, a long-term preservation service, etc;*
Requirements: Type E centres need to fulfil the requirements mentioned in chapter 2 where they do apply.
Agreement: CLARIN ERIC will sign a Service Level Agreement to specify type and characteristics of the offered services.

Most of the services offered by these centres are crucial so that CLARIN ERIC will sign Service Level Agreements with the corresponding centres, that specify the characteristics of the offered services, and will take measures to monitor the accessibility of them. In case of knowledge centres CLARIN ERIC will want to assess the quality of the advice that is given etc. In general Service Level Agreements with centres offering infrastructure type of services which are crucial for the whole will be formulated with a high expectation on availability.

[4] For CLARIN ERIC

Several CLARIN centres may give a mixture of service types, i.e. it is possible that very strong centres offer the resources stored by them (Type B), give advice about CLARIN relevant matters such as standards (Type K) and also offer some infrastructure type of services (Type A). This simply means that such centres take over more responsibilities.

There will be many more institutions that have interesting language resources and tools to offer, but who are not able or do not want to fulfil the CLARIN requirements and thus cannot offer core services. These can roughly be classified in two types:

1.5 Metadata Providing Centres (Type C)

Task: Offer machine readable metadata in a stable and persistent way allowing service providers to harvest their metadata and making them browsable, searchable and combinable;
Requirements: Type C centres are not requested to fulfil the requirements mentioned in chapter 2.
Agreement: there will be no Service Level Agreement being signed, i.e. researchers cannot rely on the availability of any service.

1.6 Recognized Centres (Type R)

Task: offer resources and tools via standard web sites, but that (yet) do not have funds to participate in the CLARIN infrastructure and that cannot give commitment statements;
Requirements: Type R centres are not requested to fulfil the requirements mentioned in chapter 2.
Agreement: there will be no Service Level Agreement being signed, i.e. researchers cannot rely on the availability of any service.

2. Requirements for CLARIN Centres (A, B, E)

The following list of requirements only holds for centres of types A, B and E

(a) Centres need to offer useful services to the CLARIN community and to agree with the basic CLARIN principles (own architecture choice, explicit statement about quality of service, usage of persistent identifiers, adherence to agreed formats, protocols and APIs).
(b) Centres need to adhere to the security guidelines, i.e. the servers need to have accepted certificates.
(c) Centres need to join the national identity federation where available and join the CLARIN service provider federation to support single identity and single sign-on operation based on SAML2.0 and trust declarations.
(d) Centres need to have a proper and clearly specified repository system and participate in a quality assessment procedure as proposed by the Data Seal of Approval or MOIMS-RAC approaches.
(e) Centres need to offer component based metadata (CMDI) that make use of elements from accepted registries such as ISOcat in accordance with the CLARIN agreements, i.e. metadata needs to be harvestable via OAI PMH.
(f) Centres need to associate PIDs records according to the CLARIN agreements with their objects and add them to the metadata record.
(g) Each centre needs to make clear statements about their policy of offering data and services and their treatment of IPR issues.
(h) Each centre needs to make explicit statements to the CLARIN boards about its technological and funding support state and its perspectives in these respects.
(i) Centres need to employ activities to relate their role in CLARIN to the research community in order to guarantee a research based status of the infrastructure and allow researchers to embed their services in their daily research work.

(j) Centres that are offering infrastructure type of services (A or E) need to specify their services for CLARIN and the terms of giving service.

Service Level Agreements will help to make all offerings explicit and describe the availability conditions. We foresee that it will take a while until all interested centres will achieve a fully CLARIN compliant state, therefore the evaluation process will associate a label (Gold, Silver, Bronze) with each centre: (1) **Gold** means that all requirements are functionally met. (2) **Silver** means that most essential criteria are met, but that there is still work to be done. (3) **Bronze** means that the centre can participate, but that essential functions are missing.

3. Centre Assessment Procedure

For all centres of types A, B and E the CLARIN ERIC shall have an assessment procedure that will check what the value of the services for CLARIN is, what the state of the services is, how the quality of the service can be evaluated over time and what kind of agreement will be required. To carry out this process the Board of Directors will set up an assessment committee, including CLARIN and external experts.

The procedure shall be as follows:

1. A negotiation phase will either be started by an interested centre or by CLARIN ERIC.
2. The centre and CLARIN ERIC will discuss the services to be offered and classify them.
3. The CLARIN ERIC will ask the assessment committee to check the state of the centre and the services.
4. A Service Level Agreement will be worked out and agreed upon.
5. The quality of the services will be assessed regularly.