Title	CLARIN ERIC Strategic Plan
Version	1.0
Author(s)	Bente Maegaard, Steven Krauwer
Date	03-04-2012
Status	For discussion by GA
Distribution	All meeting participants
ID	CE-2012-0001

CLARIN ERIC Strategic Plan

Introduction

This document presents the strategic plan for CLARIN ERIC for the coming years. |It consists of five parts.

In the first part ("The vision") we introduce the eight pillars that form the basis of the CLARIN concept and we indicate for each of them what the ultimate goal is that we want to reach.

In the second part ("The mission") we summarize the tasks to be undertaken in order to reach the goals. It follows the same eight pillars.

The third part formulates a number of guiding principles that should help us in developing and implementing the strategy, and in making choices and setting priorities.

The forth part presents for each of the eight pillars a brief sketch of the proposed strategy and formulates tentative goals for the first three and for the first five years.

The fifth part gives some concluding remarks.

The vision

The CLARIN vision is based on the following eight pillars. Each of them has been given a short label for the ease of reference in the rest of this document.

(a) ***Coverage:*** Eventually every scholar in the Humanities and Social Sciences (HSS) in every EU and Associated country should have direct single sign on access for research purposes to every single digital data collection containing language based material owned or made available by public bodies.

(b) ***Legal issues:*** No other restrictions on the use for research purposes should apply than those following from confidentiality, privacy or ethical considerations. The rights and legitimate interests of data owners should be protected at all times.

(c) ***Integration of data:*** Metadata and content search should allow scholars to find data if it exists, and they should be able to build virtual collections of material originating from different sources in different countries, and use them as if they were all residing in the same place and using the same representation standards.

(d) ***Integration of services:*** In addition to this scholars should have access to advanced language technology facilities (covering all modalities, such as text, speech, gestures) in the form of web services that allow them to annotate, explore, exploit, enhance,

analyse, manipulate and visualize these data in order to support their research. Web services should be able to operate on data from various sources and should be easily combinable to complex chains and structures to perform complex operations.

(e) ***Preservation:*** It should be possible to feed results of research projects as well as results obtained through application of services on data back into data collections in order for other researchers to use them. Data and results should be preserved in a sustainable way, and be provided with persistent identifiers to so that they can be retrieved later on to replicate results or conduct new research. In addition there should be persistent links to publications using or documenting the resources.

(f) ***Ease of access:*** Scholars should be able to understand and exploit the facilities offered by CLARIN without technical obstacles.

(g) ***Crossing borders:*** The CLARIN infrastructure should be embedded in the global research infrastructure landscape and should actively invite to crossing borders between disciplines, other infrastructures, countries and continents, as well as borders between academia and industry.

(h) ***Sustainability:*** The infrastructure should be financially, technically and organisationally sustainable over a longer period of time, but at the same time open to changes taking place in the infrastructures landscape.

The mission

In order to make the vision come true, work has to be done along a variety of dimensions. We sketch the main ones here, following the elements of the vision presented above.

(a) ***Coverage:*** Providing access to all data to all scholars in EU and Associated states requires bringing in all the countries, ensuring that in all countries all data collections become accessible, and ensuring access across national frontiers.

(b) ***Legal issues:*** Taking away legal obstacles requires developing and implementing a light licensing system, harmonization of IPR legislation across countries, and providing a proper access and authentication system.

(c) ***Integration of data:*** Metadata and content search across collections require adoption of common standards for metadata and data representation, and search facilities that can access a broad variety of data collections. This requires integration of resources both at the national and at the European level.

(d) Integration of services: Providing web services that can operate on data from a variety of sources and that can be combined with other web services requires a very high degree of interoperability, which can only be accomplished by means of standards. This requires integration between services and resources, and between services from various sources, both at the national and at the European level.

(e) ***Preservation:*** Feeding and feeding back (enhanced) data into the collections requires facilities for scholars to deposit their resources, and long term preservation facilities and persistent identifiers to ensure that they will be preserved and remain accessible and referable.

(f) ***Ease of access:*** Scholars should not need to be transformed to technicians, and access to the facilities and to the use of web services should be low threshold. User friendly interfaces, visualization methods, and good support services are needed. At the same time investments should be made in the training and education of future researchers in the use of the infrastructure are required, especially in those subareas where the use of digital methods has only a very short tradition.

(g) ***Crossing borders:*** CLARIN is European, and has its focus on the role language plays in humanities and social sciences research, but it should be careful not to lock itself up on a small, isolated island: relevant language data does not only exist in Europe but also in other countries and on other continents, humanities and social sciences scholars are rarely only interested in language, and language can also play a role in other disciplines. Academic research is not the only form of research, and in order for research results to be beneficial for European competitiveness, or for the European citizen bridges have to be made between CLARIN and players outside the academic community who could benefit from or contribute to CLARIN.

(h) ***Sustainability:*** Members have committed themselves to CLARIN ERIC for (in principle) 5 years, but the horizon of CLARIN lies far beyond this period. At the end of the 5 years period the future of CLARIN has to be secured, both organisationally and financially, but also taking into account new developments at the technological and infrastructural level, and our place in the landscape of infrastructures.

Some principles

Here we formulate a number of principles that should guide (but not dictate) us in developing and implementing our strategy.

(i) Separation of governance and coordination tasks on the one hand, and operational tasks on the other: the construction and operation of the technical infrastructure is the financial and organisational responsibility of the member countries. The rationale is that setting up central services would require new investments at the central level, would require an increase of the annual fees and create an additional flow of cross border funding. Making central services dependent on CLARIN ERIC funding would also make them more vulnerable from a sustainability point of view.

(ii) Keeping the size of the central coordination point small, and delegate tasks to teams in member countries where possible and desirable. Rationale: offices have a tendency to grow, and involvement of member teams in central tasks keeps the distance between central coordination and the work floor small.

(iii) Aiming at making access to and use of the infrastructure free for researchers in member countries. Rationale: contrary to industry where financial investments in research may eventually result in more profit, in academia the use of research facilities such as CLARIN should pay off in higher productivity or better quality, but not in cash.

(iv) Production of digital language data and tools is the primary responsibility of the members and will normally be guided by national research priorities. CLARIN ERIC will not dictate countries what to do, but will insist on compliance with CLARIN standards and it will

offer a platform for (voluntary) coordination of such activities between members so that synergies can be exploited.

(v) All data, tools and services offered through the CLARIN infrastructure will remain the property of the original owners. Depositing data in a CLARIN centre will not change ownership conditions.

(vi) CLARIN is open, and participation in centrally organised committees, events or dissemination activities is by default open to the research community at large unless this would be in conflict with the very nature of the event.

(vii) CLARIN should not duplicate anything that is already done by others or could be done by others.

Action lines and targets

For each of the pillars we will sketch very briefly

- what the starting point is at the time of beginning of CLARIN ERIC;
- our main strategy and instruments to move things forward;
- our provisional goal after 3 years;
- tentative success criteria for our self-evaluation after 5 years.

We will use the labels introduced above in the Vision and Mission sections as short section headers.

Coverage

Starting point

CLARIN ERIC starts with nine founding members: eight countries (AT, BG, CZ, DE, DK, EE, NL and PL) and one international organisation DLU. All were actively involved in the CLARIN Preparatory Phase (CLARIN-PP) through participation of at least one team. National CLARIN consortia are already operational in those countries or emerging. On the basis of criteria for CLARIN centres formulated by CLARIN-PP candidate centres are being identified and will be in the process of adapting themselves and their operations to the criteria.

Strategy

First priority is to consolidate the situation in each of the countries by establishing what exactly their contribution will be and ensuring that their contributions will be CLARIN compliant. The main instrument here is the CLARIN Agreement that will be made between CLARIN ERIC and the legal entity representing the national consortium. This will be done on a bilateral basis.

Next priority is to ensure that CLARIN centres are established or recognized in the member countries. We will set up an Assessment Committee for CLARIN Centres that will recognize centres or give recommendations for improvements on the basis of the criteria that have been formulated under the CLARIN-PP project.

We will proactively reach out to countries that have not yet joined as members and promote and support joining CLARIN. Researchers and institutions in those countries will be invited to participate in workshops and events organised by CLARIN, and we will maintain a network of

individuals and institutions to disseminate information about CLARIN, aimed at both providers and users.

Countries considering joining will be encouraged to join as Observers, and CLARIN ERIC can make specific collaboration and access agreements with institutions in non-member countries.

Goal 3 years
In 3 years time we want to have at least 15 members from EU and associated states.

Goal 5 years
After 5 years we want to have 20 members, with collaboration agreements with specific institutions in all other EU and associated states.

Legal issues

Starting point
The current legal situation is a jungle. In the absence of overarching European legislation comparable to the 'fair use' principle in the US (however poorly defined) Europe is faced with widely varying national IPR legislations and EU legislation where countries can still decide about their own exceptions. In addition to that many data collections are subject to restrictions imposed by the owners, or by ethical or privacy considerations. License conditions are non-uniform, and often inconsistent, incomplete or often even completely undefined.

Strategy
We foresee a 3-track strategy:

- building on work done in CLARIN-PP we will further develop a number of standard license templates that should cover most 'normal' cases, and we will encourage and promote the adoption of these templates for new resources developed by the national consortia; we will aim at harmonizing these templates with sister initiatives both from the HSS, from the language and speech technology community, and, if possible with the emerging new data initiatives;
- for legacy resources, for which the legal situation should be clarified before they can be integrated in the CLARIN infrastructure we will promote the adoption of the new templates, but we have to accept that in a number of cases the old licences remain in force;
- through CLARIN ERIC we will continue to bring the IPR problems to the attention of national and European legislators.

To protect the rights and legitimate interests of the owners of resources CLARIN will impose the use of a proper Authorization and Authentication (AA) system to ensure that only properly authorized users will get access to data with access restrictions.

The standing committee for the technical centres will be the body responsible for the coordination and implementation of the Authorization and Authentication system for CLARIN.

Goal 3 years
All new material created under the auspices of the launching members will be made available under the standard licensing templates, and all integrated legacy material will be provided with proper, complete and consistent licenses (legacy or CLARIN).

AA system fully operational in launching member countries.

Goal 5 years
Same for all members that have joined during the first 3 years.

Integration of data

Starting point
A considerable collection of data and tools exists, spread over hundreds of institutions all over Europe, but they are hard to find, hard to get access to, using a wide variety of encodings, annotations and metadata, and varying widely in quality and level of documentation. The situation is unbalanced in that written language data is far better represented than data in other modalities.

Strategy
The main action lines will be

- a standards action plan, whereby standards that have emerged from the CLARIN-PP project will be consolidated and adopted by the national consortia;
- as CLARIN will base itself on existing de facto and formal standards multiple parallel standards may co-exist in specific areas, and this will require the development of tools that allow for mapping between different standards;
- as all data will have to be brought to the same quality in terms of standards and metadata, tools will need to be developed that support the curation of data.

These activities will constitute a significant part of what will be described in the CLARIN Agreements, and are in principle the responsibility of the national consortia, but they should be coordinated at the European level in order to ensure overall coherence, consistency and completeness, and to avoid duplication of efforts by sharing tools and facilities.

The volume of this work should not be underestimated, as the curation and annotation of data in other modalities than written text is extremely demanding in terms of human resources.

Priority areas will be identified in order to be able to offer initial services on the basis of coherent virtual collections of data at the end of the first 3-year period. These areas will be chosen on the basis of

- an analysis of the offerings through the CLARIN Agreements;
- their expected importance for the launching users of CLARIN;
- and their capability to provide showcases that will attract more users, also from other parts of HSS.

Goals 3 years
Coherent sets of data should be in good shape and available from the launching members to researchers who want to conduct cross-lingual or cross-country research in specific areas. Areas to be determined in the second half of 2012.

Goals 5 years
Far broader range of topics, countries and languages covered, based on what members can offer.

Integration of services

Starting point
The situation is similar to the integration of data as described above. CLARIN-PP has led to a number of experimental web services that can already be offered to certain classes as users, and that can be expanded to cover a larger variety of services, located in different centres, and working on distributed data collections.

Strategy
The strategy follows to a large extent the standards action plan described under integration of data, but the focus will be on the further development of the service oriented architecture:

- interoperability;
- identification of existing services that lend themselves for application on virtual collections and for chaining and compounding;
- identification of services that can be used for showcases.

It should be noted that CLARIN ERIC will not be in a position to instruct national consortia which services they should provide, because that will normally follow from their own national research priorities. Wherever possible existing tools will be encapsulated in programs that ensure that their input and output formats are compliant with CLARIN, so that they can operate as services on CLARIN compliant resources.

Goal 3 years
A coherent set of convincing web services in place for specific research areas, including, but not restricted to linguistics. Target based on analysis of offerings through CLARIN Agreement.

Goal 5 years
Broad spectrum of services serving all areas of HSS.

Preservation

Starting point
Dedicated repositories aimed at preservation exist in some places, but not as a structural facility. Preservation of results is not normally included in research agendas of individual researchers or even of funding agencies. As a result many research results end their lives on private websites, on media that are gradually becoming obsolete, or in desk drawers. Valuable investments get lost, and research results cannot be replicated.

Strategy
As a short-term solution for every country at least one centre should be identified where local researchers can deposit research results with a view to long-term preservation and access. Such centres may or may not be located in the same country as the researchers.

They may also be shared with neighbouring data infrastructures, such as the other HSS infrastructures.

A change of culture is needed in the research community. Funding agencies should be persuaded to make proper curation and preparation of data for long term access one of the required components of project proposals.

CLARIN is actively involved in infrastructure projects such as DASISH and EUDAT, that aim at addressing common problems within HSS (DASISH) and more generally for data infrastructures (EUDAT), with a view to finding and implementing common solutions. For longer term solutions close links need to be established with emerging global data infrastructure initiatives, and CLARIN is an active participant in this process (DAITF).

Goal 3 years
All data and tools emerging from publicly funded projects in member countries can and will be deposited in repositories, and remain accessible for researchers and referable through PIDs. Solutions should be based on results of DASISH and EUDAT if feasible.

Goal 5 years
CLARIN will be collaborating with emerging European or possibly even global data infrastructures.

Ease of access

Starting point
The distribution of knowledge and expertise is very uneven. Some disciplines (e.g. computational linguistics, speech processing) are technically quite advanced or even self-sustaining, whereas in other areas penetration of digital methods is close to zero.

Strategy
A key component of the CLARIN infrastructure is the so-called knowledge sharing infrastructure, which should work along various dimensions:

- awareness actions, showcases and demonstrators for the uninitiated;
- training courses and technical support for those who want to use CLARIN;
- curriculum development for the coming generations;
- advanced support through physical or virtual centres of expertise.

A central CLARIN Portal will be set up that gives easy access to all CLARIN facilities, ranging from data and services to expertise and showcases, serving a variety of audiences with different backgrounds.

On the technical side there will be a strong emphasis on the creation of good interfaces that guide the users towards their goal and on visualization of results, both as a discovery tool of new patterns and as a way of presenting the results to a wider audience, so that the results will be better visible and can have a bigger societal impact.

Goal 3 years

- well-visited portal giving access to all CLARIN facilities;
- regular national awareness actions in all member countries;
- regular training courses and tutorials al humanities events;
- helpdesk and FAQ in place;

- five centres of expertise established;
- first curriculum plans developed.

Goal 5 years
Same, but with broader scope. Centres of expertise in place for all major areas, possibly in collaboration with other RIs

Crossing borders

Starting point
Border crossing activities are incidental, and are not part of a broader strategy, but some collaborations with parties outside CLARIN are emerging, such as network of endangered languages archives, movements towards European and global data infrastructures, joint activities with sister infrastructures in and outside Europe on HSS themes of common interest.

Strategy
A long-term vision will be developed. Existing collaborations will be taken as a bottom-up starting point, and collaboration frameworks at the EU level should be exploited to support them The fact that in most disciplines language is an important vehicle for storing and transferring knowledge and information will be used as a starting point for collaboration with other disciplines.

Goal 3 years
Initial agreements in each of the following areas

- inter-RI collaboration within Europe, both with thematic RIs and with e-Infrastructures;
- international, collaboration both at the RI and at the thematic level;
- first exploratory contacts academia-industry collaboration.

Goal 5 years
To be formulated based on the experiences and developments during the first 3 years.

Sustainability

Starting point
At this moment 9 members have joined CLARIN ERIC for in principle a period of 5 years, and more countries are expected to follow. What they all have in common is that their support for CLARIN is organised in the form of competitive calls for proposals, resulting in projects with typically a 3-5 years duration. This is not sustainable, as it will simply re-create the problem that CLARIN is supposed to solve, i.e. that access to resources is project based.

Strategy
The strategy is based on two key elements:

- first of all demonstrating CLARIN's societal impact, because without that there will be no justification for continuation; success indicators and instruments to measure them should be developed
- secondly a review of possible models that could be adopted in isolation or in combinations.

A number of alternative sustainability models will have to be explored. Possible ingredients to be considered:

- Continued funding from national sources;
- EC funding;
- Subscription or transaction fees;
- Joining forces with other RIs;
- Joining forces with institutions that are inherently sustainable;
- Joining forces with commercial parties.

Goal 3 years
After 3 years the first outline of a sustainability strategy should be formulated. This should include success measures and a plan to make them operational.

Goal 5 years
Sustainability for the following 5 years should be guaranteed.

Concluding remarks

In this strategic plan we have given an overview of the main pillars and principles on which the CLARIN infrastructure will be built.
An important task for the coming months is to make a thorough analysis of the needs for the construction of CLARIN as have emerged from the CLARIN Preparatory Phase, in combination with the offerings to CLARIN as will be reflected by the CLARIN Agreement, and the progress that has already been made in some of the countries, where work on the construction of CLARIN has already started.

A first draft will be available for discussion at a plenary meeting with the Forum of National Coordinators and the Standing Committee of Technical Centres immediately after the summer.

In order to align both the implementation plans and the minds of the teams working in the participating countries the first Annual Conference, scheduled for late 2012, will be shaped as a large, internal kick-off workshop, where all national consortia in CLARIN ERIC member countries will be represented.