CLARIN

Centre Types

2009-02-04 - Version: 5

A European Research Infrastructure

Editors: Peter Wittenburg, Martin Wynne

The ultimate objective of CLARIN is to create a European federation of existing digital repositories that include language-based data, to provide uniform access to the data, wherever it is, and to provide existing language and speech technology tools as web services to retrieve, manipulate, enhance, explore and exploit the data. The primary target audience is researchers in the humanities and social sciences and the aim is to cover all languages relevant for the user community. The objective of the current CLARIN Preparatory Phase Project (2008-2010) is to lay the technical, linguistic and organisational foundations, to provide and validate specifications for all aspects of the infrastructure (including standards, usage, IPR) and to secure sustainable support from the funding bodies in the (now 23) participating countries for the subsequent construction and exploitation phases beyond 2010.

WG2-1
CLARIN Centre Types

CLARIN-2008-1

EC FP7 project no. FP7-RI-2122230

Deliverable: D2.1 - Deadline: 1.4.2008 (postponed to 1.8.2008 due to late start)

Responsible: Peter Wittenburg

Contributing Partners: MPI, INL, OTA, RACAI, WROCUT, UPF, ELDA, ILSP, ILC, USFD, DFKI, CSC, UIL-OTS, ULeuven, ATILF, UTuebingen, HASRIL, CST, UTartu

Contributing Members: ULeipzig, CELTA, TILDE, Meertens, DANS, SBGöteborg

Scope of the Document

This document describes the categories of centres that CLARIN will define as it builds up the research infrastructure. It is based on a thorough discussion about the roles of centres in CLARIN and the requirements emerging from the various services to be offered.

This document will be discussed in the appropriate working groups and in the Executive Board. It will be subject of regular adaptations dependent on the progress in CLARIN.

This document expands the discussion of the nature and role of centres which is defined in the document *CLARIN centres.*

CLARIN References

• CLARIN Centres	CLARIN-2008-3	May 2008

Contents

1. Introduction

The CLARIN infrastructure will not be a monolithic institution offering all of the Language Resources and Technologies (LRTs) for use in the Humanities and Social Sciences. CLARIN recognizes that the creation, deployment, development and support of LRTs is not restricted to members of the consortium, or even to members of the wider network. The structure of CLARIN must allow innovation and variety in resource creation and provision, and that users have a choice of services and can select the relevant one for a particular task. The CLARIN infrastructure therefore needs to have a distributed architecture, allowing many different levels of involvement in infrastructure activities.

This document describes five classes of centres within CLARIN that are necessary to build and operate the infrastructure. This also takes into account that the transition to a full-fledged CLARIN LRT Federation will be a gradual, step by step process reaching out to the construction phase, since the funding for centres will come from various sources, and will therefore come at different times.

It is important to note that being a CLARIN 'centre' in the senses which are defined here is not the only way to be involved in CLARIN. Many vital roles in the preparatory phase and beyond will be played by centres of expertise in the creation, development and use of LRTs. Furthermore, many vital roles in surveying and research, the development of guidelines and standards, training, dissemination, administration, etc. will continue to be played by key members of the CLARIN network, who will not necessarily be service centres in the infrastructure.

Therefore, not all CLARIN member institutions will be CLARIN centres in the senses defined here. The criteria for classification as a centre at certain levels defined below are more appropriate for centres to be service-oriented institutions such as a language resources repository, a Grid computing centre or other computing service. Other activities are more easily carried out by research institutions or academies. Yet many of the most important centres of expertise in the CLARIN consortium are university departments, which are more oriented towards fixed-term research projects, and these projects and activities tend to be reliant on particular individuals. Academic departments are by no means excluded from centre status, but they will have to pay special attention to the issues of business models, funding and sustainability. Furthermore, academic departments and other types of institutions will continue to play key roles in projects, provide representatives to boards, working groups and committees, and work in the development of standards and guidelines etc.. Being a centre is by no means a precondition for CLARIN-related funding, or participation in activities, however, from a centre we expect a certain degree of commitment to deliver specified services over a long period.

The preparatory phase in particular requires the input of the individuals and institutions with the highest levels of expertise and understanding of the research practices. Many participants will play a role in developing services which will be deployed in other centres outside of their own institutions.

2 Overview

We distinguish between five classes of CLARIN centres:

- Infrastructure centres - CLARIN Type A Centres
- Service centres - CLARIN Type B Centres
- Metadata Centres - CLARIN Type C Centres
- Respected Centres - CLARIN Type R Centres
- External Centers - CLARIN Type E Centres

In a short table we can summarize the differences except for external centres:

	R	C	B	A
Online services	X	X	X	X (opt)
Services and harvestable metadata accessible via CLARIN portal		X	X	X (opt)
Fully integrated CLARIN-conformant services			X	X (opt)
Core, essential infrastructure Services with service level definitions				X

The last three levels of service need to be associated with high availability and long-term commitments, which means typically a statement of support until 2020.

During the lifetime of CLARIN centres can change their type in both directions. When commitments can no longer be given the state will change, however, this needs to be notified to the Executive Board early enough. There should always the option be used that services are transferred to another institution to keep long-term availability in the focus. In case of infrastructure services this would be a must as well as when new commitments can be made. Also the requirements for these centre types will change over time, since new insights in the pillars of an infrastructure will come up. Also in this respect centres may change their type, but formal discussions with the Executive Board are required. For the CLARIN centres we will introduce a candidate status during the preparatory phase taking care of the fact that it will take a while until the requirements are met.

The categorisation of centres should not be seen purely in evaluative terms, such that A is better than B, which is better than C, etc., and in teleological terms, such that all institutions participating in CLARIN must strive to move up the ladder towards A status. This is not necessarily desirable because the most efficient and effective way to organise the infrastructure is not likely to be with only type A centres. An ecosystem of centres at the various levels is much more likely to provide the necessary variety and flexibility that users will require, and that will be necessary for continuous evolution and enhancement.

On the other hand, we do need to build up an impressive constellation of services, and for appropriate institutions, there will be rewards in terms of prestige and allocation of resources for those who have the willingness and the ability to build effective centres as backbones of the infrastructure to be built. CLARIN needs to create an environment where success in building and enhancing centres is rewarded, but where there are also rewards for other types of activity.

The following detailed description of centre types aims to help institutions to define the type of centre that they might wish to become in the CLARIN infrastructure.

3. Detailed Description

3.1 Type R Centres / Respected Centres

We expect from all centres registered[1] in CLARIN that they offer at least resources and/or tools. For Type R centres it would be sufficient, if these resources and tools will be explained and offered via traditional web-sites. No further requirements are defined except the registration of the web-address and a contact person. However, there should be an interest to become a centre actively participating in the CLARIN LRT Federation.

3.2 Type C Centres / Metadata Providing Centres

In addition these centres offer their services via harvestable metadata and have a portal where the access ways are explained.

- Typically a human and machine readable catalogue and a schema describing the structure of the metadata descriptions is provided.
- To support harvesting either the base address for OAI PMH based harvesting or the base address for XML harvesting is provided.
- It must be possible to access the resources from the metadata description and there need to be clear statements how access to the resources can be achieved.
- If a processing component is offered, users should be able to access it via a web site.

3.3 Type B Centres / Service Centres

In addition these centres are full members of the CLARIN LRT Federation; however, they still act as individual centres not taking over responsibilities for the federation. There needs to be a clear long-term commitment for their services.

[1] At the Web-Site there will be a prominent place which shows the registered centres, the type of services they are offering and the nature of the commitment statement.

For **data resource** centres we can describe the following criteria:

- Their resources will be maintained in a well-structured and documented repository system with a long-term commitment. Versioning will guarantee that references will remain valid.
- The repository system is associated with an accepted PID[2] (persistent and unique ID) service (either in the institute or by making use of registrations at another accepted PID service site). The institute takes care that resolving the PIDs will lead to the correct object.
- The repository system will interact with a Shibboleth resource provider instance to participate in the distributed CLARIN AAI federation. For its own users it will either be member of a national/organizational identity federation or will setup its own SAML2.0 based identity provider and link it with its local authentication system.
- The centre offers various access paths to the resources amongst which there is a structured access to the complete resource for those who will need this type of unfiltered access.
- The access to the resources of the centre will comply with the license templates outlined in CLARIN.

For **processing components** offered by centres we can describe the following criteria:

- There needs to be a well-structured and documented architecture in which the services are embedded.
- The processing component needs to be augmented by web services specified by WSDL or REST descriptions indicating the available methods which are accessible via the metadata descriptions for example.
- There needs to be a clear statement about accessibility and service quality.
- The centres need to support a SAML2.0 based resource provider instance to allow users to access the services via their home identity. For its own users it will either be member of a national/organizational identity federation or will setup its own SAML2.0 based identity provider and link it with its local authentication system.
- The access to the services of the centre will comply with the license templates outlined in CLARIN.

3.3 Type A Centres / Infrastructure Centres

These centres take over responsibility for helping to manage the federation in one way or another. These centres normally extend their services described under Type B Centres; however, these centres could also be computer centres, for example, taking over some service of high relevance within CLARIN such as running a PID registration service open for all. There is a whole range of possible roles such centres can take over such as:

- running a national metadata registry, harvesting/maintaining national providers, offering a specific portal and providing a OAI PMH gateway for central harvesting
- running a (national) PID registration service
- offering large scale computing facilities to the community to execute compute-intensive tasks
- offering virtual collection services allowing to integrate collections from different sites
- offering workflow service to specify and execute chains of operations from different sites
- offering central services for the AAI federation
- etc

3.4 Type E Centres / External Centres

A further category of centre forming part of the complete infrastructure is proposed, to cover centres offering key services which are necessary to the successful operation of the CLARIN infrastructure or centres that are from different domains such as a national libraries etc. An example of the first would be a national certification authority. Numerous CLARIN centres will be reliant on these certification centres for dealing with the bulk of access, authorisation and authentication requests, but the certification authorities themselves will not be CLARIN centres.

Looking out from CLARIN more widely, there will also be related infrastructures and services which form a part of the research infrastructure environment, of which CLARIN is a part. These are categorised as Type E

[2] PIDs can be either URIs or Handles of which the persistence is guaranteed.

centres, since they may provide services of which CLARIN centres and users make use. For example, a sociolinguist may make use of social science numeric datasets accessed via the CESSDA infrastructure, and may wish to combine this data with linguistic datasets acquired via CLARIN. It is anticipated that CLARIN will work towards appropriate levels of interoperability with other infrastructures to enable successful resource discovery, access and processing, but CESSDA will remain outside of the CLARIN infrastructure.

To give another example, a repository in another continent might offer CLARIN-conformant services, such as resource discovery metadata, or text processing tools, and could thus usefully be categorised as an External Service Provider.

The E Type of centres will give us a possibility to also list them formally in the centres registry.