| | |
|---|---|
| **Title** | State of the technical infrastructure (services and centres) |
| **Version** | 3 |
| **Author(s)** | members of the SCCTC, Dieter Van Uytvanck (editor and chair of the SCCTC) |
| **Date** | 2012-10-29 |
| **Status** | For submission to GA |
| **Distribution** | SCCTC, CAC, BoD, GA |
| **ID** | CE-2012-0062 |

# 1. Introduction

This document summarizes the state of the technical infrastructure services at the level of the CLARIN ERIC's centres (section 2). As many of the national consortia are still in the starting phase it is not yet an in-depth analysis but rather an outlook on the current situation by the national representatives in the centre committee.

Because there are some generic CLARIN services available that are not directly related to the individual centres (at least not content-wise) we have included these into a separate section (3).

# 2. State of the technical infrastructure at the ERIC member level

### Austria

In the initial phase, CLARIN-AT will consist of one centre, the CLARIN CENTRE Vienna (CCV). CCV is a joint venture of the Centre of Translation Studies of the University of Vienna and the Institute of Corpus Linguistics and Text Technology of the Austrian Academy of Sciences, thus involving two major academic institutions of the country.

We are working towards establishing A+B+C centre status. We expect to achieve C-status by the end of the year. The OAI-PMH interface delivering CMDI metadata should be operational by then. According to current plans, assessment for A+B status will be possible in the course of 2013. There is data available (in particular lexicographic as well as textual corpora) and there are web services up and running. Prototypical implementations are currently being developed into consolidated services.

The process of joining the CLARIN AAI federation has been started. However, the engagement of local partners has turned out to be a somewhat lengthy process. We are also working on the accession to the CLARIN Service Provider Federation, the power of attorney – which is the first step in the process – has already been signed.

## Bulgaria

At present CLARIN-BG is officially represented by one centre of type A+B. This is the Institute of Information and Communication Technology (IICT), BAS. Since we are at the very initial phase, and are negotiating with the government about the promised local funding, the consortium is set on volunteer principles. We collaborate with representatives from the Institute of Mathematics and Informatics, BAS; New Bulgarian University; Shumen University; Tarnovo University. Their status as well as the inclusion of more members (centres) will be clarified, when the local funding is available.

STATUS at IICT:

- server is available and connected to the GRID cluster;
- collaboration is established with GRID group at IICT-BAS;
- compatibility of services with WebLicht format - the services include MorphoSyntactic Tagger, Lemmatizater, Dependency Parser;
- web search service are available for some resources (WebCLaRK - www.webclark.org).

## Czech Republic

The LINDAT-Clarin Centre (the Czech Republic only Centre) is a virtual centre to which four organizations contribute (Charles University, Universities in Brno and Pilsen and the Institute of the Czech Language in Prague). Physically, its repository is located at the Institute of Formal and Applied Linguistics at the Charles University. The Centre aspires to be type A+B (and possibly K in the future) Centre.

Funding for the LINDAT-Clarin infrastructure is provided by the Ministry of Education in the Czech Republic. Currently, it is being funded for an initial period, 2010-2015.

Technically, the Centre has a running repository based on dSpace (v. 1.8), currently with 52 corpora. There have been roughly 15,000 views of the repository in 2012, most of them to the Prague Dependency Treebank family of corpora. The repository has been complemented with a licensing database; the process of assigning licenses encourages users to use the open (Creative Commons) licenses, while allowing for legacy licenses or simply more restricted uses. It uses EduGAIN and service provider federation authentication, implements EPIC PIDs. Using these services and additions, it offers full CMDI-compliant metadata creation & data upload workflow. In addition, it contains compatibility fields in the repository metadata to be (minimally) META-SHARE compliant (and thus to allow for metadata exchange). Metadata are openly harvestable using OAI-PMH. The repository is running on a publicly accessible server in the infrastructure's "demilitarized" zone, with its own (external) SSL security certificate, data and power backup system, connected to the optical university network in the Czech Republic.

The Centre offers also free, simple (browsable) presentation of some of the corpora it hosts or their samples. It also implements several services, such as the PML-TQ treebank search interface, albeit not yet in the pure form of web service (e.g. as a Weblicht- compliant one).

The Centre is currently in the process of fulfilling the few missing features in order to be Type-B compliant and certifiable (i.e., the Federated Content Search is being prepared, as is a specific policy page with all the rules and policies the Centre is bound to adhere, services are being migrated to the full REST-compliant status). The process of getting the DSA – with features that support the rest of the Clarin requirements – has already been started.

## Denmark

At present CLARIN-DK has one centre, which applies to be assessed as type B. This centre is hosted at University of Copenhagen and currently funded until early 2017 though the national infrastructure initiative DIGHUMLAB.  More centres may be included during the funding period.

Clarin.dk has established a repository system based on Fedora and eSciDoc, with a functioning OAI-PMH provider endpoint. Metadata is currently delivered in OLAC; this will be changed to CMDI during the next months. All metadata are publicly available.

Clarin.dk is using the AAI service WAYF (National AAI service), which ensures single-sign-on. A licensing system is currently implemented using PUB, ACA and RES licenses.

A user interface to the repository is available at www.clarin.dk with metadata search and a number of other services, including a workflow planner tool and text annotation tools.

## Dutch Language Union

The Dutch-Flemish HLT Agency is currently situated at INL (Institute for Dutch Lexicology) and we make language resources available via the INL's technical CLARIN infrastructure. From January 2013 onwards, the HLT Agency will be situated at the Dutch Language Union. The new situation with regard to the technical CLARIN infrastructure is to be decided.

Due to the HLT Agency's involvement in e.g. the European DAM-LR and CLARIN projects, our work towards obtaining the Data Seal of Approval and the certainty that there will be funding for the next 5 years, we will be able to provide CLARIN services in 2013 based on a (new) technical infrastructure.

As the HLT Agency is operating on a professional basis a service desk for the linguistic resources it maintains and distributes, the HLT Agency in addition aims at becoming a CLARIN K-centre in order to make its specific expertise (including IPR and related to language policy) available for the entire CLARIN community.

## Estonia

Estonia aims to setup a type B centre at the Centre of Language Resources (CELR).

CELR is a consortium of the University of Tartu, the Institute of Cybernetics at Tallinn Technological University□and the Institute of Estonian Language, established in December 2011. The CELR centre will provide access to tools and resources by these institutions, as well as some others created in the national funding programmes. The centre infrastructure is in currently the beginning of being established, aiming to be assessed for type B in 2013.

The repository for CELR has not yet been set up. The metadata is being converted to CMDI, including some that will be converted from META-SHARE format. An OAI-PMH endpoint and formally described web services will be made available when the repository has been set up. There have been discussions on sharing DOI PIDs with other Estonian infrastructures, but the current plans are to use the handles from EPIC.

The Estonian Identity Provider Federation TAAT was established this year, using SimpleSAMLphp for authentication. So far from the CELR University of Tartu has joined TAAT as an Identity Provider, the other institutions are in progress of joining. CELR has set up a Service Provider for TAAT.

### Germany

Since its start in May 2011 the CLARIN-D consortium consists of 9 centres (foreseen type between brackets):

- BAS, LMU München; Munich; German speech and multimodal data (Type B)
- BBAW; Berlin; German language, lexica, diachronic corpora (Type B)
- IDS; Mannheim; German language, big German corpora (Type A+B)
- MPI; Nijmegen; resources of minority languages, multimedia and multimodal data (Type A+B)
- University of Tübingen; Tübingen; annotated corpora, linguistic knowledge components and web services (Type A+B)
- Saarland University; Saarbrücken; multilingual corpora and corpus tools (Type B)
- University of Hamburg; Hamburg; multilingual and sign corpora and lexica (Type B)
- University of Leipzig; Leipzig; sevices and specialized reference corpora (Type B)
- University of Stuttgart; Stuttgart; corpora and corpus tools (Type B)

7 of these have already established a repository system with a functioning OAI-PMH endpoint that delivers CMDI metadata. 6 have setup a Service Provider. 7 offer CLARIN compatible web services. 5 feature a functional Federated Content Search endpoint. 5 have issued handles for their resources.

Looking at the current state of progress it is expected that most of the centres will largely fulfil the technical B-centre requirements by the beginning of 2013. In that respect the focus will from now on shift somehow to the sustainability and assessment requirements. Going through the Data Seal of Approval procedure will be one of the main objectives for the CLARIN-D centres in the next year.

## Netherlands

Currently the CLARIN-NL consortium consists of five centres (foreseen type between brackets):

- Data Archiving Networked Services; The Hague; (Type B)
- Huygens Institute; The Hague; (Type B)
- Instituut voor Nederlandse Lexicologie; Leiden; (Type B)
- Max Planck Instituut voor Psycholinguïstiek; Nijmegen; (Type A+B)
- Meertens Instituut; Amsterdam; (Type A+B)

*Please note that the Max Planck Institute is also listed as a German CLARIN-D centre.*

All of these have already established a repository system or are setting this up. All have a functioning OAI-PMH endpoint that delivers CMDI metadata. All organizations have an Identity Provider and 4 have setup a Service Provider. Two organizations currently offer CLARIN compatible web services, the others are preparing for this. Four organizations feature a functional Federated Content Search endpoint. Four organizations are assigning persistent identifiers, three are using handles for their resources, one is using URN:NBN. The other organization is preparing to use the Handle System.

Looking at the current state of progress it is expected that most of the centres will largely fulfil the technical B-centre requirements by the beginning of 2013. In that respect the focus will from now on shift somehow to the sustainability and assessment requirements. Going through the Data Seal of Approval procedure will be one of the main objectives for the CLARIN-NL centres for the year ahead.

## Poland

LTC CLARIN-PL will initially consist of one B-type centre located at Wrocław University of Technology. Prototype of the centre is available at http://nlp.pwr.wroc.pl/clarin . Three web services are operational right now: TaKIPI-WS (morphosyntactic tagging), SuperMatrix-WS (access to corpus-based similarity between words) and plWordNet-WS (providing access to plWordNet). Web services are based on SOAP/WSDL. Metadata is harvestable by OAI-PMH. We are currently working on construction of other web services for: morhological analysis, morpho-syntactic tagging, chunking and recognition of relations between chunks, named entity recognition, recognition of semantic relations between named entities, anaphora resolution and word sense disambiguation.

We are still in the process of acquiring funding for LTC CLARIN-PL. We plan to work together with 6 academic partners on developing: system for long time data preservation, corpora, tools for advanced searching through corpora, lexical-semantic resources, shallow and deep parsing, information extraction, text summarization, and text mining applications focused on humanities and social science users.

We plan to implement all of the formal requirements of CLARIN B Centre. We also plan to establish one K centre.

# 3. State of the technical infrastructure in general

Next to those services offered at the centre level there are also some CLARIN-wide infrastructure components that have been developed during the preparatory phase or in one of the national consortia. The list below, enumerating these generic services, is intended to be a kind of a baseline measurement, rather than being an in-depth analysis.

## Production services

- OAI-PMH metadata harvester
- Metadata exploration: Virtual Language Observatory
- European Persistent Identifier Consortium (EPIC) service – API version 1
- WebLicht (resource processing)

## Beta services

- Centre registry
- Metadata creation tools:
  - Arbil
  - ProForma
- Metadata exploration: Meertens CMDI search
- Service Provider Federation, including:
  - CLARIN Identity Provider
  - Easy-to-use discovery service
- ISOcat Data Category Registry

## Alpha services

- Federated Content Search
  - validity checker
  - aggregator
- Relation Registry
- Schema Registry
- Vocabulary registry
  - OpenSkos / CLAVAS
  - SMC - Semantic Mapping Component
- Virtual Collection registry
- European Persistent Identifier Consortium (EPIC) service – API version 2
- Metadata exploration: CLARIN-AT repository search and web service