Report CLARIN centres workshop Prague

All presentations given during this workshop can be retrieved via http://www.clarin.eu/node/2931

1. Thursday, Nov. 5

1.1 Forenoon: introduction and context sketch (http://www.clarin.eu/node/2932)

Centres in Research Infrastructures - perspectives from other communities (Peter Wittenburg)
Peter gave an overview in which way centers are being discussed in other communities such as CESSDA, DELAMAN, centerNet, LHC (Large Hadron Collider), DEISA and PARADE. In all these initiatives from very different disciplines centres play a major role and are seen as the backbone of a persistent network of services. When we want to overcome the ”download-first paradigm” to a true cyberinfrastructure it is the general conviction that this will only work when we can rely on stable services offered by centres who gave commitment statements.

The user perspective (Erhard Hinrichs and Hans Bennis)
Erhard presented the requirements for centres from the perspective of the typical researcher. It is obvious that the researcher wants to include the services offered by centres in the daily and unpredictable workflows. One well-known example is to create all sorts of annotations (automatic and/or manual) which require that centres offer flexible and reliable mechanisms for workspaces, referencing etc). It is also obvious that these services must be available 24h/7d and that the interfaces must be simple to achieve an acceptance in the researcher community. Other relevant aspects are the availability of documentation and user support, the offer of training courses to allow steep learning curves and proper mechanisms for the protection of the workspaces and for acknowledging the work. The latter two are determining whether trust can be established. Centres in the traditional style with more static services will not be sufficient.

Discussion
In the discussions a number of additional points were raised:

- A priority list for centres should be made very clear (see below)
- Online tutorials would be very helpful.
- It should be possible in the CLARIN infrastructure to include services where certain payments will be required (for example to access a lexicon or to execute a tool). This was widely supported.
- At all levels there should be an indication which services and resources are free of charge and which are not, to prevent frustrations. In this respect it was stated that access to fragments of resources (sentences) can be free although the resource as a whole is not free. It was agreed that an “Academic Use” principle needs to be established in Europe.
- Guidelines for versioning of resources and tools and for the granularity of persistent identifiers are urgently needed.
- Not all countries have national identity federation plans. CLARIN needs ways to include those centres as well. This will be taken care of.
- There was a discussion whether web services can be Shibbolized. This is not directly possible. It is recommended to not invest time in this issue at this moment. Shibboleth is a solution that only works well for web browsers. Local applications that require access can use the practice to work with Short Lived Certificates (SLCS) as developed in the Grid world. For web services different solutions are proposed and a suitable choice will be worked out in

collaboration between MPI/CLARIN and the Dutch Big Grid project. It is hoped that in 2010 components can be made available to everyone.

- There was the question which kind of metadata should be provided. In the CLARIN documents it is said that (a) the LRT inventory is currently being used as an ad-hoc solution to make your resources and tools visible under the Virtual Language Observatory (but this option needs to be replaced by real metadata), (b) IMDI and OLAC are currently being accepted as metadata formats to be harvested, (c) CLARIN will adopt CMDI (Component Metadata Infrastructure) and stepwise turn over to this solution. CLARIN needs to take care that IMDI and OLAC descriptions will be converted into CMDI. New ideas such as automatic profile matching will require more detailed metadata descriptions which can be expressed with the help of CMDI.

Priorities for Centres

The list of priorities for the setting up of centres is as follows:

1. Set up a proper repository system that incorporates all resources in a structured, reproducible, accessible and persistent way which includes a solution for the granularity and versioning aspects.
2. Set up a framework for high quality metadata that makes your resources (and services) visible and that allows any service provider to harvest your metadata records.
3. Set up the system so that the centre can participate in a national identity federation and in the emerging CLARIN service provider federation.
4. Register all resources with a persistent identifier system to guarantee persistent references.
5. Take care that the centre has a long term preservation solution.
6. Prepare your setup to participate in a formal quality self-assessment procedure as specified by the Data Seal of Approval method.

1.2 Afternoon: each centre presents its current state

(terminology is explained at the end)

Name	**ATILF (LRT centre) + INIST (library repository know-how)**
Repository	busy setting up a fedora-based system (end 2009)
LTP	?
Metadata	LRT inventory (quickly), TEI headers will be integrated in fedora and from there OLAC metadata records will be generated (planned but no date yet); OAI PMH installation will come with Fedora.
Web services	-
AAI	IdP (planned in collaboration with RENATER), SP (not yet clear)
PID	plans for handles using EPIC service (no target date)
Notes	talking with publishers about long-term archiving

In general there is quite some progress at ATILF. It needs to be sorted out when an SPF connection can be scheduled. It should happen early in 2010.

Name	BBAW
Repository	Are setting up a fedora-based system (end 2009)
LTP	plans for long-term support for resources (not the services) by the academy (50 years not yet clear)
Metadata	Will generate OLAC/DC (end 2009); OAI-PMH coming with Fedora implementation is functioning
Web services	XML-RPC (for access to dictionaries), tokenizer, tagger, parser, NER; all being moved to web services
AAI	IdP: sandbox in place for access to DFN, are member of DFN AAI; they are also member of prototype SP Federation
PID	Intend to use urn:nbn (resolver: http://www.persistent-identifier.de/?link=610)
Notes	

In general there is quite some progress at BBAW and they are ready to take part of the CLARIN federation. BBAW should check whether the URN PID service provider supports the functionality described by CLARIN, otherwise it will be recommended to also register Handles for their resources.

Name	CLARIN-DK (= University of Copenhagen (coordinator), University of Aarhus, University of Southern Denmark, Copenhagen Business School, Royal Library, National Museum, Danish Language Council, Society for Danish Language and Literature)
Repository	eSciDoc, target September 2010.
LTP	
Metadata	OLAC, DC and CMDI (when ready)
Web services	A number of tools will be available as web services (e.g. tokeniser, pos-taggers, lemmatiser) for Danish and other languages. The first web services will be made accessible no later than April 2010.
AAI	WAYF as IdP, but local IdP is being considered.
PID	eSciDoc assigns UUID's to all resources by ingest, but as for PIDs there are currently no plans.
Notes	

In general there is quite some progress at CLARIN DK. The remaining question is when CDK is ready to participate in the CLARIN federation.

Name	CSC, U Helsinki
Repository	All resources are accessible via organized file system,
LTP	
Metadata	metadata is web accessible, will provide IMDI/CMDI for the Multi-lingual Research Collection, OLAC will be made available for the Finnish language bank resources; OAI PMH has been set up thus all resource metadata should be harvestable
Web services	-
AAI	IdP has been set up within HAKA: SP component has been installed
PID	Plan to regsiter Handles (via EPIC)
Notes	

In general there is quite some progress at CSC and they are ready to take part of the CLARIN federation.

Name	DANS
Repository	Easy (own repository system), are working on a migration to fedora
LTP	Have a longterm preservation strategy
Metadata	DIDL, DC, all harvestable (http://easy.dans.knaw.nl/oai?verb=Identify)
Web services	-
AAI	IdP is ready via SURFnet; SP has been installed
PID	Will use urn:nbn (resolver: http://persistent-identifier.nl)
Notes	their repository and archive is subject of quality assessments

In general there is quite some progress at DANS and they are ready to take part of the CLARIN federation.

Name	HASRIL
Repository	currently reorganizing the whole repository which will take some time
LTP	
Metadata	Yet not ready to deliver metadata due to fragmentation of resources
Web services	

AAI	IdP integration is planned, SP set up is planned
PID	A URN prototype is offered from the library
Notes	

With respect to the setup of a repository system there is good progress although it is not yet evident when things will have been setup. Other steps will follow then. With respect to the URN system it should be checked in how far it fulfills the functional requirements of CLARIN.

Name	**IDS**
Repository	An own system called COSMAS is being used for some time already.
LTP	Own LTP-strategy. IDS is now a member of nestor. (http://www.langzeitarchivierung.de/)
Metadata	OLAC generation almost ready (target: end 2009); OAI PMH setup also being finished.
Web services	Allow already now building virtual collections of own resources
AAI	IdP setup is ready; a Shibbolization of COSMAS is planned for February 2010 which will allow IDS to participate in an SP federation; an installation of the Globus Toolkit is planned to participate in the German Grid system
PID	Plan to use Handles
Notes	For the installation of Shibboleth IDS is using the service of the company 'DASII International' in Tübingen.

In general there is quite some progress at IDS and they are ready to take part of the CLARIN federation. The installation of GTK will not help directly at EU scale due to the different middleware stacks being used.

Name	**ILC**
Repository	currently ILC is reorganizing its whole repository, not yet clear which system will be taken
LTP	
Metadata	ILC plans to create CMDI components for lexica and offer their metadata records as CMDI; scheduling of OAI PMH is dependent on repository system
Web services	Are working on web services to allow access to their lexica; want to Shibbolize web services
AAI	IdP is almost ready, SP setup is planned
PID	No plans yet

Notes	

With respect to the setup of a repository system there is good progress, however which software will be used and it is not yet exactly clear when the transition will be finished. Metadata deliverable will be independent and can thus happen earlier. It is not fully clear when the SP component setup can be realized.

Name	ILSP
Repository	ILSP has a repository system running. Currently also installed and experimenting with DSpace 1.5.2.
LTP	
Metadata	ILSP plans to use component-based metadata (CMDI), all XCES descriptions will be moved to CMDI components in early 2010; Through DSpace OAI PMH with DC metadata is also supported.
Web services	Have implemented tool chaining via UIMA methods;
AAI	Shibboleth is supported but needs configuration.
PID	No plans yet
Notes	awaiting the results for the national CLARIN proposal

ILSP seems to be ready to provide HQ metadata very soon. It is not yet clear when the AAI setup is scheduled. Much will depend on national CLARIN funds.

Name	INL
Repository	INL is reorganizing its repository system
LTP	
Metadata	IMDI (currently off-line), component metadata will be supported; the OAI PMH is supported
Web services	-
AAI	IdP and SP will be set up end 2009; need to go via U Leiden
PID	Have already used handles (currently off-line)
Notes	

In general there is quite some progress at INL and they are almost ready to take part of the CLARIN federation.

Name	Latvia

Repository	The LRT centre is busy with reorganizations; the D-Space software will be deployed in Q1 of 2010
LTP	
Metadata	everything is available in the LRT inventory ; have some data associated with TEI headers which can easily be transformed; plans for OLAC and OAI PMH in begin 2010, (if national project is supported)
Web services	- Experimental REST service for word morphological analysis. - Text-to-speech SOAP service.
AAI	Don't have a national identity federation yet; SP is available now via SimpleSAMLphp
PID	Some plans, related to the setup of a repository
Notes	There is a very modern datacenter in the institute, with a cloud computing infrastructure ready to be launched as a commercial service.

It seems that U Latvia will be ready soon to offer metadata. It should also be possible to integrate them in a test into the CLARIN federation. Other transitions will take more time.

Name	**Leipzig**
Repository	Are using an own system based on relational databases
LTP	
Metadata	They plan to provide IMDI and CMDI based metadata and setup of OAI PMH should be simple for them
Web services	Offer many web services and want to describe them by CMDI
AAI	IdP setup is planned; a SP component is not necessary since their services and resource fragments can be accessed freely
PID	No plans for PIDs yet
Notes	Main focus on web services

U Leipzig seems to be able to provide CMDI metadata records soon and to participate in federation plans.

Name	**Lund**
Repository	Are using IMDI+LAMUS which working since several years
LTP	
Metadata	They generate IMDI and the OAI PMH protocol is supported

Web services	
AAI	IdP setup is finished; a SP component has been installed and only needs to be upgraded
PID	Handles are being registered
Notes	

U Lund is ready to join the CLARIN federation.

Name	**Meertens Institute**
Repository	Are using an own system based on relational databases;
LTP	
Metadata	Already provide some IMDI, will provide records using component metadata; OAI PMH will be setup in early 2010
Web services	-
AAI	IdP setup is ready; SP setup does not seem to be necessary (no closed resources)
PID	Plan to use Handles
Notes	

MI seems to be ready to provide high quality metadata soon and it should be possible to integrate them soon in federations.

Name	**MPI**
Repository	Use an own system for several years based on IMDI and LAMUS which makes use of a file system as primary storage (http://corpus1.mpi.nl/)
LTP	MPI has an open deposit policy and all resources are copied to four other centers, i.e. a long-term archiving strategy is in place.
Metadata	All metadata is provided as IMDI records, OLAC is being generated and the OAI PMH has been implemented (data provider and service provider side), all will be transferred to CMDI when this has been stabilized
Web services	several web services have been implemented (metadata search, lexicon access and visualization, resource access and visualization, ISOcat,...)
AAI	IdP and SP setup have been finished
PID	All objects are associated with handles, a handle service has been installed

Notes	

MPI is ready in all respects.

Name	Oxford Text Archive
Repository	Use an own system based on TEI concepts
LTP	yes
Metadata	Internally they use TEI headers, but they are creating OLAC records; OAI PMH has been installed; OTA can also create CMDI records
Web services	-
AAI	planned for end summer 2010
PID	persistent URLs for metadata records, plans for PID for the resources themselves
Notes	

OTA is obviously ready to provide metadata and can quickly join the CLARIN federation. Some important features will be available at the end of the summer of 2010.

Name	Charles University Prague
Repository	CU currently uses a mixture of file system, SVN and relational databases, but can move to D-Space or ePrints – both have been tested already
LTP	
Metadata	Can easily generate all kinds of metadata from internal representations; almost ready for harvesting via OAI PMH
Web services	-
AAI	IdP setup is finished; SP setup will depend on the choice of the repository system
PID	Use handles since they are supported in D-Space
Notes	need answers to questions on granularity PIDs

CU is busy to restructure their holding to make it fit for the infrastructure; it seems that they can easily setup things. It would be needed to know from CU when exactly they can provide metadata according to one of the recommended paths and when they can participate in the CLARIN federation.

Name	Sheffield
Repository	No repository system is necessary – only web services will be offered

LTP	
Metadata	waiting for web service metadata specification for web services and will create them according to the CMDI standard; an OAI PMH instance will be installed
Web services	Several web services are offered
AAI	A SP component is not necessary since all services will be free; Sheffield U is part of the UK AAI federation
PID	Yet no plans
Notes	

Since Sheffield U is only providing free services they can easily participate. Metadata will be offered soon.

Name	**Språkbanken**
Repository	several systems are in use; currently much restructuring and harmonization takes place; much is based on TEI P5
LTP	
Metadata	OLAC will be extracted from TEI first, OAI PMH will be set up; all should be ready end of 2009
Web services	There are web applications giving access to the data
AAI	low priority since resources are either completely open or completely closed
PID	Are very much interested in PIDs; tend to use handles
Notes	need recommendations on PID fragment identifiers; SB has a deposit service

It seems that SB can easily participate as centre.

Name	**Tartu**
Repository	They are using reorganizing their repository and are talking with the university about a D-Space installation; transition is planned for 2010; it is planned to move all data to the computer centre.
LTP	
Metadata	TEI headers are used; end 2010 metadata should be available when CMDI components are finished
Web services	Corpora are accessible via web applications

AAI	There are plans for the use of SimpleSAMLphp to install an IdP and an SP
PID	Intend to use handles that are issued by D-Space
Notes	

It seems that U Tartu will be able to offer metadata and join the CLARIN federation in 2010.

Name	**Tübingen**
Repository	eSciDoc will be used as repository system
LTP	
Metadata	OLAC records will be offered end of 2009; OAI PMH will also be setup in January
Web services	Several together with U Stuttgart, see D-SPIN weblicht pilot
AAI	IdP setup is finished; SP setup is planned for end of 2009
PID	urn and handles (used by eSciDoc) will be used
Notes	

It seems that U Tübingen will be ready in all respects in January 2010.

Name	**Vienna**
Repository	Use a own system (PHAIDRA) which is fedora-based, the transition is still ongoing and will be finished in summer 2010; see https://phaidra.univie.ac.at/
LTP	
Metadata	OLAC metadata will be extracted and it is planned for summer 2010
Web services	-
AAI	Plans but not yet concrete
PID	fedora-based actionable PIDs based on Handles
Notes	

It seems that U Vienna will need until mid 2010 to be integrated in the CLARIN federation.

Name	**Wroclaw**
Repository	Use own system
LTP	

Metadata	Will create CMDI based metadata for web services; should be ready soon; setting up OAI PMH has been finished
Web services	Offer several web services
AAI	AAI is not necessary unless hosting of third-party resources will be done as Polish centre; in Poland there is no national IdP
PID	Will use the EPIC handle service if necessary
Notes	

It seems that U Wroclaw can easily be integrated at the metadata and service level.

2. Friday, Nov. 6

2.1 Forenoon

Discussion about the requirements short guide – see the new version of the document.
Persistent Identifier API: http://www.clarin.eu/node/2935
Authentication and Authorization: http://www.clarin.eu/node/2933
Component Metadata and OAI-PMH: http://www.clarin.eu/node/2933

Centres Requirements short guide – the main points that came out of the discussion were:

- The need for clear formulations and explicit statements (e.g. for the timing, for infrastructure requirements)
- Have good references to existing documents (e.g. the one describing the centre types)
- Slightly weaker minimum requirements for the migration of data and services when a centre stops.
- A clear versioning policy (also for web services) needs to be made available for each centre.
- An explicit statement that the aim is to support researchers will be added to the short guide.

2.2 Afternoon

The legal set-up of the CLARIN Service Provider (Start-up) Federation:
http://www.clarin.eu/node/2933

2.3 Summary

At the end Peter gave a short summary of the meeting. He stressed that

- about 17 centres will obviously be ready to participate in the first month of 2010
- about 9 will need more time to set up the relevant components
- a short report will be sent to every centre to continue the interaction'
- a wiki forum will be setup with color codes to indicate the state of the work
- video conferences will be used in 2010 to continue the direct discussions
- hopefully reports will be created (INL has one) about how to do certain installations
- CLARIN is ready to organize follow up training courses if this is necessary
- CLARIN would look for experts that can help in setting up things

Further, he mentioned that

- the Virtual Language Observatory will be maintained and used to show the harvested resources and tools
- everyone should check if all their resources and services are visible
- everyone is welcome to participate in the CLARIN use case where a distributed search will be realized
- all interested experts will be asked to contribute to a recommendation document about the granularity and versioning issues

Terminology

Repository

The entry "repository" indicates whether one can speak about a structured and persistent organization of the data allowing people to access the resources that are identified by metadata and persistent identifiers. This should be documented by mentioning a software system that supports the holding. Many different software solutions are being used currently. A versioning policy is expected to be supported by the repository system.

LTP

The entry "Long-term Preservation" indicates whether the centre has already a strategy of how to preserve their data, i.e. replicating data at different locations, checking authenticity etc.

AAI

The entry "AAI" (Authentication and Authorization Infrastructure) indicates the installation and integration of components such as Shibboleth to allow the centre to participate in a domain of trust. This basically has three aspects: (1) signing an agreement to be member of a trust federation; (2) setting up an identity provider component to become part of an identity federation; (3) setting up of a service provider component to become part of the CLARIN service provider federation.

Metadata

The entry "Metadata" indicates the persistent availability of well-structured metadata records describing the stored resources according to the CLARIN requirements (CLARIN Metadata Infrastructure) and the availability of a service allowing metadata harvesting which in general is done by providing an OAI PMH port.

PID

The entry "PID" indicates the usage of PIDs to uniquely and persistently identify all kinds of LRT objects in the repository. The final task must be to set up the repository in a way that if a user selects a PID the repository will give exactly the resource back that is meant.

Web Services

Some centers will primarily not store data resources but offer services to the community. The entry will indicate this. Yet there are not as clear specifications of how to do this. They should be described by metadata as well and in future we foresee that also services should be identified by PIDs.