Minutes of the first CLARIN-DE and DARIAH-DE

Technical Advisory Board Meeting - Berlin, 10.04.2013

Participants:

Host

Wolf-Hagen Krauth (excused, BBAW)

Technical Advisory Board

Jonas Beskow (KTH Stockholm)
Jan Hajic (Charles University, Prague)
Eduard Hovy (CMU)
Michael Lautenschlager (DKRZ)
Toma Tasovac (Digital Humanities Center Belgrade)
Claire Warwick (University College London)

CLARIN-D

Sebastian Drude (MPI Nijmegen)
Erhard Hinrichs (University Tübingen)
Thomas Zastrow (University Tübingen)
Dieter van Uytvanck (MPI Nijmegen)
Gerhard Heyer (University Leipzig)

DARIAH-DE

Peter Gietz (DAASI)
Daniel Kurzawe (GWDG/SUB)
Heike Neuroth (SUB)
Stefan Schmunk (SUB)

Computing Centre (partner of both projects)

Johannes Reetz (RZG)
Ulrich Schwardmann (GWDG)
Daniel Mallmann (FZJ)

Summary:

Introduction and function of the TAB (Sebastian Drude, Heike Neuroth, Erhard Hinrichs)

Since the Chair Wolf-Hagen Krauth (BBAW) could not attend, Heike Neuroth (SUB) welcomed the TAB members and projects on behalf of the BBAW.
Erhard Hinrichs gave a short introduction about the function of the TAB: The main goal of the meeting was to inform the TAB about the progress of the two projects and to get some advice

for future development. Therefore, both projects presented their technical infrastructures and pointed out the similarities and differences.

The CLARIN-D Infrastructure (Presentation: Dieter van Uytvanck)

Dieter van Uytvanck gave an introduction to CLARIN-D with respect to the technical aspects of the CLARIN-D infrastructure.
One of the major goals of the technical infrastructure in CLARIN-D is to make resources and tools accessible for a broader community in a simple, user friendly and transparent way via web applications and Service Oriented Architectures.
The TAB asked about the type of resources. Online resources are the focus of CLARIN. They could be accessible via the Internet or other ways (e.g downloadable software for installation, etc.). For enhancing interoperability, CLARIN is using the defined standards and protocols, which are recommended to the users. Metadata is mapped to IsoCAT. Centres are not monolithic, but could be specialised. In Germany there are nine CLARIN centres at the moment. All are "Class A+B", which means that they provide access to their own resources and tools as well as providing services for the whole infrastructure. There are fixed criteria for becoming a CLARIN centre to ensure their sustainability.
Each centre hosts a repository, which stores language resources and tools. There are also recommendations for the usage of standards in terms of dataformats and schemata. A Federated Content Search (FCS) allows the user to search directly in the textcorpora of all nine CLARIN-D centers from one common user interface.
The TAB asked if you can also access the data in other centres from all user sites. It is possible, but at the moment just for open access data or data which is free available inside the academic community. In the future there will be a solution for limited resources. But this depends on future developments in the AAI.
As one metadata schema will not be sufficient for the needs of a specific community, CLARIN supports the CMDI metadata framework with links to IsoCAT. All centres provide access via an OAI-PMH interface and are searchable via the Virtual Language Observatory (VLO).
The TAB asked about the complexity and realisation of the metadata harvester. There were also comments about the complexity and the future developments of metadata schemes. As metadata definitions may change in the future, a monolithic system has to adapt accordingly. This should be taken into consideration in future development.
EPIC PIDs are used in CLARIN. NAGIOS is the solution for monitoring the components and services. WebLicht is used as SOA, and web services can be integrated into other web applications.

The DARIAH-DE Infrastructure (Presentation: Stefan Schmunk)

Stefan Schmunk presented the DARIAH-DE project with respect to its technical infrastructure. He also gave some examples of its usage. DARIAH-DE is focused on text, data and source material in numerous disciplines in the area of the arts and humanities. There are five work packages. The first three (E-Infrastructure, Research and Teaching, Research Data) represent the primary areas of tasks in DARIAH-DE, while the other two work packages (Consortium Management, Participation in DARIAH-EU) are responsible for the project organisation and the contribution to the European DARIAH project DARIAH-EU.

After an overview of the work packages he presented the details of the technical infrastructure. The DARIAH-DE portal is the central access point to DARIAH-DE. It is reachable via browser and is integrated into the AAI system (shibboleth). In the portal, the following are represented: teaching activities within DARIAH-DE, curricula, research and scholarly data, and also information and services for the technical infrastructure.
It is possible to access four different kinds of hosting services from generic virtual machines (VMs) to pre-configured VMs and services. There are also higher-level services that are integrated directly into the liferay portal system. Some of them are demonstrators for functionalities of the DARIAH components and their application in the arts and humanities.
The TAB asked about the decision process about the selection of the demonstrators and who pays for the integration and hosting. At the moment, most of the demonstrators are developed by affiliated projects or DARIAH-DE itself.
There are also generic services for accessing and working with content data: the Schema Registry and the Collection Registry. Basically, the Collection Registry provides a high quality database of accessible collections. The Schema Registry supports the mapping between different metadata schemata. As opposed to the CLARIN-D service, the Collection Registry maps between two selectable schemata.
Here the TAB asked about scalability of an m:n mapping in a service like the schema registry. So far, neither performance or mapping problems occur in the schema registry.
The TAB also asked if there is any interest from the humanities in building up ontologies based on the schema registry, and shouldn't they be encouraged to use existing standards? The reason for the schema registry that there are a lot of different standards in different interpretations. It is necessary to give the researchers a tool to access all these data. DARIAH-DE is also using the EPIC PID Service and is working together with partners in Europe in DARIAH-EU.

Common structures in DARIAH and CLARIN (Presentation: Johannes Reetz)

Johannes Reetz presented the similarities between the two projects from the computer centre's point of view. He pointed out the common technologies and some differences in the two infrastructure projects. All of the computer centres presented are active in both projects, but the role differs between the projects. In CLARIN-D, the computer centres are resource providers in terms of hardware, basic services and service hosting. They also support generic service hosting. In DARIAH-DE, the computer centres are more involved in the project. The AAI structure has a lot of communalities and both projects are using EPIC PIDs. There are also a lot of similarities regarding the authorisation and authentications: both use an SSO Federation via SAML and Shibboleth with unique personal identifiers from the identity providers (IdPs). But DARIAH-DE has a central instance for managing privilege groups and the IDs are used in different way, while authorization in CLARIN takes place locally at each centre.
Both projects provide a storage federation: DARIAH-DE uses an iRODS federation and defines its own DARIAH Storage API for handling objects, while CLARIN-D is using the EUDAT Storage Service.
CLARIN-D uses the centre registry for monitoring, while in DARIAH-DE the structure is configurable. Both monitoring systems are based on NAGIOS.

Final discussion

The TAB gave recommendations about the openness of the infrastructures: open and well documented services make it easy for users to contribute and integrate their work. But abuse could be an issue in an open infrastructure.
While DARIAH-DE is more user oriented, CLARIN-D focuses on the stability of the infrastructure. The ideal situation is somewhere between these two approaches.
The TAB asked about the scaling of the projects: If the projects are scaling up in size, who will pay for the scaling in support and infrastructure? DARIAH answers that ESFRI requires the national ministries to commit to this responsibility when they join.
The situation of the bit preservation in the two projects is not clear. There should be a common working group for the representation and common concepts regarding long term preservation.
At the moment, data sets are not a scientific outcome, but they are important in the research process. We need something that includes datasets as a scientific outcome.
The TAB asked about the level of collaboration between CLARIN-D and DARIAH-DE in Germany in contrast to Europe. DARIAH-DE answered that there is not such a strong collaboration in all of the EU countries. Both projects are recommending this for the EU partners.
The TAB remarks that the actual status of both projects is not clear at all. In the next meeting we should get more into the actual progress.
Multilingual information material is important to get all the people involved who don't speak English.
A clear statement about the applicability of the infrastructure and services and its support in founding could help to promote the projects. Unification could help to reach bigger communities.
From a technical point of view, the identity management is important: eduGAIN should be supported in both projects in future; it is a simple solution that is starting to become available.

Action Points

- Create TAB mailing-list (TLA)
- I-AG meetings in fall 2013
- Make overview of communalities and low hanging fruits
  For instance, some low hanging fruits were mentioned:
- AAI mutually include the others
- Controlled Vocabulary Services
- Unified / mutual access to content search?

Next Tab meeting

The next TAB meeting is scheduled to take place at the DH 6-12 July in Lausanne, on 7th July in the afternoon.

Further Information:

https://dev2.dariah.eu/wiki/x/95Hf