GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

2nd CLARIN-D/DARIAH-DE Technical Advisory Board Meeting Mannheim, February, 17 2015

Meeting Metadata

Location:

IDS Mannheim
R 5, 6-13
D-68161 Mannheim

Room: “Großer Vortragssaal” (auditorium), ground floor

Date and Time:

Date: 2015-02-17

Time: 11:00 – 16:30

Participants:

Technical Advisory Board:

- Jonas Beskow (KTH Stockholm)
- Jan Hajic (Charles University, Prague)
- Eduard Hovy (CMU)
- Michael Lautenschlager (DKRZ)
- Gerhard Schneider (Universität Freiburg, Rechenzentrum)
- Toma Tasovac (Digital Humanities Center Belgrade)
- Claire Warwick (University College London) via Skype

excused: Melissa Terras (UCL), Carol Goble (University of Manchester)

From the infrastructures:

CLARIN-D:

- Dirk Goldhahn (University of Leipzig)
- Erhard Hinrichs (University of Tübingen)
- Oliver Schonefeld (IDS Mannheim)
- Thorsten Trippel (University of Tübingen)
- Dieter van Uytvanck (MPI Nijmegen)

DARIAH-DE:

- Mirjam Blümm (SUB Göttingen)

- Peter Gietz (DAASI Tübingen)
- Andreas Henrich (Uni Bamberg)
- Heiko Hütter (DAASI Tübingen)
- Stefan Schmunk (SUB Göttingen)

Computing Centres:

- Daniel Mallmann (Forschungszentrum Jülich)
- Ulrich Schwardmann (GWDG, Göttingen)
- Thomas Zastrow (Rechenzentrum Garching)

Agenda items

Welcome

Ludwig Eichinger, head of the IDS Mannheim, welcomes the participants.

Role call of participants

The participants introduced themselves and their affiliation.

Welcoming the Technical Advisory Board from the projects and introducing CLARIN and DARIAH

The participating groups were introduced.

The Technical Advisory Board was introduced as an advising group for consulting on the collaboration between CLARIN-D and DARIAH-DE on the technical level.

The Digital Research Infrastructure DARIAH-DE was introduced by Mirjam Blümm by explaining the structure and the achievements since the last TAB meeting. The question of the dissemination of the infrastructure was answered by showing an overview of the cooperation projects (about 40) and the figure with user statistics, which shows the constant increase of infrastructure users with many services used productively for several years. Dissemination is also ensured by presentations on conferences, regular publications and community engagement activities e.g. by a stakeholder board and members of the consortium.

With CLARIN-D the hosting infrastructure was introduced by Erhard Hinrichs by explaining the structure and the achievements since the last TAB meeting. The question on the commitment of depositing data was answered by stating that the commitment varies by institution, ranging from 10 years at the universities to 50 years as bitstream preservation. The IDS has a longer scale commitment without fixed timeframes. It was seen as problematic to give a long time commitment if the heads of the institutions are not involved, but this concern was resolved as all participating institutions as such are also showing their mutual interest by financial commitments. Questions on the transition from the pilot implementation phase to the usage phase were also postponed.

The computing centres and their role both for CLARIN-D and DARIAH-DE were presented by Ulrich Schwardmann. The centres provide essential technology solutions such as an AAI infrastructure, and monitoring solutions despite of structural difference. DARIAH-DE for example has central authorization policies while in CLARIN-D user-management is connected to the home institutions of the users, so it is distributed in CLARIN-D.

The TAB members pointed out that CLARIN-D and DARIAH-DE show synergies on the low level and middle layer of technologies, while on the higher level DARIAH-DE includes the computing centres in the core technical developments, while CLARIN-D utilizes them as service providers. One reason seen was that CLARIN-D centres have another layer of institutional support, for example in the computing centres of their institutions, while DARIAH-DE sees the need to include domain experts both from the research projects as well as the computing centres to use the research infrastructure at its full potential. Following up on the advantages and disadvantages of decoupling research infrastructures and computing centres it was stated that there is justification for both approaches for different purposes. For example large parts of the digital humanities have low computing requirements but are data driven, causing additional requirements based on restrictions of data use. Also, long term archiving becomes an issue for collaboration. Hence computing centres are contracted as service providers with a contracting template with Service Level Agreements but at the same time the centres are integrated by a link between services provided and required from the infrastructures.

Common infrastructural components of CLARIN-D and DARIAH-DE

The common infrastructural components of CLARIN-D and DARIAH-DE were introduced by Peter Gietz and Dirk Goldhahn. Among the common components are: AAI, PID, monitoring, storage, metadata harvesting, search functionality in the content of the resource. It was stated that in DARIAH-DE, the monitoring often showed surprising usages of services and unexpected allocations of user interest, which showed a high demand for a more detailed monitoring and statistics. Therefore, a structure was created in DARIAH-DE which researches the critical success factors for VREs.

Peter Gietz additionally demonstrated the sustainability concept of DARIAH-DE with the DARIAH-DE eHumanities Service Unit (DeISU). It was stated by the board that this concept could be very beneficial as user demand and resource provisioning is strongly coupled. Monitoring of user statistics is an important aspect of the developing process of the common research infrastructural components. For that e.g. DARIAH-DE started to monitor the access for some of the technical components (e.g. Wiki, Etherpad, several scientific tools etc.) and will increase the efforts to evaluate more specific and detailed their usage.

General discussion of the Technical Advisory Board and the consortium

Following the general overviews, the Technical Advisory Board members used the opportunity to ask questions to which the consortia responded.

Q: At the last meeting of the TAB various action items were discussed, such as common services by CLARIN-D and DARIAH-DE. Are there any results?

A: The computing centres' involvement helps to avoid having independent services for each project. Some progress has been made in this respect. The scope is on generic services that are prioritised for collaboration, more specific services with regards to specific applications may follow.

It was suggested to provide community based services based on research questions, while general services such as depositing services, authentication and authorisation could be continued to be generalized.

One recommendation in the area of AAI is to collaborate with edugain. CLARIN-D explained that the users of DARIAH-DE not having access to the infrastructures via their home institutions, referred to as homeless users, can use CLARIN by using the DARIAH-DE IdP. DARIAH-DE made clear that on the technical level its AAI is integrated with eduGain. The major problem lies within the organizational level, i.e. the attribute release policy of the IdPs that prevent sending of needed identifiers to the SPs. With the operation of a DARIAH IdP for homeless users this issue can be circumvented. The experience of CLARIN with eduGain shows that an opt-out policy integration per country would be beneficial for providing access to the infrastructures rather then current institution-based opt-in solutions.

CLARIN and DARIAH increased their cooperation, for example in the integration of the Data Protection Code of Conduct in the TERENA (now GÉANT) context. The Service Providers in DARIAH and CLARIN support the CoC. DARIAH-DE already became member of the DFN AAI, CLARIN-D is working with the Identity Providers so that they agree to the Code of Conduct and join eduGain.

The cooperation with eduGain covers countries currently not contributing to one of the infrastructures. Consequently the question arises which services should be provided for those countries that participate in eduGain, but are not contributing to one of the European Infrastructure Consortia.

The TAB members stressed that after the construction phases of CLARIN and DARIAH, the tools should have sufficient impact and the users should lobby for paying for the infrastructure. The consortia explained the option of having a difference between value added services for paying participants and possible clashes with Open Access. It was expressed that fee based models are seen sceptically in the humanities, but that providing privileged access for those who support the infrastructure may be an option. The TAB members pointed out that open access to the basic services and data may be also appropriate for the general public. Another option introduced by the consortia is that scholars may bring their own data services or their own data. This results in a distinction between open data options and open services, depending on restrictions of material and infrastructure.

Q: The quality of the technical infrastructure and access options are aspects of the use of the infrastructure, but how can the infrastructures reach out to the community?

A: The consortia explained that both in DARIAH-DE and CLARIN-D there are various outreach activities, starting with teaching, providing teaching materials, participating in summer schools, organizing discipline specific working groups, etc. Another level of outreach is also established by cooperation with other ESFRI projects (CESSDA, etc.), and projects on the European level (DASISH, PARTHENOS, etc.). Contacts with national groups that have a different scope and timeline often provide options for an exchange of technology. An example of such an institution is GESIS.

For individual researchers, user scenarios, FAQs, training materials and web applications are being prepared. The question for the procedure after the end of funding of the current implementation phase of the infrastructures was discussed with the result that the goal is to convince the scholars of the usefulness of the infrastructures and good communication strategies for example by improved websites. It was noted that the quality of the backends has already been established, hence in the remaining part of the implementation phase, the frontend should be more in focus, addressing questions of expected tools and killer applications. It was discussed that higher level tools and applications such as PID services have been in focus and that the integration of existing tools is currently on the way. The maintenance of the tools requires programmers as well. An example of the integrated applications are exploratory tools (such as the CLARIN VLO), the content based search functions based on SRU/CQL extensions (e.g. CLARIN FCS), virtual data collections and their integration with web-processing pipelines. Smart analysis software and large national corpora would be applications for users.

As outreach and impact in the community was out of the primary scope of the TAB, this question was not further elaborated.

Q: A new problem appears if somebody reuses existing data and adds additional levels of analysis. What happens to added information if somebody works on data that exists and wants to enhance the data?

A: The discussion shows that the standard procedure is to use versioning of data, in the case of annotations to add additional annotations levels with a PID as reference. A principle problem to be solved then is if the annotations can be added back in a repository as various property rights might be affected. Additionally questions of quality assurance would have to be addressed. As an example the BRAT-fork WebAnno was given, which allows adding additional annotation levels by users. It is unclear what should happen with the results of such an added annotation process.

Q: The infrastructures currently are based on project funding. Which services would have to be discontinued if funding is reduced and certain services and maintenance work such as quality assurance cannot be funded any longer?

A: The discussion shows that a business model needs to be implemented to support the infrastructure. As an example the DARIAH-DE e-Humanities Infrastructure Service-Unit (DeISU) was used, which defines service units with fixed cost structures and service levels. Funding organizations such as the BMBF need to take into account the requirements of priorities and service levels. Current discussion points to the direction that a third of the cost is in archiving, while two thirds of the costs are in the retrieval of the data.

Q: A lot of infrastructure components are available for free, such as unix. Is too much of the systems rebuilt in the infrastructure projects?

A: The infrastructures reuse existing tools and services as they are available. Though it was agreed that OpenSource is not always the better answer, with possible large initial investments, a tendency to use open source is apparent. Funding agencies paying for open access journals may also be willing to pay for releasing data and software. But sometimes the requirements are different. For example for language material LDC and ELRA use PIDs for citation only, while research infrastructures also require PIDs as an actionable item for toolchains in the "data fabric". Hence CLARIN-D and DARIAH-DE use the handle system as it is the most flexible and modern system. A redirect or alias to map PID systems onto each other is also possible.

Another system that could be a candidate to serve in a research infrastructure is for example an ownCloud system for personal workspaces. This is seen as a temporary storage solution with unclear quality and not guaranteed long term preservation. Workflows of research infrastructures need to also provide for long term maintenance. This is achieved by a cooperation with projects such as EUDAT and community based backends such as openSKOS to replace the custom build ISOcat infrastructure. CLARIN-D and DARIAH-DE use as much off the shelf software with large community uptake as possible and add components as required for the application contexts.

Closing remarks

Mirjam Blümm and Erhard Hinrichs for DARIAH-DE and CLARIN-D, respectively, thank the participants for the fruitful discussion.

The projects express their appreciation for the services provided by the hosting institution.