WP2: Voortgangsverslag

Datum: 2015-11-18

Auteur: Gertjan Filarski

Wat doet WP2?

In WP2 wordt de overkoepelende infrastructuur voor CLARIAH gebouwd. Deze infrastructuur verbindt linked-open-datasets uit WP3, 4 en 5. De focus ligt op 'wie', 'wat' en 'waar' data (personen, concepten en locaties) met identificerende kenmerken.
Zo bouwt WP2 bijvoorbeeld uit de onderzoeksdata van de drie domeinen een generieke personen dataset met lijsten van namen, geboorte & sterfdatum en plaats, en geslacht. De datasets in WP2 worden in RDF aangeboden en gekoppeld aan externe datasets als dbpedia.nl, Europeana en de erfgoedsector. Geesteswetenschappelijke onderzoekssoftware kan naar deze data linken en op die manier toegang krijgen tot de achterliggende specifieke data uit het onderzoeksdomein.

Progress Report CLARIAH WP 2: Technology

1. Organization & communication

1.1 Overview

In 2015Q1 the CLARIAH Technical Plan was finalized. This high-level document details the activities (tasks), budgets and partners that make up WP2. The current status report will follow the general layout of the Technical Plan closely. The partner institutes wrote detailed project proposals for the various tasks during 2015Q2-3. Many facilities also used this half-year to recruit qualified IT engineers. Development on most tasks only commenced in 2015Q4. Although this start is roughly six months later than expected the consequences are limited and under control. The engineering capacity in 2015 is in most instances added to 2016 and – to a lesser extend – 2017 in order to deliver the core and user interaction infrastructure as expected early 2018.

1.2 Data & Infra Committees

WP2 has a bimonthly meeting of two committees. A team of data specialists and scholars discusses modelling and curation tasks in the Data Committee, while the engineers participating in the Infrastructural Committee report technical and infrastructural tasks. Before the summer of 2015 the committees came to the joint conclusion that many of the discussions required expertise from both groups. In October a combined meeting was held and this will be repeated in December. Starting 2016 a single committee meeting will become the MO.

Discussions in the meetings are open and to the point and have evolved into sessions that not only govern WP2 but also connect WP3, 4 and 5. During the committee meetings ideas and know-how is exchanged between the work packages. Attendance is growing beyond the original coordinating members to also include engineers working on various tasks inside and outside WP2.

1.3 User Engagement & Community building

The WP2 committee meetings have shown that there is a clear desire for community engagement. This engagement is channelled through three initiatives within WP2. First of all there is a growing central mailing list to inform partners on the general developments in CLARIAH and the WP2/committee meetings specifically. Besides this rather directive medium, WP2 has setup a Basecamp project tool to accommodate debate. This environment is extensively used by the WP2 community for discussions on best practices, knowledge exchange etc. Thirdly, WP2 has setup a

central CLARIAH github account for code sharing. The account has become popular among the CLARIAH community and features ten code- and data repositories at the time of writing this report.

The WP2 committee members and the community have frequently expressed the need for an improved CLARIAH website that can integrate these WP2 initiatives and to allow the community to expand beyond the immediate partner institutes.

2. Components

Project plans are currently being finalized and will become available on the WP2 Basecamp-site.

2.1 Core infrastructure

Plans for tasks 11.100 – LOD infra and 11.200 – Maintenance Tools have been combined with 41.100 – Data Access Environment and 42.100 – Software Plugins & Libraries (both components of the User Interaction Infrastructure) collectively known as ANANSI. Huygens ING has hired Jauco Noordzij as lead- engineer for the development of this part of the project. The team will be a partnership between Huygens ING, DANS and VU. Jauco has started in October 2015. Task 11.300 – Data Conversion is scheduled for 2016 or 2017 and is not yet initiated. Meertens Institute has started development of Task 11.400 – CLARIN CMDI interoperability.

For tasks 13.100 – Identity Management, 13.200 – Central User Management and 13.300 – Homeless Users NISV has recently hired Themistoklis Karavellas and a project plan will be made available soon.

2.2 Structured data

For task 21.100 – Person Entities a model is being drawn up by Sebastiaan Derks (Huygens ING). Besides person records from the research domains (WP3, 4 and 5) the inclusion of several external datasets are currently subject of debate. This resulted in a positive reply from 'Biografisch Portaal' and a negative answer from 'HSN' (historic sample of the Netherlands). Discussions to include NIOD data are on-going. Richard Zijdeman (IISH) filed a project plan for task 21.200 – Location Entities and Katrien Depuydt (INL) did the same for task 21.400 – Diachronous Semantic Dutch Lexicon. Work on both tasks has recently started. Given the small amount of work Task 21.500 – Document Entities has been put on hold for now.

2.3 User Interaction Infrastructure

Task 41.100 – Data Access Environment and 42.100 – Software plugins & libraries make the core infrastructure accessible for users. For this reason the development of these tasks are combined in ANANSI – as discussed above. INL, Tilburg University and Radboud University have started work on task 41.400 – OCR/TICCL pipeline in November 2015. This project ties closely to work scheduled in WP3. WP2 agreed with Antal van den Bosch and Martin Reynaert that improvements to the algorithms in the TICCL pipeline are the domain of WP3 while work related to turn TICCL into production level software – like optimization, documentation etc. – are the responsibility of WP2. There is evidently a grey area in managing these activities and WP2 and WP3 agreed to deal with these issues pragmatically.

2.4 Standards & Best Practices

NISV is drawing up a plan to setup a standardization process as detailed by Task 52.100. DANS has filed the proposal for task 54.100 – Guidelines for Documentation, Data & Software Sustainability and INL for task 55.100 – CLEVER. In the WP2 committee meetings the community has expressed the urgency for all three tasks and the necessity to accommodate a discussion that will evolve standards and guidelines. A discussion platform should be integrated in the new CLARIAH website and made available to the public before the 2016 open CLARIAH call.

3. Relations & integration

3.1 Internal CLARIAH

Work in WP2 has a significant focus on structured linked open data, which creates an overlap with WP4. In order to tie the two work packages closely together WP2 and WP4 have arranged regular

technical meetings; Rinke Hoekstra (lead engineer at WP4) is appointed as consultant in WP2 ANANSI; and WP4 engineers and coordinators participate strongly in the WP2 committee. The relationship between WP2 and WP5 is comparable to that between WP2 and WP4. Participation of WP5 engineers and coordinators is very strong and the two work packages recently had the first of a regular series of technical meetings. WP5 has invited WP2 for a day of brainstorming and plan making in November. Integration between WP2 and WP3 is more procedural compared to WP4 and WP5. This may be the result of the fact that – unlike the other two packages – it is formalized in a specific task (11.400 – CLARIN CMDI interoperability). WP2 would like to strengthen its partnership with WP3 and for this reason a technical meeting will be arranged in December 2015 – comparable to the meetings with the other two work packages. It is the intention of WP2 to develop this into a regular series as well. Engineers working in the work packages are very welcome to actively participate in the WP2 discussion meetings and the community.

Each of the domain work packages has a technical officer (WP3: Antal van den Bosch/Piek Vossen, WP4: Frank van Harmelen and WP5: Maarten de Rijke). According to the WP2 Technical Plan the connection between WP2 and the technical officers is maintained through Frank van Harmelen/Rinke Hoekstra in the committee meetings. Piek Vossen has joined the WP2 committee meetings directly. WP2 would like to open a more structural line of communication to exchange ideas and knowledge during the actual development stage of CLARIAH (2016- 17) or at the request of the technical officers.

3.2 External integration

Integration activities in WP2 have so far mostly focused on the internal relations in CLARIAH. For 2016 this focus will expand to include external partners as well. In 2015Q3 WP2 has initiated discussions to integrate the CLARIAH infrastructure and dataset with dbpedia (NL) and Europeana. It is the intention of WP2 to conclude these discussions early 2016 in order to start experiments in 2016Q3-4. Likewise in 2015 WP2 has had a limited connection to both the CLARIN and DARIAH ERICs. It is the intention to participate more deeply in workgroups of both infrastructures starting 2016. Finally, in the spring of 2015 WP1 initiated a closer cooperation with CLARIAH AT (Austria). WP2 and CLARIAH AT will meet in Vienna in December in order to investigate the potential to integrate activities and infrastructures in 2016.