Interim Evaluation CLARIAH-CORE

Self-Evaluation Report

V3.1 2017-07-10

Contents

1 Introduction 4
2 Highlights 4
3 Summary 4
3.1 Goals 5
3.2 Assessment 5
4 Evaluation per WP 7
4.1 WP1 (Overall Management) 7
4.1.1 Goals 7
4.1.2 Overall Assessment 7
4.2 WP2 8
4.2.1 Overall Assessment 8
4.2.2 Assessment per task/subgoal 8
4.2.3 Recommendations 13
4.3 WP3- Linguistics 15
4.3.1 WP3 Goals 15
4.3.2 Overall Assessment 16
4.3.3 Recommendations 17
4.4 WP4 18
4.4.1 WP4 Goals 18
4.4.2 WP4 Assessment 18
4.4.3 Recommendations 20
4.5 WP5 21
4.5.1 WP5 Goals 21
4.5.2 Assessment 22
5 Appendix Detailed Reports 24
5.1 Detailed Report WP1 Overall management 24
5.1.1 Assessment per task/subgoal 24
5.2 Detailed Report WP3 Linguistics 34
5.2.1 Introduction 34
5.2.2 General Overview 34
5.2.3 Detailed Overview per Partner and Task 39

5.3 Detailed Overview WP4 Social Economic History 58
5.3.1 Overall Assessment 58
5.3.2 Assessment per task/sub goal 59
5.4 Detailed Overview WP5 Media Studies 62
5.4.1 Overall Assessment 62
5.4.2 Assessment per task / sub goal 63
6 Appendix: Acronyms 67

1 Introduction

This is the CLARIAH-CORE Self Evaluation report for the Interim Evaluation. We first list some highlights of the CLARIAH-CORE project so far (section 2), provide a short overall summary of this report (section 3), and then discuss the status of each of the work packages (section 4). The status reports have been kept short on purpose, but more details can be found for several work packages in the Appendix (section 5). The descriptions of WP2 and WP5 are slightly differently structured. They contain all information but may receive an update in the coming weeks.

Factual data on the CLARIAH CORE project can be found in the CLARIAH CORE Fact Book.

2 Highlights

- Despite some delays in the start-up phase mainly due to difficulties in recruiting (especially engineers) and some technical issues, the project as a whole is reasonably well on course to reach its goal before the planned end date.
- Call for Research Pilots Launched (and in the meantime 16 projects awarded funding).
- Joint Call with NL eScience Center launched (proposal evaluation is underway).
- ANANSI beta version made available
- Upgraded version of the treebank search application PaQU ready and in use.
- Tools to transpose csv or excel into Linked Open Data both for less and more advanced users (QBer, CoW) (best WHISE 2016 paper award)
- Tools for querying across multiple long-term macro and micro datasets, and sharing and executing those queries online (best SALAD paper award)
- There is a lot of cooperation across WPs, e.g. WP4 shares data with Anansi (WP2) and focuses on cross-WP research (Hack-a-LOD 2016 audience award with WP3). A lot of cross-disciplinary work in the research pilots.
- Launch of the first version of the Media Suite, a virtual research environment for accessing audiovisual data and tools for collection analysis, search, annotation, analysis, and visualization.
- Successful CLARIAH kick-off, ‘toog’ and ‘tech’ days held.
- Successful CLARIAH Linked Open Data workshop held.
- Successful international Linked Open Data for Linguistics Workshop organized in Utrecht.
- Presentations at Digital Humanities 2016 conference on teaching television history with digital data and tools, annotating audiovisual heritage in a media studies context and representing missing data in collections.

3 Summary

3.1 Goals

The CLARIAH-CORE project aims to create a research infrastructure for humanities researchers as part of the European research infrastructures CLARIN and DARIAH. It will provide generic infrastructural facilities (WP2) and facilities for three core disciplines: linguistics (WP3), social-economic history (WP4), and media studies (WP5). With these three core disciplines, CLARIAH covers their dominant data types: text (linguistics), structured data (social-economic history), and audiovisual data (media studies). CLARIAH-CORE actively pursues cross-disciplinary synergy. It aims to disseminate its results widely, to involve the intended users (humanities researchers) in design and development decisions, to educate and train students and researcher in using the infrastructural facilities in their research, and to test the infrastructure in research pilot projects (WP1). CLARIAH-CORE evidently has great potential for societal impact, inter alia in the Creative Industry domain and the identified top sectors, as well as for language, migration and economic policies. It also covers multiple aspects of the National Science Agenda.

3.2 Assessment

Overall, it can be stated that the CLARIAH-CORE project is well underway after two years despite a somewhat delayed start for most WPs. Nevertheless, there clearly is a delay in spending the budgeted money in most WPs, mostly due to problems in obtaining the required (engineering) capacity. Cross-disciplinary synergy is visible in concrete cooperation projects across disciplines, such as the Athena project (WP3, WP4), the occupation detection project by WP3 and WP4, and very strongly in the awarded research pilot projects. CLARIAH plays an active role in the European infrastructures CLARIN and DARIAH. Events to disseminate the results are organized regularly, there is strong support for relevant events organized by others, as well as for CLARIAH participants to visit conferences and workshops to share their work with their peers. Several papers and other contributions at conferences and workshops have been assigned awards. Educational events have been organized and this will be increased in the coming years since there is now functionality to educate and train students and researchers in.

In WP1 the governance for the project has been set up. Good relations with the European infrastructures CLARIN and DARIAH are being maintained. Numerous other projects have received support from CLARIAH CORE in many different forms. We have been active in the process for the national roadmap for large-scale scientific infrastructure. CLARIAH has been put on the national roadmap and coordinates all Humanities activities for the roadmap. Two calls have been launched, one for research pilots (16 projects awarded funding and about to start) and one in cooperation with the NL eScience Center (evaluation of the submissions is ongoing). A lot of events have been organized, supported or attended, and many CLARIAH participants have been supported in visiting workshops and conferences to report on their work for CLARIAH and discuss it with peers.

The overall status of WP2 is positive. The main projects are progressing, although with some delays. These delays are mostly caused by a lack of engineering capacity at major partners like NISV, Huygens ING and INT. Halfway through the project, we have no immediate cause for concern – projects are advancing, although in some cases slower than planned. About a third of the total available budget of €1.8M is spent. Since most projects are scheduled to end late 2017 or early 2018 we expected to have

spent about two thirds of the budget instead. However, there is time to compensate for these delays. Mid-2017 we will reassess the situation to see if more pressing actions are necessary.

WP3 is in general well on schedule, though there some delays at INT (delayed start-up of projects due to reorganisation) and at UU (problems with personnel falling ill and leaving). Several tasks of Meertens require re-evaluation and perhaps a change in character, in part by independent developments. WP3 is currently in the process of redefining these tasks. WP3 is involved in cross-WP tasks (specifically cooperating with WP4 and with Athena). WP3 members play an active role in the European CLARIN infrastructure

WP4 is progressing as planned. In the first year several tools and applications were delivered. In the second (past) year, a more stable version of the HUB has been built within the International Institute of Social History (IISG), which will host the HUB in the second half of CLARIAH and at least 5 years beyond. A number of datasets have been transposed into Linked (Open) Data and tools were created for transposing data and querying data. Finally, we are engaged in various cross-WP initiatives (WP2, WP3). It was not easy to find qualified personnel, especially for the knowledge representation part. Another issue with the project is the learning curve and coordination needed to transpose datasets into Linked Data in a *meaningful* way.

The start of WP5 was marked by a long stage of inventory and design in 2015 and early 2016, which featured monthly meetings and various additional design sessions and user studies (aimed at defining and refining requirements for the Media Suite). In 2016 we started building the foundations and components of the first version of the Media Suite. The initial plan was to make the tools and collections listed for version 1 available to researchers from Q1 2017. This has been postponed with three months; version 1 will be launched on 4 April 2017 so is available from Q2 2017 – still in time for usage in the Research Pilots.

4 Evaluation per WP

4.1 WP1 (Overall Management)

4.1.1 Goals

The goals of WP1 are setting up the governance structure, maintaining good relations with and play an active role in the European infrastructures CLARIN and DARIAH, supporting other infrastructure and research projects and project proposals, organizing and supporting dissemination and outreach, as well as education and training, and setting up a call for research pilots.

4.1.2 Overall Assessment

The governance has been set up and consists of the following bodies:

- Board
- Executive Board
- Supervisory Board (‘Raad van Toezicht’)
- International Advisory Panel

The executive board is a subset of the board. The tasks and responsibilities of the governance bodies has been defined in by-laws and approved by the Supervisory Board. The CLARIAH office has been set up to support the activities in WP1. The governance structure functions well.

Good relations with the European infrastructures CLARIN and DARIAH are being maintained. National coordinators for CLARIN (Jan Odijk) and DARIAH (Henk Wals) have been appointed. Many participants in CLARIAH play an active role in CLARIN and DARIAH committees and task forces. CLARIAH is well represented at the yearly CLARIN and DARIAH conferences.

Numerous other projects have received support from CLARIAH CORE in different forms.

We have been active in the process for the national roadmap for large scale research infrastructures. CLARIAH has been put on the national roadmap and coordinates all Humanities activities for the roadmap. A proposal for a successor project will be submitted in June 2017.

A lot of events have been organized, supported or attended, and many CLARIAH participants have been supported in visiting workshops and conferences to report on their work for CLARIAH and discuss it with peers.

Although only one call for projects was planned, actually two calls have been launched, one for research pilots (16 projects awarded funding and about to start) and one in cooperation with the NL eScience Center (evaluation of the submissions is ongoing).

An interdisciplinary project (ATHENA) to develop a data portal that will hold information on historical context of human – nature relationships for a broad variety of plant and animal species and the

landscapes and ecosystems they live (d) in, was proposed and accepted by the CLARIAH Board. It is an excellent example of the interdisciplinary (hence cross-WP) approach that CLARIAH enables.

There are a number of budgeted activities (in particular around brain gain and IPR) that still have to be planned and executed in the coming two years of the project.

4.2 WP2

4.2.1 Overall Assessment

The overall status of WP2 is positive. The main projects are progressing, although with some delays. These delays are mostly caused by a lack of engineering capacity at major partners like NISV, Huygens ING and INT (formerly INL). NISV (understandably) prioritizes available engineers for CLARIAH WP5; Huygens ING was at full capacity but faced resignations after moving to Amsterdam in 2016Q4 and INT went through a significant reorganization and downscaling of its activities in 2016.

Halfway through the project, we have no immediate cause for concern – projects are advancing, although in some cases slower than planned. About a third of the total available budget of €1.8M is spent. Since most projects are scheduled to end late 2017 or early 2018 we expected to have spent about two thirds of the budget instead. However, there is time to compensate for these delays. Mid-2017 we will reassess the situation to see if more pressing actions are necessary. If so, we will realign funding or add partners to existing projects in order to reach the objectives before end 2018.

4.2.2 Assessment per task/subgoal

4.2.2.1 Project: ANANSI

Tasks:

- 11.100 RDF Infrastructure
- 11.200 Maintenance Tools
- 11.300 Data Conversion
- 41.100 Data Access Environment
- 42.100 Software Plugins

Partners:

- Huygens ING
- VU
- DANS

Planned dates: 2015Q4 - 2018Q4
Effective dates: 2016Q1 - 2018Q4

Status: project is ongoing but with a slight delay due to a late start. Project is expected to be delivered on-time.

The project planned to release an ANANSI beta-version in 2016Q4. The software was eventually released in 2017Q1 and can be found at http://anansi.clariah.nl. The non-public alpha-version was first presented at the CLARIAH Techdag (2016Q3) - https://vimeo.com/186090384. The beta was first demonstrated at the CLARIAH Toogdag (2017Q1), and later at the symposium organised by the Platform Linked Data Nederland also in 2017Q1.

We expect data from WP3, 4 and 5 to become available through ANANSI over the course of 2017.

The beta-version is the result of a year-long platform-driven development effort by DANS and Huygens ING. In 2017 development will switch into a user-driven mode – ANANSI will be tested by users and their feedback will direct further development. At this point we already see people starting to experiment with the beta-version, a CLARIAH ANANSI user workshop is planned for 2017Q2. E.g. the current version provides a lot of importing and curation tooling to get data in the system. We now expect users to guide us to their preferred (graph) analytical and visualisation tools and focus the development effort on implementing those.

Also in 2017Q1 the VU will re-join the project. During 2016 the team focussed on related work in CLARIAH WP4. VU participation in ANANSI will concentrate on integration of WP2 and WP4 structured data technology.

4.2.2.2 Project: CMDI Interoperability

Tasks:

- 11.400 CLARIN CMDI Interoperability

Partners:

- Meertens Institute

Planned dates: 2015Q2 – 2018Q2
Effective dates: 2015Q2 – 2018Q2
Status: project is ongoing and on track.

A virtuoso server exposing CMDI data as RDF is available. Integration with the CLARIAH infrastructure has been provided as a Virtuoso OAI-RS plugin by the ANANSI project. Implementation by MI is scheduled for 2017Q2. The project team now continues to work on mapping RDF back to CMDI and has paper presentations on the subject planned.

4.2.2.3 Project: Central User ID Management

Tasks:

- 13.100 Identity Management

- 13.200 Central User Management
- 13.300 Homeless Users / OpenID

Partners:

- NISV (Beeld & Geluid)

Planned dates: 2015Q4 – 2016Q4
Effective dates: 2015Q4-?
Status: project is ongoing but delayed.

NISV has significant problems hiring engineers – available capacity is primarily dedicated to development in WP5. This has caused a delay in the delivery of the security server. The CTO is working closely with NISV to address the situation. We have effectively agreed to an extension of the project – a new end date will be set after the test environment has been made available to the CLARIAH community early April 2017. The intention is to have a fully operational production platform ready before 2017Q4. The system will be thoroughly tested by VU (WP4) and the ANANSI project (WP2) before acceptance as a production platform. In 2017Q1 all formal agreements for sharing of user metadata information with CLARIN have been arranged.

4.2.2.4 Project: H-PEP

Tasks:

- 21.100 Person Entities

Partners:

- Huygens ING

Planned dates: 2016Q1 – 2018Q4
Effective dates: 2016Q1 – 2018Q4
Status: project is ongoing and on track.

HI has prepared a standardized data model for persons based on the Biografisch Portaal collection as planned. This dataset has been curated and transformed into RDF and is currently available in a test environment. The full version will become operational in ANANSI as scheduled by 2017Q2.

4.2.2.5 Project: GeoTide

Tasks:

- 21.200 Location Entities

Partners:

- IISH
- Huygens ING

Planned dates: 2016Q1 – 2017Q4
Effective dates: 2016Q1 – 2017Q4
Status: project is ongoing and on track.

Geotide creates linked open data on historical locations. The project is split between IISH – data from 1812 onwards – and HI – earlier data. These two sets are fundamentally different. IISH is representing historical country borders (from 1950 onwards) and Dutch municipality borders from 1812. In collaboration with the Dutch Central Bureau of Statistics, Triply and Hic Sunt Leones, so far IISH has managed the following: transposition of historical country borders worldwide since the 1950's into Linked Open Data. For the Netherlands it has created the missing border polygons from 1997 up until today. These polygons are transposed for municipality borders and can be downloaded from http://www.gemeentegeschiedenis.nl. This section of the project is virtually done and what remains is the data distribution via ANANSI. Huygens has concentrated on the person project H-PEP first, and is now preparing the second section of the project – gathering location data from historical databases. This is fully on schedule and ready to be delivered by the end of this year.

4.2.2.6 Project: DiaManT

Tasks:

- 21.400 Concept Entities

Partners:

- INT (formerly INL)

Planned dates: 2015Q4 – 2018Q4
Effective dates: 2015Q4 – 2018Q4
Status: project is ongoing and on track.

INT has made a thorough assessment of datasets that will be traced diachronically – this assessment does not only include readily available linguistic corpora but also commonly used search terms by scholars – e.g. those based on the most popular searches in Delpher. The work is done in cooperation with tasks from the VU in WP3. A prototype will be made available on schedule in 2017Q2. The project led to several presentations and publications:

- *Diachronic Semantic Lexicon of Dutch,* poster, pitch (Katrien Depuydt) + demonstration (Jesse de Does) at *Digital Humanities 2016*, in Krakow 12-15 July 2016.
- Naar een diachroon semantisch lexicon van het Nederlands, Katrien Depuydt. Lecture at the information session: *CLARIAH Call for Research Pilots en Lezing Katrien Depuydt (INT)* in Utrecht 22 September 2016,
- Naar een diachroon semantisch lexicon van het Nederlands, presentation by Katrien Depuydt and demo by Jesse de Does, at the visitor day of INT at 12-10-2016.
- Naar een diachroon semantisch lexicon van het Nederlands, presentation by Katrien Depuydt and demo by Jesse de Does, for advice council of INT at 2-12-2016

4.2.2.7 Project: PICCL

Tasks:

- 41.400 OCR/TICCL Pipeline

Partners:

- University of Tilburg
- Radboud University Nijmegen
- INT (formerly INL)

Planned dates: 2015Q4 – 2017Q4
Effective dates: 2015Q4 – 2017Q4
Status: project is ongoing and on track.

During 2016 this project was executed in combination with work funded through WP3. This integration focused on the general improvement of the TICCL-algorithms, avoided duplication of activities and tied WP2 and WP3 together. In 2017 the two projects will be more separate as the focus for the PICCL project will shift to the delivery of a stable production pipeline hosted by INT. A project meeting to this end is planned for late 2017Q1.

4.2.2.8 Project: QSODA

Tasks:

- 54.700 Documentation, Data & Software Sustainability

Partners:

- DANS
- Radboud University Nijmegen

Planned dates: 2015Q4 – 2018Q1
Effective dates: 2015Q4 – 2018Q1

Status: project is ongoing and on track.

QSODA delivers annually updated guidelines to promote data and software sustainability across CLARIAH. The first version of these guidelines was delivered on schedule late 2016 and can be found on the CLARIAH GitHub page: https://github.com/CLARIAH/software-quality-guidelines and included an interactive survey where users can rate software packages based on the guidelines.

4.2.2.9 Project: CLEVER

The project aims to evaluate the current situation with regard to availability and performance, and to make an estimate of future human and financial resources, and functional requirements. This will be done in terms of a balance between the achievable quality of the software, the use of computing power at a reasonable price and an acceptable quality of research, teaching and the wider public.

Tasks:

- 55.100 Performance & Availability

Partners:

- INT (formerly INL)

Planned dates: 2016Q1 – 2016Q2
Effective dates: 2017Q1 – 2017Q2
Status: project is significantly delayed and scheduled to commence 2017Q1.

Due to reorganization and severe engineering capacity problems at INT – this project was delayed by a year. WP2 and INT have agreed to start CLEVER late 2017Q1. Since the size of the project is limited and is planned to take only half a year, the effects of this delay are minor.

4.2.2.10 Project: Standardization

Tasks:

- 52.100 - Standardization process: study and implementation

Partners:

- NISV

Planned dates: 2015Q4 – 2016Q4
Effective dates:? - ?
Status: project is significantly delayed

This task involves the study and implementation of a consensus-based decision process that can be called upon in case of emerging standardization issues that likely will arise in respect to the implementation of the core infrastructure. CLARIAH aims to facilitate a means to discuss such standardization requests within an organisational structure based on well-established standardisation working bodies such as W3C.

NISV has significant problems hiring engineers – available capacity is primarily dedicated to development in WP5. Within WP2 we have agreed to dedicate available capacity to the Central User ID Management since several software projects across CLARIAH are dependent on this solution for their security authentication and authorization. This project will be released in April 2017 as a test environment and at that stage we will start discussions with NISV on progress on the standardization process. The eventual deliverable should be a whitepaper.

4.2.3 Recommendations

In the second half of the project the focus should be on consolidation and implementation activities instead of the development of new technology. Various platforms that have been constructed over the first one and a half years, now reach a stage in which users can start interacting with them. Further development should therefore be guided by user feature requests, workshops, training etc. instead of being platform-driven.

A second recommendation regards deeper integration of the work packages – and this can only carry over into a subsequent grant. At the start of CLARIAH various domains started in highly different conditions. WP3 has had about 6 years of CLARIN funding and evidently is much further ahead than e.g.

WP4 – that had to start almost from scratch. Also WP2 needed time to get a central infrastructure ready for integration. Presumably these differences are smaller in the second half of the project and hopefully in CLARIAH2. It is now much more visible what the various domains and WPs have to offer and it is important to now start an integration drive. E.g. we originally planned that WP2 would get data from the various other WPs. This data connection and the surrounding infrastructure (ANANSI) is now available as beta. Instead of only acquiring data, we should now investigate how other software systems, both within CLARIAH - e.g. the WP5 media suite – and elsewhere could further improve the CLARIAH infrastructure. Resources for this drive might be available within current funding schemes or should get priority in the CLARIAH2.

4.3 WP3- Linguistics

4.3.1 WP3 Goals

WP3 aims to provide research infrastructure facilities for carrying out linguistic research. It aims to cover each phase in a typical linguistic research project, and therefore provides facilities for obtaining data and tools (both finding existing data and creating new data); for automatic, semi-automatic and manual enrichment of data with various linguistic annotations; for search in and analysis of the data; for visualization of search and analysis results; for publishing data and software in the CLARIAH infrastructure; and for creating and publishing enhanced publications on the research. Interoperability is an important topic that plays a crucial role for all data and tools in all these phases.

The CLARIAH-CORE WP3 plan describes, for each of these aspects, what is needed, what is already available (from earlier projects such as STEVIN, CLARIN-NL, Nederlab and CLARIAH-SEED), and which facilities must be created anew or upgraded.

Obtaining data and tools. The major activities here involve specifying requirements for metadata, metadata curation, and creating tools for browsing and searching in metadata. For creating new data crowdsourcing and survey software will be developed. Research data management guidelines will be defined. Further development of and support for often used formats and protocols will be carried out (CLAM, FoLiA). A tool will be created to enable researchers to easily make metadata for often occurring resource types.

Enrichment of data. This includes upgrades of various tools that automatically assign linguistic annotations to data (UCTO, TICCL, Frog, Frog Generator), as well as further development of tools for manual annotation (FLAT). The existing workflow system TTNWW will be upgraded and a new workflow system for annotation will be set up (PICCL, in cooperation with WP2).

Search. This includes upgrades to search applications (CELEXweb, GrETEL, PaQu, OpenSoNaR, MTAS, MIMORE, AutoSearch) to search for (examples of) linguistic properties. It also involves integrating the Nederlab data and software. It also covers various specific forms of federated search, and chaining search (i.e. combined search in heterogeneous resources).

For extraction of non-linguistic information from data the work includes extraction of occupations from texts in (older versions of) Dutch (in close cooperation with WP4), as well as extraction of events and emotion detection. Cooperation with WP5 on extracting metadata relevant for media studies from text is also planned.

Publishing data and software. This includes the creation of a metadata SPARQL endpoint, facilities for dealing with data and tools with IPR and/or ethical restrictions, archiving and ingest functionality and

setting up a deployment framework. It also includes hosting of the PaQu and GrETEL search applications by INT.

Enhanced publications. No activities are planned for creating and publishing enhanced publications.

Interoperability. This includes determining CLARIAH-formats, adding facilities for supporting multiple formats in all tools (converters). For ensuring semantic interoperability an upgrade of OpenSKOS (used by the CLARIN Concept Registry) will be made. Shared vocabularies will be developed and vocabulary items linked to entities.

4.3.2 Overall Assessment

Overall. The work package organization was originally organized along themes, each with a team leader, but in practice management has been carried out directly with the organizations involved. The tasks for federated search were originally to be assigned to Meertens institute, but have actually been assigned to INT. The cooperation among the partners is generally excellent. There are monthly progress meetings, and regular meetings with WP2. The management team has been holding regular meetings, since the end of 2016 once a week. A number of tasks will have to be reconsidered due to independent developments. This is planned for Q2 2017. Since the created functionality is distributed over a number of institutes it is important to show that together this forms a single coherent research infrastructure. This will be given special attention as of 2017.

WP3 members were active with CLARIAH as a whole in various respects, organized and co-organized several events, and held presentations, posters and/or demos on all CLARAH events. WP3 also plays an active role in the European infrastructure CLARIN.

Obtaining data and tools. Initial versions of requirements for metadata have been formulated. Metadata curation is ongoing for a selected number of data collections from the Netherlands. Work has been done on creating metadata for software. Both metadata curation and creating metadata for software are delayed because of staffing problems. The impact of this delay is very limited. The creation of tools for browsing and searching in metadata has been postponed and will be reconsidered due to the fact that CLARIN Austria, with which we closely cooperate, has developed such a tool that we can also experiment with. Work on the survey software is progressing as planned, but the development of crowdsourcing has been postponed because of independent developments with regard to crowdsourcing at the KNAW. It will be reconsidered mid 2017. The work on research data management guidelines has been postponed since many university libraries are also working on them. Development of and support for CLAM and FoLiA has progressed as planned. Development of the tool for easy creation of CMDI metadata has been postponed, awaiting clear specifications.

Enrichment of data. Upgrades of enrichment tools (UCTO, TICCL, Frog) have been made. Frog Generator has been created, which enables one to create one's own linguistic annotation tool on the basis of a corpus of manually annotated data for an arbitrary language. This tool has been tested on Classical

Greek. Development of the FLAT tool for manual annotation is progressing as planned, as is the work on the workflows for annotation, PICCL and TTNWW.

Search The planned work on CELEXWeb and PaQu has been finished. WP3 will invest part of its reserve in a further extension of PaQu, adding an interface with a menu of predefined queries for syntactic profiling. This will make it possible, e.g., to study syntactic development of school age children.
The development of upgrades to the search applications MTAS and MIMORE is progressing as planned. The work on GrETEL is progressing, but in a slower pace than originally planned due to limited engineering capacity. This is not a problem since there are no dependencies and the planned work can be carried out before the CLARIAH project end date. Work on OpenSoNaR and AutoSearch started later than originally planned but is now progressing in accordance with the revised planning. Integration of Nederlab is progressing as planned. Plans for the work on federated search and chaining search have been elaborated, but the work is still to start.

For extraction of non-linguistic information from data, a tool for extraction of occupations from texts in (older versions of) Dutch has been developed and is being tested in close cooperation with WP4 and the Athena project. Extraction of events and emotion detection is planned for 2017, and the cooperation with WP5 for 2018.

Publishing data and software. The metadata SPARQL endpoint has been delivered on time. Work on the deployment framework is ongoing. Archiving and Ingest software has been created in the context of The Language Archive (TLA). It is called FLAT (not to be confused with the manual annotation tool with the same name). Work on facilities for dealing with data and tools with IPR and/or ethical restrictions still has to start. Hosting of PaQu and GrETEL by INT still has to start.

Enhanced publications. No activities were planned.

Interoperability. Most of this work has started later than originally planned but otherwise on schedule. Work was done on tools to convert NAF into FoLiA (and vice versa). Many tools (PaQu, GrETEL, AutoSearch) have extended their support for different input formats or are in the process of doing so. Various formats for modeling lexical data were considered, in particular the LEMON model. Several diachronic lexicons in this format were presented. A module for fine-grained entity typing was developed. It turned out that many entities mentioned in humanities texts are not present in any repository, which led to work on discovering information on such ‘dark entities’.
A workshop on interoperability was organized, leading to concrete tasks and deadlines, which will be followed up at the end of 2017. The upgrade of OpenSKOS has been delivered on time.

4.3.3 Recommendations

These have been included in section 4.3.2.

4.4 WP4

4.4.1 WP4 Goals

1. Deliver a 'structured data hub' that forms a single point of entry to a live repository of interconnected and (partially) harmonized datasets pertaining to the field of socio-economic history (SEH).
2. By providing interlinked, integrated datasets, the hub should facilitate formulating and answering **cross-dataset research questions**.
3. The hub should grow to become a research infrastructure that is a **major resource** in the field of SEH. To achieve this, we envisage that it should support curation, storage, finding, linking, selecting, visualization and analysis of structured data. To what extent the infrastructure will provide rather than enable these tasks should be made clear through the requirements analysis.
4. The hub should demonstrate its **value** (and the value of its data) to non-experts in the domain.
5. The project will deliver a **critical mass** of data made accessible through the hub. The datasets in the hub have to meet a number of criteria pertaining to relevance, importance and quality. This is to maximize the impact of the data hub on current research questions, and attract individual researchers that want to **contribute** and/or **study** the data it hosts.
6. The hub should be **integrated** with the central CLARIAH infrastructure delivered within work package 2 of the project. This is safeguarded by membership of key WP4 personnel in WP2 fora, and physical proximity at regular intervals.
7. The hub should be **hosted** at and **integrated** with the infrastructure of the IISG.
8. The hub should integrate and **adopt lessons learned** from its predecessors. Most notably CEDAR, HSN and ClioInfra. This involves migration of CEDAR, HSN and ClioInfra data and tools to the CLARIAH infrastructure, to the extent that this contributes to CLARIAH goals.
9. The hub should be up to date and reflect the latest version of datasets as much as possible. It should therefore be **resilient to changes** to any datasets that it depends on.

4.4.2 WP4 Assessment

Ad 1. In year 1 we have built a prototype HUB and evaluated it. Now we have built an improved version enhancing stability and scalability.

Ad 2. We have transposed various datasets into Linked Data. On the CLARIAH Toogdag 2017 we have demonstrated how we were able pull data from 4 different international datasets to answer a research question on the influence of religion on someone's socio-economic position, controlled for a country's economic development. With the increasing number of datasets we plan to transpose opportunities for cross-querying datasets will be increasing over the next two years.

Ad 3. We are advocating the HUB nationally and internationally, e.g. through sounding board meetings, at workshops and conferences, training Master's and PhD students and Post-docs with their data and 'our' tools. Internationally we are now focusing on advocating the HUB on social and economic conferences, e.g.: expert-workshop on linking Swedish Historical Person data (April 2017), Digital Humanities Benelux (July 2017), Social Science History Association conference (November 2017), World Economic History Conference (under review). We have also been very active in supporting historians with their CLARIAH call proposals, in order to retrieve maximum feedback on our tools, resulting in four financed projects.

Ad 4. We have attracted a grant to hire a Master's Student demonstrating the use of Linked Data for contemporary questions on migration flows (paper presented at AISB, Bath 2017). For hands-on use by non-experts we are closely collaborating with WP2 regarding front-ends and usability. We have dedicated an engineer 1 day a week to WP2 for exchange related issues and these front-end issues.

Ad 5. We are well underway with transposing datasets into Linked Open Data. For various datasets we had some minor issues, such as retrieving the actual owner of the datasets and ask for rights. For a number of datasets that are restricted we now also have restricted Linked Data and are able to demonstrate this to the respective parties and see whether they would be interested to share their data this way.

Ad 6. We have dedicated an engineer for one day a week to WP2. This has resulted in the fact that at the moment ANANSI is able to extract all Linked Open Data from our HUB. The communication is however still experimental and will be fully fledged in the coming year.

Ad 7. We built the HUB from inside the IISH environment already allowing for perfect integration. Tooling is still based at the VU University. In the coming year, we will focus on moving the tooling to the IISH as well.

Ad 8.

- CEDAR: CEDAR is available from our HUB. In addition to migrating the data and endpoint we had to align domain names.
- HSN: a sample of the HSN has been transposed to Linked Data. Because of privacy issues, the data are non-open and require registration with HSN.
- ClioInfra: the data in ClioInfra have been transposed to Linked Open Data.

Ad 9. We are still working on a workflow to properly adopt changes in files. We are able to pull data from Dataverse and then transpose it to Linked Data. This could even be automated with event hooks. (Thus recreate the Linked Data once a new version of the dataset appears in Dataverse). The question however is, what to do with the Linked Data of the old version? Obviously you want to preserve it, but you also do not want multiple versions of the same data in the graph, as it reduces speed.

4.4.3 Recommendations

For the second half of the project there are a number of issues we recommend looking at. For one, most cross-WP collaboration is still very fragile. We have engaged with all WP's, but we are not yet able to answer a research question with data from texts (WP3), audio/video (WP5) and structured data (WP4). In addition to tech-days we could think of theme-days, where we would try and connect multi-type data on a particular theme. Even if it would not succeed, it would clearly show the hurdles to take for cross-WP interaction.

We also need to get much more user engaged. The KNHG organized a 'interrogate our tools' workshop, where tools were presented and then users could apply them and give feedback. That seemed like an excellent format to get better feedback on our tooling, WP specific, as well as for the entire CLARIAH pipeline. So we would recommend organizing at least one such a day in the coming year.

To give users, in some cases including CLARIAH-internal users a better overview of the available functionality a user-friendly front-end or front-ends should be built in the second half of the current project. The augmentation of for example structured data with information derived from texts (and vice versa), will in multiple disciplines be a major leap forward. However, this also may require broadening the scope beyond the academic discipline and add in sources and knowledge from cultural heritage institutes.

4.5 WP5

4.5.1 WP5 Goals

Work Package 5 (WP5) is engaged in setting up an infrastructure for access to audiovisual sources and related contextual material (program guides, RTV ratings, photographs, etc.). It aims to consolidate the functionalities of five existing tools for exploratory and targeted, contextual media research (CoMeRDa, AVResearcher, Trove, Dive and Oral History Today, developed as prototypes in the context of NWO CATCH, CLARIN-NL and CLARIAH-Seed) in a research environment, to advance these functionalities based on scholarly requirements and to train scholars in using them. For this purpose, we have designed the Media Suite: a virtual research environment in which scholars find the data and tools for their research and which offers them the opportunity to build collections, bookmark and enrich them.
The CLARIAH-CORE WP5 plan describes these aims. In 2016 we have decided to exchange the original, tool-oriented approach for a modular one, whereby the key components of the original tools are rebuilt and offered as a stand-alone mini-tool. In addition, we provide the original tools in the form of 'recipes' that combine the functionalities of the mini-tools into a comprehensive interface, that allows for more complex research. This modular approach facilitates maintenance (and thus is more sustainable) and provides researchers with a better insight into the workings of the tools (exposing rather than blackboxing them, so allowing researchers to take a tool-critical perspective). Finally, the Media Suite provides APIs for more advanced users to directly query the raw data. Below, we list the main goals of our agenda, organized along the stages of a typical media scholarly research project.

1. **Obtaining data and tools**: We aim to provide tools to **browse and search for data** in the CLARIAH infrastructure that researchers may want to use in their research. Since most of the audiovisual datasets in CLARIAH are restricted because of copyrights and ethics, these tools have to be made available in a **closed, authenticated environment**. Additionally, we need **facilities to incorporate an existing dataset** that is not yet part of the CLARIAH infrastructure into the Media Suite. This requires a service for registering, converting and describing selected resources, including a workflow and documentation that assists collection managers and researchers in adding, curating and describing their data. Also needed is a clear **interface that guides researchers towards the tools** that best match their research goals, as well as guidelines that basic CLARIAH Media Suite-compatible software must meet.
2. **Search**: There must be facilities for a researcher to **select a specific corpus and connect it to a specific search engine and to make it searchable** through existing search engines and applications. It should also be possible to browse, select, group, sort, filter and combine the results and extract them for usage in analysis tools.
3. **Enrichment and annotation of data**: Required is, first, a facility (**user space**) that allows users to create annotations that can be persistently linked to the non-exportable data in the Media Suite. Second, we aim to offer a **range of software which enables researchers to enrich their data with all kinds of annotations** such as speech and speaker recognition software. The software should include tools to carry out both manual and automatic enrichment, as well as tools for manual addition, verification and/or correction of the (automatically enriched) data.
4. **Analysis**: We aim to provide **dedicated tools for audiovisual data analysis** in combination with **tools for analyzing media-related collections** of textual, visual and structured data (RTV program

guides, subtitles, radio bulletin transcripts, newspapers, photo collections, RTV ratings), in part inside the tools provided in the Media Suite (e.g., to analyze search results), in part (for data which may be exported) in existing external analysis tools such as NVivo, ELAN, R, SPSS etc.

5. **Visualization**: Should offer visualization options integrated in the search/analysis tools or as separately callable functions. It should also be possible to combine the visualizations and switch easily between them.
6. **Infrastructure and governance**: Sustainable access to audiovisual data and tools, as well as media-related collections, requires developing an **infrastructure** at the Netherlands Institute for Sound and Vision (NISV), the CLARIAH data center for audiovisual data, and **collaborations** with other relevant stakeholders (CLARIAH WPs, collection owners, developers and users).
7. **Dissemination.** In order to increase uptake, training and dissemination are key. There is an urgent need for training modules that improve digital literacy by focusing on 'data and tool criticism': a reflection on the ways in which the specific nature of digital sources and computational tools for research structure the research process and impact on the results.

4.5.2 Assessment

Ad 1 Obtaining data and Tools: we have made **tools** for collection selection, browsing, exploration and searching (text search, multiple query search) available in a robust format under a single umbrella application with a clear interface: the Media Suite. The tools are available both as stand-alone 'components' and integrated in the 'recipes' that represent the original tools. Currently the AVResearcherXL tool is available as a recipe ('comparative search'); shortly we will make available the DIVE+ Linked Open Data browser and the tools Oral History Today and CoMeRDa. The **data collections** that are made accessible within the Media Suite are registered in the CLARIAH WP5 Data Registry (a CKAN instance); this is also the **facility for adding new, external collections**, for which documentation will be developed in 2018. So far, we have made possible the use of the following data sets in the Media Suite: NISV catalogue of RTV data and (open) news videos; Dutch oral history collections from various owners; the EYE Jean Desmet collection; Meertens Soundbites collections; KB radio news bulletin scrips; objects from the Amsterdam Museum and Tropenmuseum collections. To be added shortly: TV subtitles and program guides (NISV). Currently negotiated: KB newspapers, EYE film collection, EU Screen collection.

Ad 2 Search: we have **made several tools for searching available** to researchers: collection selector (allows users to select which collections to use in search/analysis tools and recipes), text search (for metadata records and transcripts; includes full-text, field-restricted, faceted, date-restricted search, similar document search, cross-collection search), multiple query search (allows users to search with multiple queries in a comparative search results view, i.e. AVResearcherXL tool).

Ad 3 Enrichment and annotation of data: We have implemented a basic version of the **user space**, where users can select, bookmark, annotate datasets and save their results. We implemented a **baseline media annotation tool** with the most important annotation features for scholarly research, including external vocabularies that aid consistent description (GTAA, DBpedia, UNESCO) and external knowledge bases that allow for linking selected audio or video fragments to related content (Wikipedia, Europeana). Currently we are working on implementing a SaaS solution for **automatic speech recognition and crowdsourcing**.

Ad 4 Analysis: We realized **playout** of audio and video with segment selection and annotation and are currently implementing the DIVE **Linked Open Data Browser** (for semantic browsing and exploration of AV collections, including AV playout) and **Linked Narrative Builder** (for constructing and annotating narrative paths). In 2017 and 2018 we will work on **integrating automatic text analysis services** and **baseline automatic audiovisual analysis services** (transcription, speaker identification, face recognition, shot detection, emotion recognition, color analysis).
Ad 5 Visualization: we implemented **various visualization tools** (timeline viewer, word cloud viewer, thumbnail viewer) that can be used as stand-alone services and are also integrated in the recipes, and in 2017-2018 will provide each of these tools with **documentation** that explains how the visualizations are generated (tool criticism).
Ad 6 Infrastructure and governance: we integrated the components of the original tools by implementing these in a **sustainable infrastructure** at NISV. We also implemented a **single sign-on solution** for managing access to privacy/IPR restricted content, as well as a method for the play-out of restricted content. We developed a generic model for annotation (cross-domain, cross-media) and implemented a basic, closed-environment user space that does not go beyond the Media Suite domain. Currently we are establishing **agreements with various stakeholders** on sustaining the Media Suite and adding more generic datasets to it and engage in various initiatives to **foster cross-WP CLARIAH collaboration**.
Ad 7 Dissemination: we **trained content-owners** to maximize the use of their collections for scholarly research by showing the benefits of central collection registration and providing strategies for preparing their collections (metadata, play-out) to allow their use in an research infrastructure. We developed **courses, workshops and lectures** for BA, MA, RMA students, PhDs and media scholars and oral historians (among others via RMeS and Huizinga). We delivered **papers and workshops** at international conferences and realized **publications** in the area of Digital Humanities, Media Studies, Information Studies. In 2017-2018 we will contribute use cases that employ AV data and tools to the DARIAH-EU project Visual Media Repository.

5 Appendix Detailed Reports

5.1 Detailed Report WP1 Overall management

5.1.1 Assessment per task/subgoal

5.1.1.1 Governance

The governance has been set up and consists of the following bodies:

- Board
- Executive Board
- Supervisory Board ('Raad van Toezicht')
- International Advisory Panel

The executive board is a subset of the board.

The tasks and responsibilities of the governance bodies has been defined in by-laws ('huishoudelijk reglement', CC 15-215, Version 2.3, July 17, 2015) and approved by the Supervisory Board.

The initial composition of the governance bodies can be found in the fact book.

Several changes have taken place since the initial composition:

- Jan Blommaart (IBM) withdrew from the Supervisory Board early in 2016.
- Dr. Henk Harmsen (DANS) withdrew from the board and was replaced by Andrea Scharnhorst (DANS) (March 2016)
- Prof. Dr. Jan Luiten van Zanden (UU) withdrew from the Board and joined the Supervisory Board. Dr. Richard Zijdeman (IISH) replaced him in the Board. (January 2016).
- Prof. Dr. Sally Wyatt (KNAW, UM) withdrew from the Supervisory Board because her appointment as member of the permanent NWO committee for large scale infrastructures created a conflict of interest.
- Prof. Dr. Franciska de Jong (CLARIN ERIC, UU, EUR) joined the Supervisory Board (Mid 2015).
- Jan Müller (NISV) joined the Supervisory Board (end of 2015) and withdrew from the Supervisory Board in November 2016 because of a conflict of interest.
- Dr. Sebastian Drude (MPI, CLARIN ERIC) withdrew from the Supervisory Board in March 2017 because of a change in his position and consequent move abroad.
- Prof. Dr. Hans Bennis (Meertens) withdrew from the Supervisory Board in December 2016 after his retirement at Meertens Director and his appointment to General Secretary of the Dutch Language Union.
- The Board has decided to invite one more member to the IAP, and the invitation has been sent.
- The Executive Board proposes to add a member to the Supervisory Board, and will make a proposal to that end in the first Supervisory Board Meeting.

The Board is supposed to meet at least 5 times a year. In 2014 (before the start of the project), there were 2 meetings, in 2015 there were 4 meetings, and in 2016 5 meetings.

The Executive Board is supposed to meet once a month. In 2014 (before the start of the project) there were 4 meetings, in 2015 there were 11 meetings, and in 2016 9 meetings.

The Supervisory Board is supposed to meet once a year. In 2014 (before the start of the project) there was 1 meeting, in 2015 there were 2 meetings, and in 2016 1 meeting.

Office The CLARIAH office was set up with Jan Odijk (UU) as CEO, Drs. Arwin van der Zwan (Huygens ING) as project secretary, and Dr. Patricia Alkhoven (Meertens) as Coordinator for external cooperation. Drs. Arwin van der Zwan fell ill early in 2015, causing problems due to understaffing. These problems were partially and for a limited period remedied by the CLARIAH-CORE CEO and PI (e.g. for preparing and making minutes of meetings) and by hiring Erica Renckens (former CLARIN-NL project secretary) for selected tasks. After a while, Annejet Landman (Huygens ING) was hired to carry out a part of the Project Secretary's tasks. Since early 2017, Arwin van der Zwan has started working again, though in a very limited way (2 hours per day), and it remains to be seen whether and when he will return in full. Pending his illness, his temporary replacement will be increased in size.

5.1.1.2 Relations with CLARIN and DARIAH

5.1.1.2.1 CLARIN

CLARIAH director Jan Odijk was the national coordinator for CLARIN and remained so in the CLARIAH-CORE project. He attends the monthly CLARIN National Coordinator's Forum (NCF) meetings (most of them via skype, but some face to face) through which the CLARIN members organize activities (e.g. the yearly CLARIN Conference), exchange information amongst them and with the CLARIN ERIC Board, and provide input to the board in CLARIN ERIC policy matters.

The NCF also acts as Programme Committee for the yearly CLARIN Conference, and this programme committee was chaired by Jan Odijk in 2014. Jan Odijk was also the editor of the *Selected Papers from the CLARIN 2014 Conference* (Odijk 2015).

Jan Odijk also attended the 2015 and 2016 CLARIN ERIC General Assemblies as expert advisor to the Netherlands delegate Drs. Alice Dijkstra (NWO).

Dr. Arjan van Hessen and Dr. Patricia Alkhoven have been active in user involvement workshops associated to CLARIN Conferences.

Dr. Arjan van Hessen is member of the CLARIN ERIC User Involvement Task Force since March 2017, and is co-organizer of the User Involvement Workshop to be held in Helsinki, Finland, 8-9 June 2017.

CLARIN ERIC regularly organizes workshops (e.g. in the context of the CLARIN PLUS project). The Netherlands has sent many delegates to these workshops. For an overview, see appendix.

5.1.1.3 DARIAH

CLARIAH Supervisory board member Henk Wals (IISH) is the Netherlands national coordinator for DARIAH. Henk Harmsen (DANS) was a CLARIAH board member and was the Chief Integration Officer in DARIAH, active in the Virtual Competency Centre (VCC) Scholarly Content Management. He resigned from both positions in 2016. Andrea Scharnhorst (DANS) is currently one of the heads of this VCC and a CLARIAH board member. Connie Kristel (NIOD) was one of the CLARIAH Directors until 2016.

CLARIAH was well represented at the DARIAH Annual Event in 2016 in Ghent, and will be well represented in the 2017 DARIAH event in Berlin.

CLARIAH has described its contributions to the DARIAH infrastructure in forms set up for this purpose by DARIAH, and these were approved by DARIAH.

5.1.1.4 Support for Other projects

CLARIAH has given support to 19 project proposals. This set includes both national (14) and international ones (5), and both research (12) and infrastructure projects (7). The form of the support differed per project but included support letters, information on the use of CLARIAH technology, on embedding project results into the CLARIAH infrastructure, recommendations of knowledge utilization, etc. etc. An overview of the project proposals involved is available, but it is confidential since it contains names of applicants and project proposals that were not awarded funding. Some of the project proposals were awarded funding and we plan to set up dissemination activities around them showing the relevance and importance of CLARIAH for these projects (through blogs, short movies etc.) in 2017.

5.1.1.5 eHumanities.NL

eHumanities.NL[1] is a national platform that brings together expertise and research in the development and use of digital technologies in the humanities and the social sciences. It unites all Dutch Universities with a humanities faculty and a number of other relevant research institutes. Lex Heerma van Voss was member of the eHumanities.NL network Executive Committee ('programmaleiding') on behalf of CLARIAH from March 2017 through January 2017. As of January 2017, Jan Odijk has replaced him in this position.

5.1.1.6 Roadmap

The NWO Permanent Committee for large scale research infrastructures has made an overview ('landschapsanalyse') of large scale research infrastructures and presented this on May 24, 2016.[2]

The committee put CLARIAH on the draft Roadmap. It then proposed that candidate projects from the Humanities should be clustered under CLARIAH. CLARIAH was asked to coordinate such clustering activities. Some organisations, which had explicitly presented themselves as research infrastructures,[3] were incorporated and presented together with CLARIAH as CLARIAH-PLUS in the final national roadmap document.

[1] http://www.ehumanities.nl/
[2] http://onderzoeksfaciliteiten.nl/
[3] KB, NISV, NA, TLA and the CLARIAH Media Suite

A sounding board committee was created by the committee in order to 'act as a counterweight for the coordinating role that CLARIAH has for NWO'.[4] It has met twice (October 7, 2016 and April 21, 2017). Its members are Lex Heerma van Voss (Huygens ING: history), Henk Wals (IISH: history), Jan Odijk (UU: (computational) linguistics), Peter Doorn (DANS), Wido van Peursen (VU: religion sciences), Piek Vossen (VU: computational linguistics), Julia Noordegraaf (UvA: media studies), Arianna Betti (UvA: philosophy), Ellen van Wolde (RU Nijmegen: religion studies), Erik Kwakkel (UL: mediaeval studies and book sciences), Olaf Andersen (NISV), Steven Claeyssens (KB), Remco van Veenendaal (NA), Rosa van Santen (NISV), Sebastian Drude (CLARIN ERIC, before that TLA/MPI: linguistics), Sven Dupré (UU: art history), Anita Hopmans (RKD – Netherlands Institute for Art History), and Martijn van Leusen (RUG: archeology).

It is important to obtain wide support for a proposal in this committee, and we are working towards this.

CLARIAH was put on the national roadmap for large scale research infrastructures onderzoeksvoorzieningen by NWO on 13 December 2016.[5] This implies that we can submit a proposal for a project in the next call. This call was published in December 2016.[6] The submission date is June 1, 2017. If successful, a new project can start immediately after the current project: 1 January 2019.

5.1.1.7 Calls

A budget of 1 million euro was reserved for organizing calls for projects. In the original plan a call was foreseen for research pilot projects. The aim is to test components of the infrastructure by carrying out a small research project. This call could be launched only after CLARIAH-CORE had created enough functionality to be tested. For this reason it was decided to launch this call in September 2016. The project proposals that are awarded funding can then start as of March 2017 and will all be finished by mid 2018, leaving some time to implement or act upon findings of the projects. All in all, CLARIAH launched two calls for proposals: the ADAH-call together with the NL eScience Centre and the Call for Research Pilots. We discuss each of them in turn.

5.1.1.7.1 ADAH Call

Early in the CLARIAH-CORE project NL eScience Centre proposed CLARIAH-CORE to organize a call together. In this call, 3 projects can be awarded funding, 100 k euro per project financed by CLARIAH-CORE, and 150k euro per project by NL eScience Centre, in total for 300K euro + 450K euro = 750k euro. CLARIAH-CORE and NL eScience Centre elaborated a call text and evaluation procedure in which the interests of both parties are secured. Unfortunately, the CLARIAH Board rejected this proposal on March 11, 2015. The board considered a call with the eScience Centre too technical in nature to involve the Humanities in it. There was also fear of insufficient control over the activities of the eScience computer scientists (based on bad experiences in other projects with NL eScience Centre). It was also doubted whether the added value of cooperation with NL eScience Centre would be large enough. The board

[4] Concept report 1st Meeting of the Humanities Sound board committee of 7 November 2016, item 1b., p.2.
[5] http://www.nwo.nl/binaries/content/documents/nwo/algemeen/documentation/application/nwo/permanente-commissie/roadmap-grote-onderzoeksfaciliteiten/Roadmap+grote+onderzoeksfaciliteiten.pdf
[6] http://www.nwo.nl/financiering/onze-financieringsinstrumenten/nwo/nationale-roadmap-grootschalige-onderzoeksfaciliteiten/nationale-roadmap-grootschalige-onderzoeksfaciliteiten.html .

concluded that CLARIAH is happy to cooperate with NL eScience Centre, but a joint call is the not most effective tool for this.

In February 2016, discussions with NL eScience Centre on organizing a joint call were reopened, and this resulted in September of 2016 in an agreement to jointly organize a call. Most of the earlier objections were taken away, in part by increased confidence in each other, in part by explicit formulations of specific requirements in the call text. The call ('Accelerating Scientific Discovery in the Arts and Humanities (ADAH)') was published on the NL eScience[7] and CLARIAH[8] websites on October 13, 2016.[9] An information session for this call was organized on November 1, 2016. Upon request by the applicants, meetings to discuss a proposal with eScience engineers were held. The number of submissions is 23. Some of these did not fully comply with the requirements and got the opportunity to submit an updated proposal within a limited amount of time. All resubmitted on time except for one. The evaluation procedure is currently ongoing. The final decision on awarding of the proposals is expected in June 2017.

The procedure was prepared from the CLARIAH side by Jan Odijk and Lex Heerma van Voss, with support from Sjef Barbiers and Antal van den Bosch. Though originally Jan Odijk was involved in organizing the call, he got involved in one of the project proposals submitted due to obligations of an independently financed project. Because of this, Jan Odijk has withdrawn from organising this call to avoid any possible conflicts of interest, and his tasks have been taken over by Lex Heerma van Voss.

5.1.1.7.2 Call for Research Pilots

The call for research pilots was planned in the original project proposal. In the project proposal, the budget for this was set to 1 million euro. Because of the ADAH call, this was reduced to 700 k euro, which would allow funding of app. 12 projects. The call was launched in September 2016.[10] An external and independent ad-hoc evaluation committee was formed. 28 submissions were made. Out of these, after evaluation and ranking, 16 projects were awarded funding. Since the ad-hoc evaluation committee recommended spending more money on this call than originally budgeted because of the high quality of the submitted proposals, the board decided to take 200 k euro from the reserve and assign it to the budget for the research pilot call, so that 16 projects can be funded. A detailed report on the whole procedure followed for this call, from publication of the call text to the recommendations by the ad-hoc evaluation committee on the evaluation and ranking of these project proposals, is available (CC 17-016). The awarded projects presented themselves briefly for the first time on the CLARIAH Toogdag 2016 held in Amsterdam on March 10, 2017. It is striking how much cross-WP and thus cross-discipline cooperation there is in the projects: in 10 out of the 16 projects people from different WPs work together, in many different combinations. Of course, WP2, which provides generic infrastructure facilities, occurs in multiple projects, but there are also combinations among the other WPs: WP3 and WP4, WP4 and WP5.

[7] https://www.esciencecenter.nl/redactional/2016-adah-project-call
[8] http://www.clariah.nl/projecten/adah-project-call
[9] The call consists of a call text, an application form, a document specifying funding conditions specific to this call, and a document on the NL eScience Intellectual Property Policy.
[10] https://www.clariah.nl/en/projects/research-pilots/the-call

When evaluating the procedure, the board concluded that some of the proposals that had been funded did not so much evaluate existing infrastructure but aimed at building additional infrastructure functionality. This is in itself useful, and it is both understandable that applicants that want to add functionality use the call in this way, and that reviewers and members of the ad-hoc committee are seduced by this. Nevertheless the board concluded that in future calls (i.e. in future projects) we should try to define the notion 'research pilot project' more rigidly and give compliance with this rigid definition a more prominent role in the evaluation procedure.

5.1.1.8 Education & Training / Dissemination & Outreach

Website Before the CLARIAH program was granted, a website was built by volunteers for various dissemination activities. Once granted, it was decided to build a more professional site, suitable for both normal computer screens and mobile devices. One of the biggest "questions" of the new site was: for whom to make it. The website has the following goals:

1. Dissemination of the program, the results (data, tools and standards) and (scientific) papers for both an academic and a more general audience.
2. Creating the CLARIAH community with humanity scholars, computer scientists and data owners
3. Platform for exchanging ideas, problems and progress between the participants of the different work packages
4. Dissemination of upcoming CLARIAH and CLARIAH-related events, blogs about these events and more. Moreover, forms were "created" for event subscriptions, cost declarations and grant requests.

It turned out that it was not possible to satisfy all these wishes simultaneously. So no.3 (exchanging ideas) was moved to 5 different Basecamp-sites.

In February 2016, the current website was up and running.

Brochure Together with a new logo and a restyled website, a new CLARIAH brochure was published in the Summer of 2016, replacing the old brochure with the seed-money projects.

Newsletter Simultaneous with the new website, a MailChimp based newsletter has been setup, sending (with an average of 6 weeks) newsletters to 380-subscribers. The effect of newsletters for the amount of subscriptions for events is noticeable. Besides the CLARIAH newsletter, there is a WP5-Media Studies newsletter (72 subscriptions).

Social Media A twitter account (@CLARIAH_NL) and a Facebook page complete the dissemination instruments of CLARIAH. The tweets are well followed by other infrastructure programs, retweeted by different scholars and organisations and showed on the website. Facebook (https://www.facebook.com/clariahinfra/) is used as well but does not have a lot of impact.

Videos. As in CLARIN (2009-2015), we have decided to make videos of impressive results (projects, software, events) in order to increase the visibility for a no-expert audience. The videos can be found at the CLARIAH website (CLARIAH-video's) and (CLARIN videos). Once the 4 work packages will have mature and showable software, new videos will be made.

Other infrastructure program. There is (of course) a good collaboration and mutual exchange of information with both the CLARIN ERIC and the DARIAH ERIC. Moreover, we collaborate with a group of 6 ERIC's from the HSS. Goal of this collaboration is to exchange ideas about "including the HSS-scholars in the infrastructure program". How to prevent an infrastructure that nobody uses?

Events

2015

- **Kickoff** On March 13, 2015 there was the official kickoff of CLARIAH. The event was visited by more than 130 people and was organised at the Netherlands Institute of Sound and Vision in Hilversum. More information can be found here: https://clariah.nl/en/events/kick-off
- **CLARIAH days** Each year we organise a so-called CLARIAH-day (spring) and a CLARIAH-techday (autumn). The general CLARIAH-day is meant for a broader, non-expert community (other humanity researchers, heritage-specialist, policymakers and funders). On such a day, an overview is given of the state of the project, the activities of the previous year, the upcoming activities, what went ok and what went wrong (and why).

  The Techday is meant for ICT and DH-specialist from inside the CLARIAH community. It is a day where "we" inform each other about progress, stagnation, problems and ways to solve them. Although the day is open for everybody interested, the majority of the audience consists of CLARIAH-scholars and CLARIAH-technicians.

2016 Last year two major events and some smaller ones were organised. In January, we had our first CLARIAH day at the RCE in Amersfoort. It was a successful day with keynote speakers from abroad (Charley Mörth, Austria) and from Utrecht (Joris van Eijnatten). For a blog see: http://www.clariah.nl/nieuw/blogs/339-clariah-dag-2016

For the launch of the Research Pilot Call (sept 2016) an information meeting was organised.

On October 7, the second CLARIAH day of 2016 was organised. It differed from the first one that was targeted on the a more general audience. The second CLARIAH-day was a CLARIAH-tech-day: focusing on an inter-collegial exchange between technicians, working in the CLARIAH realm.

In October 2016, CLARIAH was present with some demos at the annual National eScience symposium.

Courses. In January 2016, CLARIAH supported (with scientists and money) the first coding for the humanities initiative: Making sense of Digital Spaghetti. It was a collaboration between the Humanity department of the Utrecht University, the University College Utrecht and CLARIAH. The course 2x 5 working days) took place at the UCU and a farm outside the city. According to the "votes" of the participating students it was a huge success (8.6 out of 10). An evaluation report of the course can be found here (in Dutch).

The UvA (Jan Don, Fred Weerman) asked information about the course: willing to do a comparable course in Amsterdam.

For 2017, we are setting up a (kind of) copy of the Spaghetti course together with the EUR (Humanities and Social Sciences).

Projects. CLARIAH-SEED had already initiated the development of a Digital Humanities Course Registry,[11] which was originally called DODH[12] standing for **Dutch** Overview of Digital Humanities. This Course registry was received very enthusiastically by DARIAH, so that it was turned into a **DARIAH** Overview of Digital Humanities, with contributions from multiple countries. CLARIAH-CORE decided to extend this with a registry for Digital Humanities projects[13], which has been created and which is being kept up to date and monitored by a student assistant during the course of the CLARIAH-CORE project. CLARIN ERIC director Franciska de Jong informed us that CLARIN and DARIAH are currently cooperating at the European level to create a new version of the course registry, which was launched in April 2017 at the DARIAH Conference. Utrecht University (DH-lab) will keep-up the course registry at least for the coming year.

Conferences. CLARIAH and the Utrecht University succeeded this year in bringing to Utrecht two Digital Humanity conferences: DHBenelux in 2017 and the big DH2019. The DH2019 will be organised by Utrecht University, CLARIAH, NISV and the National Library of the Netherlands.

Course Task Force. The CLARIAH Course Task Force has been convened twice since its foundation in 2015 (18 June 2015 and 7 April 2016). Each university or humanities institute in The Netherlands is represented as well as Belgium (KU Leuven). Main topics of discussion were the need for exchanging information about DH courses and the possibility to exchange course materials. Using Basecamp, it is now possible to store course materials and keep in contact with all members. In 2017 we will visit each university to inform about CLARIAH, demonstrate tools and collect ideas and feedback for cooperation (CLARIAH on Tour). Goal is to better involve students in CLARIAH and to embed CLARIAH data and tools in the curriculum of each university or research institute.

Contributions at conferences. Many researchers and developers from the CLARIAH community contributed to International conferences and symposia, such as DH2016, CLARIN Annual Conference and LREC.

[11] https://dh-registry.de.dariah.eu/ and http://www.clariah.nl/projecten/dodh/395-dodh#course-registry-2
[12] http://www.clariah.nl/projecten/dodh/395-dodh
[13] http://www.clariah.nl/projecten/dodh/395-dodh#project-registry-2

Further, the CLARIAH community held many presentations, posters etc. for national symposia and conferences. (See, attachment for overview of activities).

Non-CLARIAH events. Often we get requests for support for activities, conferences and events that are CLARIAH-related but are not organised from within CLARIAH. If relevant for CLARIAH, these activities are supported with money, publicity and sometimes a lecture in exchange for publicity about CLARIAH, a blog and access for a couple of CLARIAH members. Some examples: the CLIN-conference (2015, 2016), THATcamps (2015, 2016).

Support for Scholars. Scholars who want to visit a workshop/conference/meeting that is clearly CLARIAH-related can get money for the fee and travel. In exchange we ask them to write a blog about their experience at the event. The blogs are posted at the website. The support can be asked via a form at the website. (15 requests last year)

Publications, conference contributions, presentations, lectures. For an overview see the Fact Book.

5.1.1.9 Brain Gain

We reserved money in the budget for 'brain gain', i.e. creating opportunities for top quality researchers who work abroad to spend time in the Netherlands, share their knowledge and expertise and make use of the CLARIAH infrastructure. So far, no activities in this respect have been organized.

5.1.1.10 IPR

We reserved money in the budget for costs related to intellectual property rights, e.g. legal advice. Though IPR causes problems, especially for audiovisual material and recent copyrighted texts (novels, journals, newspapers), which is available for research only in limited amounts. Though a working group for IPR was set up at the request of WP5, its members currently are limited to two persons (Jan Odijk (WP1) and Eva Baaren (WP5)), though Henk van den Heuvel and Nelleke Oostdijk (both RUN) from WP3 have offered support and it has not done any work yet.

5.1.1.11 Athena

As a follow-up of a CLARIAH-SEED funding, a proposal (Athena) was submitted by Jan Luiten van Zanden to carry out a project in which the CLARIAH core disciplines are integrated. ATHENA (www.athena-research.org/) is an interdisciplinary project to develop a data portal that will hold information on historical context of human – nature relationships for a broad variety of plant and animal species and the landscapes and ecosystems they live(d) in. Following the three leading disciplines of CLARIAH, i.e. language studies, media studies and socio-economic history, ATHENA will combine different data types: text, media objects (images) and (semi-) structured data respectively. The project will develop a data portal where mutually linked historical databases holding information on flora and fauna can be approached and data can be combined, integrated and analysed. The aim of the ATHENA project is not only to bring different research fields from within the humanities together, but to look beyond. By integrating the fields of history, archaeology and ecology in an innovative research environment, scholars from multiple scientific disciplines will be able to study human-nature relationships in meaningful ways.

A budget of 150,000 euro was requested, and the Board agreed to fund this project. WP3 had reserved some money for cross-discipline activities and made 10,000 euro from this budget available for this project for cooperation between WP3 (VU) and WP4 to extract information on fauna and flora from textual documents. The remaining 140,000 euro was drawn from the reserve kept apart in the CLARIAH-CORE budget.

What has been done. The project started with the front-end development (provided in html) to ensure that the envisioned concept was realized in the final product. Parallel to that, database and hosting requirements for linking the interdisciplinary datasets were defined. Moreover, follow-up meetings with data providers were held to ensure the requirements could be achieved (Month 1-6).

The next step was acquiring and harmonizing the datasets for which no API was available into a single database structure. The different datasets were linked based on species entities. To that end, the taxonomic dataset was compiled and enriched with name variations (Month 4-8). Finally, the back-end structure, meeting the hosting and database requirements, was chosen and developed. Moreover, the first datasets (modules) have been implemented (ongoing).

What needs to be done. Most datasets will be implemented within the back-end structure within the current project time frame. Unfortunately some datasets cannot be implemented before that time because of data availability/accessibility issues, while others are linked with a temporary fix as no API is available yet. Moreover, the newly created database structure will be hosted at ING Huygens. The database transfer, however, will take more time because of the conversion to RDF.

The project has requested (and received) extensions in time and is targeted to finish by the end of 2017.

5.2 Detailed Report WP3 Linguistics

5.2.1 Introduction

This document describes the status of CLARIAH Work Package 3 (WP3) after the first two years (2015-2016) of the project. It also reflects on this status and indicates where revisions of the plan are under discussion, desired, needed or have already been made. We start with a global overview (section 4.3.2), which is followed by a more detailed overview per partner and task in section 4.3.3.

5.2.2 General Overview

In this section, we first describe the overall management of WP3 (section 4.3.2.1), followed by a brief description of events organised or co-organised by WP3 (section 4.3.2.2). We next describe the relations WP3 members have to related international projects (section 4.3.2.3). Finally, this is followed by a global overview of the activities and status of each partner involved (section 4.3.2.4).

5.2.2.1 Overall Management

The overall management is carried out by Sjef Barbiers (formerly Meertens Institute and Utrecht University, currently Leiden University), Daan Broeder (Meertens Institute), and Jan Odijk (Utrecht University). They each cover different aspects and have a different focus: Sjef Barbiers: scientific leader, overall management and policy; Daan Broeder: technical infrastructural matters; Jan Odijk: elaboration of specifications and overall organisation.

Originally we planned to partition WP3 into 4 themes, each with a theme leader. In practice however, governance has been carried out directly with the organisations involved, not along the lines of themes.

At the end of 2014, the global plans for WP3 as described in the proposal were elaborated in detail by the main WP3 theme leaders (Van den Bosch 2015, Brugman 2015, Kemps-Snijders 2015, Vossen 2015). These were integrated, together with known requirements and desiderata from other sources (e.g. from the CLARIN-NL project) into an overall plan for WP3 in (Odijk et al. 2015). In this plan, all steps in a typical research project are sketched, and it is indicated which research infrastructural means are desired or required in each step. Parts of these infrastructural means are already available from CLARIN-NL or other projects, so the plan indicates which functionality is to be added, to be extended, improved, integrated, or made interoperable. For each of these aspects it defines tasks and assigns them to (teams of) organizations involved in CLARIAH WP3.

On the basis of this plan, the available budget was distributed over the tasks and the partners (Odijk 2015e), and commitment letters (‘toezeggingsbrieven’) were sent out by Meertens Institute, the leading organization (‘penvoerder’) of WP3 to the partners involved.

Originally, a large part of the activities for federated search were assigned to Meertens Institute, since they had worked on this topic also in CLARIN-NL and had built up a lot of knowledge and expertise in this area. However, Meertens indicated not to be willing to continue with federated search, preferring instead aggregated search. Most of the federated search tasks were then assigned to INT, which has ample experience with search in corpora and lexica, and is therefore an excellent alternative.

A new partner, Taalmonsters ,[14] which was not part of the original consortium, was requested to work on the frontend of the OpenSoNaR search application, in close cooperation with INT.

Monthly progress meetings are held, usually via skype, but occasionally face-to face. These are intended to coordinate the work among partners, exchange information, monitor progress, discuss and address problems, etc. Every partner submits a written progress report on the current state, which is discussed at the meeting. There are different views within the WP on how the development of software should be planned and monitored, and on what type of documents should be used to support such planning and monitoring (e.g. tabularly organized , or more narratively textual), so the reporting within WP3 has not been uniform, and we are still trying to find an optimal form for such planning and monitoring.

The management team held regular meetings during the whole period, and intensified this since the end of 2016: a weekly skype meeting is held to discuss the status and plans.

Several meetings were held with specific partners, to elaborate their plans, discuss problems and decision points, and to coordinate the work with the work by other partners. Other meetings were held to discuss specific topics. See the appendix for an overview.

Daily communication runs mainly via Basecamp and e-mail. A chat system has been set up (especially for the developers to communicate among them) but its use is very limited.

Cooperation among the partners is essential for some topics, e.g. between Nijmegen and VU for NAF-FoLiA conversions, between Utrecht and Groningen for shared functionality for the closely related search applications they work on, etc. Overall such cooperation has been excellent, though occasionally (and partially related to the shifted role of theme leaders) there were problems due to miscommunication and different expectations of the partners involved. These, however, have been solved immediately when they became clear.

CLARIAH WP3 works, given its embedding in the CLARIN infrastructure and its origin in the CLARIN-NL project, on a distributed infrastructure. Its distributed nature brings with it the challenge to ensure that an infrastructure is created that is perceived as a coherent infrastructure rather than as a fragmented landscape of unconnected tools and initiatives.[15]

This requires the full attention of WP3, and being half way through the project we should take the opportunity to reevaluate goals and objectives and realign tasks towards a common goal. Important aspects that should be considered in this context include more focus on interoperability, more focus on same look and feel of distributed functionality, branding, a single entry point for the infrastructure, and, for selected cases, centralization of functionality.

WP3 will consider this problem in the coming months and, where desirable and possible, adjust its goals.

[14] http://www.taalmonsters.nl/

[15] And this problem holds of course more generally for the CLARIAH infrastructure as a whole.

5.2.2.2 Events Organised

WP3 organised or co-organised several workshops on specific topics and attended relevant workshops organized by others within CLARIAH. Concrete examples are the international workshop on linked data for linguistic resources, the WP3 workshop on interoperability, and the workshop on linked data in CLARIAH.

VU organised and hosted the WP3 workshop on interoperability on Sep 16. 2016,[16] in which all partners of WP3 and some representatives of WP2 were present to discuss how to tackle the problem of interoperability. It resulted in a concrete list of actions assigned to the participants with a target date.

On February 6 and 7 2017, CLARIAH WP3 organised a workshop to discuss the application of Linked Data for linguistic research.[17] The workshop invited presentations from a number of foreign experts and from representatives from CLARIN centers that had acquired some experience using Linked Data in their projects. The workshop had a pragmatic approach, discussing pros and cons of Linked Data usage for a number of current linguistic research topics that were introduced by linguists from WP3.

VU was co-organiser and host of the CLARIAH Linked data workshop.[18] This workshop was intended to discuss the use of Linked Data as technology for connecting data across the different CLARIAH work packages (WP3 linguistics, WP4 structured data and WP5 multimedia).

The goal of the workshop was twofold. First of all, to give an overview from the 'tech' side of these concepts and show how they are currently employed in the different work packages. At the same time we wanted to hear from Arts and Humanities researchers how these technologies would best suit their research and how CLARIAH can support them in familiarising themselves with Semantic Web tools and data. The workshop was very successful, not only because of the knowledge transfer and the lively discussions, but also because of the feeling it created that we are all working on the same project. A follow-up workshop with a more hands-on setup is scheduled for 1 May 2017.

5.2.2.3 Relations with International Infrastructure Projects

Several WP3 members are also active at the European CLARIN level. In particular:

- Daan Broeder: member Centre Assessment Committee, member CLARIN ERIC standards committee, member Metadata Curation Task Force
- Marc Kemps-Snijders: Member Standing Committee for CLARIN Technical Centres
- Jan Odijk: member Metadata Curation Task Force, member CLARIN ERIC standards committee
- Menzo Windhouwer: member CLARIN ERIC standards committee

Daan Broeder is heavily involved in EUDAT2020[19] by his work for this project on community requirements and engagement (WP4). In addition, Meertens is part of the EUDAT Collaborative Data Infrastructure (CDI).

[16] http://www.clariah.nl/en/events/calendar/30
[17] http://www.clariah.nl/en/new/blogs/591-linked-data-for-linguistic-research
[18] http://www.clariah.nl/en/new/blogs/549-clariah-linked-data-workshop
[19] https://eudat.eu/

Marc Kemps-Snijders is involved in the Parthenos project[20] as task leader of Task3 in WP6 (Services and Tools).

5.2.2.4 Global Overview of the Status By Partner

5.2.2.4.1 INT

The INT has started most of its work in WP3 considerably later than foreseen. This was due to reorganisation and redundancies, an IT outsourcing project, and a project to transfer the materials of the Dutch-Flemish Human Language Technology agency from the Dutch Language Union to the INT. At the end of 2016, preparations were made for carrying all out all the remaining work in 2017 and 2018.

5.2.2.4.2 Meertens Institute

The Meertens Institute has primarily focussed on those tasks for which clear synergy effects could be achieved given the roadmaps of ongoing activities of projects. OpenSKOS activities were aligned with CLARIN Plus and NDE roadmaps, with the intention of providing a broader joint platform for future collaboration. The development activities with respect to Mtas (see section 1.1.3.2.8) were aligned with Nederlab and the Nederlab roadmap to 1) deliver the necessary functionality in a timely manner to allow Nederlab developments to benefit from the outcome and 2), in case the Mtas result was incapable of producing the desired result, allow the Nederlab project to identify and implement alternative routes.

With respect to T02, local version of VLO, improvements of the CLARIN VLO functionality have prompted a re-evaluation of this task. Where, at the start of the CLARIAH project, it was felt that the usability of the CLARIN VLO was considered suboptimal for CLARIAH purposes it's further development in the CLARIN Plus project showed sufficient progress.

The crowdsourcing task, T09, was divided into two equal subparts: integration of Meertens's survey environment and integration of crowdsourcing transcription environments, previously developed in several non-CLARIAH projects. While the first part is proceeding as planned, further development of crowdsourcing transcription environments was taken up as part of a more strategic discussion at the KNAW level on the future of crowdsourcing applications. Awaiting the outcome of these discussions, which are expected to have a wider impact on the community, this part of the project was postponed.

Remaining tasks that have not been filled in yet are to be re-evaluated in collaboration with WP3 coordinators to ensure that goals and deliverables reflect the current state of affairs and contribute towards common CLARIAH WP3 overall goals.

5.2.2.4.3 Radboud University Nijmegen (RUN)

The Radboud University team has been growing according to plan from the onset of the project to six persons, with Ko van der Sloot (KvdS) and Maarten van Gompel (MvG) in the largest capacities, and with part-time contributions from Antal van den Bosch (AvdB), Henk van den Heuvel (HvdH), Nelleke Oostdijk (NO), and Martin Reynaert (MR). Overall, activities have progressed faster than planned.

[20] http://www.parthenos-project.eu/

In general, with respect to the development of software we would like to note that the way that the team operates is not entirely in sync with the original plan with milestones. The key difference is that most tasks have started before their planned start date. We are continuously working on our software stack, and have started tasks as soon as developers were in place (rather than starting them according to our plan in a non-overlapping order). In essence all our work is in 'on track'. All software is functional, documented, and continuously available (https://github.com/proycon/LaMachine). We welcome tests of recent versions of our software and are completely open to contributions, bug reports, wish lists and suggestions for improvements. All of our software is Open Source (GPL v.3).

This implies that on the one hand all tasks have started and can even be seen as finished to a certain extent, while on the other hand work on tasks that were planned to have ended still continue, due to our intention to continue implementing improvements as time allows. For our developers, this means a type of distributed attention over many different goals, which has their preference over a more piecemeal division of time. On a regular basis, with at least weekly meetings between subsets of team members, changes in activity plans are discussed and agreed upon. Usually, short 'sprints' are defined on one of the software packages, where one or two developers focus more on one goal for a few weeks. Ad-hoc activities may also be caused by bug reports or feature requests from outside the team.

All NLP software mentioned below (Frog, Ucto, CLAM, FLAT, FoLiA libraries, optionally TICCL) is available under one meta-package called **LaMachine**. The code, including instructions on how to build virtual machines (e.g. Vagrant, VirtualBox), a Docker instance, or a virtual environment of LaMachine and all packaged software on the most common operating systems (Windows, macOS, Linux), can be found here: https://github.com/proycon/LaMachine

5.2.2.4.4 Taalmonsters

Taalmonsters has been working on WhiteLab Version 2, the front-end for the OpenSoNaR+ search application. The work is on schedule and will be finished mid 2017. It is partially dependent on the developments for BlackLab by INT.

5.2.2.4.5 University of Groningen (RUG)

RU Groningen worked on PaQu, an application to search in syntactically annotated corpora. They finished their work and carried out additional unplanned activities, which will be described in detail below.

5.2.2.4.6 Utrecht University

Utrecht University is involved in a number of tasks which will be described below in more detail. Overall, Utrecht is experiencing some delays in the execution of the tasks due to a number of causes: (1) work on CLARIAH WP1 leaves JO too little time to work on WP3 aspects, esp. metadata curation. This will remain so until at least mid 2017 (until after the new CLARIAH-PLUS proposal has been submitted); (2) employee falling ill, causing delays especially on metadata curation for tools. In addition, the main developer for GrETEL will leave the project by June 1, 2017. Replacements will have to be found to continue this work. Fortunately, there are little dependencies on the tasks by Utrecht in WP3, so there are no consequences for WP3 as a whole.

5.2.2.4.7 VU

Except for the late start due to contractual issues, the work has progressed according to plan. We achieved mappings across NAF, FOLIA and ALPINO which are major representation formats for data and tools within the Dutch language community. For lexical representation, we studied international standards, adapted the revised LEMON model of the W3C committee and SKOS for LOD publication of conceptual vocabularies. These formats handle needs at different levels and for different users. We extended the LEMON standard for diachronic and regional usage and tested the architectures for a variety of use cases. The next phase will focus on linking lexical repositories. Furthermore, we developed different semantic taggers for annotating text: professions, entities and emotions. This work will continue in the next part of the project. The remaining funds are sufficient to achieve the objectives. The next phase specifically focuses on the Text2RDF pilots. Up to now, VU established a firm basis for this through various workshops on interoperability and the pioneering work on this topic.

5.2.3 Detailed Overview per Partner and Task

5.2.3.1 INT

INT is involved in a large number of tasks that have been grouped here in a number of clusters around the same topic: Search, Metadata for Tools, GrETEL Upload, AutoSearch, OpenSoNaR, and WebCELEX.

5.2.3.1.1 Search

This task is the combination of the following tasks:

- T06 Internal formats
- T07 Language-related formats
- T12 Extend OpenSkos
- T13 OpenSkos in search
- T35a Federated search token-annotated corpora
- T35b Federated search treebanks
- T35c Federated search lexicons
- T36 Chaining search
- T76 Local search

This task aims to make a large set of linguistic resources available for an integrated content search. These resources belong to three main types: corpora with token-based annotations, treebanks and lexica. CLARIAH distinguishes three levels of accessibility and interoperability:

1. *Local searchability*: the resource is available as a web service, on its own terms
2. *Federated content search*: resources of the same type can be queried as a single resource
3. *Chaining search*: information from heterogeneous sources can be combined, and sequential search workflows can be executed.

An extensive working plan was finalised in November 2016. Work is under way. This very large task will be executed in 2017 and 2018.

5.2.3.1.2 T19 Metadata for tools

Metadata are used to describe resources so that those in the outside world who are interested can find them. The metadata is therefore published and collected by search engines. The INT publishes the metadata according to the CMDI standard as required by CLARIN. CMDI has many different modules for different types of resources. Within the CLARIAH project it has been proposed to transfer the metadata for tools to a special CMDI module. Therefore we transferred the metadata from our own modules to a new module, proposed by the University of Utrecht.

The work was finished on time in 2016 (not, as planned originally, in 2015).

5.2.3.1.3 T31 Gretel upload

This small task (0.5 person months) involves work to start hosting the GrETEL application, once the new functionality (upload, analysis) has been added by UU, and can be combined with the task INT has anyway in their role as Dutch-Flemish HLT Agency / CLARIN B Centre, i.e. to host the GrETEL application. It can start only when a stable new version of GrETEL is available. A natural point to start some experiments with this is June 1, 2017, because it is targeted to release a first version of the GrETEL extensions integrated in GrETEL 3.0 by then.

5.2.3.1.4 AutoSearch

This task is the combination of the following tasks:

- T33 Autosearch upload
- T49 Autosearch metadata search

AutoSearch allows users to upload data for corpora, after which the corpora are made automatically searchable in a private workspace. The search application is powered by the INL BlackLab corpus search engine. The search interface is the same as the one used in for example the Corpus of Contemporary Dutch / Corpus Hedendaags Nederlands.

Status The working plan was written for 90% in 2016. Work will start in May 2017, with October 1 2017 as planned end date.

5.2.3.1.5 T41 OpenSoNaR update search

OpenSoNaR is an online system that allows for analysing and searching the over 500 million word Dutch reference corpus SoNaR. For OpenSoNaR two versions were created in CLARIN-NL: the original version called OpenSoNaR and an upgraded version called OpenSoNaR+, which contains a new frontend, adds support for spoken corpora and includes the Spoken Dutch Corpus (CGN). Through this task many improvements have been and will be implemented: the speeding up of problematic queries, more options for searching, grouping and sorting.

The work on OpenSoNaR concerns work on the backend (BlackLab). The work on the backend is carried out by INT. There is parallel work on the frontend (WhiteLab), carried out by Taalmonsters (Matje van de Camp). There is close cooperation between INT and Taalmonsters.

Status About 66% of the work was finished on time in 2016. Work is under way for the remaining 33%, with May 15 2017 as planned end date.

5.2.3.1.6 T43 WebCelex

WebCelex is a web based interface to the CELEX lexical databases of English, Dutch and German. The application was hosted at the Max Planck Institute for Psycholinguistics in Nijmegen. In this project the INT took over the hosting. The browser compatibility and the graphical user interface were checked, and some basic checks took place of the software and data. The application was placed behind the CLARIN login. WebCELEX is used by Taalportaal to provide the reader with examples of linguistic phenomena described there.

Status The work was finished as scheduled and on time in June 2015.

5.2.3.2 Meertens Institute

5.2.3.2.1 T02 Local version of VLO

Status: Changed plan

One of the main motivations for this task was the perceived poor state of the CLARIN Virtual Language Observatory. In the startup phase of this task specific task requirements were discussed with WP3 coordinators/stakeholders. In the meantime new versions of CLARIN's Virtual Language Observatory reflecting some of the desired changes were created. The course of further action is to be discussed further with WP3 coordinators; whether to pursue this task or reallocate the funds towards other activities. Discussions on this are ongoing.

5.2.3.2.2 T03 Improved metadata

Status: Ongoing

For Meertens Institute's resources a critical review of all metadata is planned as part of the transition towards the new FLAT repository. This includes feedback received from the CLARIN EU metadata quality coordinators. Planning of this project is aligned with delivery of new versions of the FLAT repository system, developed in collaboration with TLA Nijmegen.

5.2.3.2.3 T04 Tools to check metadata requirements

Status: Changed plan

Execution of this task has been postponed until the deliverables from the CLARIN Plus project have become available. In joint collaboration with CLARIN EU this task will further be taken up. The deliverable of this task will result in a software package or service that can be reused by others, e.g. by integrating it into the ingest workflow of a repository.

5.2.3.2.4 T79 Deployment framework

Status: ongoing

This task focuses on ease of deployment of delivered services. It is integrated into as many of the other tasks as possible to facilitate ease of use during the deployment of processes either as demonstrators or as production services. The main focus of this task is deployment via Docker containers. Deployment via Docker containers has been implemented for the Mtas task where a demonstrator setup is being distributed via GitHub. Also, OpenSKOS currently runs in a Docker configuration for the CCR and CLAVAS. For TTNWW, some services have already been dockerized, e.g. via LaMachine (https://github.com/proycon/LaMachine), which have been delivered by the technology providers of individual services. Dockerization of the Taverna service is to be investigated further.

For Meertens Institute's activities dockerization activities are synchronized with internal projects where, in collaboration with I&A (KNAW internal information services) and HuC partners, a dockerized production environment is prepared. This has currently entered a testing phase.

5.2.3.2.5 T08 Facilities for Creating CMDI metadata for often occurring metadata types.

Status: Halted

Pending further discussions with WP3 coordinators this task is halted until sufficient clarification is obtained on how to approach this task, more specifically, which CMDI profiles will need to be supported and how the results of this task are to be embedded in other ongoing activities. The goal of this task is to produce an easy to use form-based environment allowing end users create various CMDI metadata records without being confronted with the complexity of the CMDI framework.

5.2.3.2.6 T19 Metadata for tools.

Status: Halted

Awaiting input from UU.

5.2.3.2.7 T51 Metadata SPARQL endpoint

Status: work finished on time

A metadata SPARQL endpoint is operational for CMDI records. This allows for querying and retrieving CMDI records in an LOD representation. Further development and additional requirements are handled through the WP2 CMDI2RDF task.

5.2.3.2.8 T01/T48 Mtas/Nederlab integration

Status: work finished on time

URL: http://www.nederlab.nl/onderzoeksportaal/
Multi-Tier Annotation Search (Mtas) delivers a scalable solution for search and analysis of large annotated corpora. In recent years, multiple solutions have become available providing search on huge amounts of plain text and metadata. Scalable searchability on annotated text however still appears to be problematic. With Mtas, we add annotation layers and structure to the existing Lucene approach of creating and searching indexes, and furthermore present an implementation as Solr plugin providing both searchability and scalability. With Mtas we present a configurable indexation process, supporting multiple document formats, and providing extended search options on both metadata and annotated text, such as advanced statistics, faceting, grouping and keyword-in-context. Mtas is currently used in production environments, with up to 15 million documents and 9.5 billion words.

Mtas offers functional support for: CQL, KWIC, statistics, frequency lists, grouping, combined metadata/content search

Suitable for: Different annotation formats using a configurable indexing process (a.o., FoLiA, ISO 24624:2016), large data volumes

Integrated into Nederlab framework
- Nederlab board has chosen Mtas as its framework of choice (... 2016)
- Total number of documents 14.5 million
  8.7 billion word tokens (words and punctuation symbols)
  70 annotation tiers

Availability:
- Source code: Github https://github.com/meertensinstituut/mtas
- Documentation: https://meertensinstituut.github.io/mtas/
- Automated build and integration environment: http://www12.meertens.knaw.nl/jenkins/

5.2.3.2.9 T12/T13/T14 OpenSKOS

Status: work finished on time

URLs: http://145.100.58.150/ccr/public/ and **http://145.100.58.150/OpenSKOS-browser/)**

The OpenSKOS platform, originally developed in the CatchPLUS project, has since February 2015 been deployed as part of the CLARIN infrastructure, both as a Concept Registry and CLAVAS. In a coordinated project approach where NDE, CLARIN EU and CLARIAH were involved the back end architecture was migrated towards a RDF triple store and, not yet represented, SKOS elements were added. CLARIAH contributions were made to extend support for SKOS relations support, including visualizations as part of the SKOS browser. Currently, the new OpenSKOS version has entered the operational phase on behalf of CLARIN.

As part of the coordinated approach the source code branches that have developed during earlier development stages are to be merged into a single main branch. This merge is planned for OpenSKOS 2.2 which will be realized as part of NDE's OpenSKOS roadmap.

OpenSKOS offers functional support for: RDF triple store, SKOS Conceptschemes, SKOSCollections, SKOSConcepts, Relations, SPARQL and REST access points

Suitable for: SKOS-based thesauri

Collaborations:
- Development co-financed by CLARIN EU
- Development aligned with Netwerk Digitaal Erfgoed (Expert Group OpenSKOS)
- CLARIN EU (CCR en CLAVAS)
- DARIAH (Backbone thesaurus)

Availability:
- Source code: https://github.com/OpenSKOS/OpenSKOS/tree/master

5.2.3.2.10 T28 Upgrade TTNWW

Status: ongoing

For TTNWW new versions of available web services and tools are expected to be deployed. Synergy is expected with task T79 Deployment Framework through the use of Docker containers.

To further raise the level of user experience the environment is further integrated with Owncloud, allowing end users to upload, annotate and share their own data files. Furthermore, the annotation results have been integrated with Mtas, providing a first step towards integration with Virtual research environments such as Nederlab. For further integration, authentication and authorization procedures will need to be further synchronized to ensure end users have only access to their own or shared data.

Currently, possibilities are being discussed with SurfSara to provide this environment as part of their ongoing B2SHARE activities.

5.2.3.2.11 T39 MIMORE upgrade

Status: ongoing

As part of the MIMORE upgrade the user interface look & feel has been aligned with Nederlab and the underlying search engine has been replaced with Mtas. All linguistic information, originally represented in MYSQL databases, has been transformed to FoLiA representations. As for the tag sets, multiple tag sets are represented (MAND, SAND, Edisyn and CGN). Only for the MAND tag set a manual mapping process is still needed to facilitate the mapping onto CGN. While awaiting this, a test setup of the environment has been made available.

5.2.3.2.12 T58 Archiving ingest functionality/T39 Crowdsourcing software

Status: ongoing

As part of this task a connection between the Meertens Institute's Limesurvey questionnaire environment and the FLAT archive is being prepared allowing for deposition of finished survey data into the repository. This task is partly filled in in conjunction with the ongoing collaboration project with TLA Nijmegen focusing on the joint development of the FLAT archiving software. CLARIN-compliant CMDI packages are prepared and deposited via a SWORD interface (hence the relation with T58).[21] As part of this process an easy to use metadata editor is produced allowing capturing of the most essential metadata information deemed necessary for deposition as part of the survey request process.

5.2.3.3 Radboud University Nijmegen

Radboud has been involved in the following WP3 tasks:

T10 (Data Curation Service), T21 (CLAM), T22 (Frog), T23 (Frog Generator), T24 (FLAT), T26 (TiCCL), T55 (Ucto), T63 (Radboud lead), T70 (RDM Guidelines), and T71 (FoLiA). We describe these tasks in more detail below, grouping some tasks.

5.2.3.3.1 T10 (Data Curation Service), T70 (RDM Guidelines)

For Data Curation RU was appointed an additional task, being the definition and implementation of a profile for collection records.

RU has built a Django user interface[22] in which metadata for a collection can be inserted in a user-friendly way. This interface has been extensively used by a student assistant who completed the records

[21] https://en.wikipedia.org/wiki/SWORD_(protocol)
[22] https://www.djangoproject.com/

for 45 collections in this way. The interface was developed such that at its backend the metadata categories are directly connected to the CLARIN concept registry and a corresponding (validated) profile in the CMDI Component registry. In this way the corresponding CMDI files can directly be generated via a menu option in the interface.

Work is in process to define relations between collections and resources in this framework and integrating the resulting collection records in the search engines that CLARIN provides. Due to this work, T70 has been delayed. We expect that the delay is also profitable since relevant RDM[23] guidelines are currently established at University Libraries (e.g. in Utrecht and Nijmegen). We are closely reading these guidelines and foresee that they can be easily adopted for CLARIAH purposes in the very near future.

5.2.3.3.2 T21 (CLAM)

Github : https://proycon.github.io/clam/

CLAM allows developers to quickly and transparently transform a command line application into a RESTful web service and web interface, with which both human end-users as well as automated clients can interact. CLAM is set up in a universal fashion, requiring minimal effort on the part of the service developer. The actual application is treated as a black box, of which only the parameters, input formats and output formats need to be described. The application itself needs not be network aware in any way, nor aware of CLAM, and the handling and validation of input can be taken care of by CLAM.

CLAM is written in Python and runs on UNIX-derived systems. A Python API is provided, but knowledge of Python is not necessary to use CLAM. CLAM communicates using a transparent XML format, and uses XSL transformation offers a full web 2.0 web-interface for human end users.

In CLARIAH WP3, work on CLAM commenced in April 2016, ahead of planning. In October 2016, January 2017, and March 2017 new versions of CLAM were released (current version 2.1.9; see https://github.com/proycon/clam/releases for more details).

5.2.3.3.3 T22 (Frog) and T23 (Frog Generator)

Github: https://languagemachines.github.io/frog/

Frog is an integration of memory-based natural language processing (NLP) modules developed for Dutch. All NLP modules are based on Timbl, the Tilburg memory-based learning software package. Most modules were created in the 1990s at the ILK Research Group (Tilburg University, the Netherlands) and the CLiPS Research Centre (University of Antwerp, Belgium). Over the years they have been integrated into a single text processing tool, which is currently maintained and developed by the Centre for Language and Speech Technology at Radboud University.

[23] Research data Management

Frog performs tokenization, lemmatization, morphological analysis, part-of-speech tagging, syntactic chunking, dependency parsing, and named-entity recognition.

In CLARIAH WP3, Frog has been updated (by KvdS with input from AvdB) in two major aspects:

- Froggen (for Frog Generator, a trainable version of Frog) has been created to allow for the generation of a functional Frog instantiation based on a lemmatized and part-of-speech tagged corpus in any language (these two levels are the current scope of Froggen; more levels will be added). A successful test has been carried out with Old Greek; more tests with historical and dialectical variants of Dutch are planned.
- MBMA and MBLEM, the memory-based components responsible for morphological analysis and lemmatization, have been qualitatively improved. These tools, based on the CELEX and eLex lexical databases, generated errors due to inconsistencies in the original data. Morphological analysis is now also 'complete' in the sense that it produces fully nested morphological analyses.

Components of Frog have been refactored or reprogrammed (e.g. the dependency parser was rewritten in C++ from Python, speeding up parsing) and a long list of wishes and bugs were implemented and fixed. Work on Frog is continuous.

A manual for Frog has been created, with external writing support from dr. Iris Hendrickx (Radboud University).[24]

5.2.3.3.4 T24 (FLAT)

Github: https://github.com/proycon/flat

FLAT is a web-based linguistic annotation environment based around the FoLiA format, also further developed in CLARIAH WP3 (see below) by Maarten van Gompel. FLAT allows users to view annotated FoLiA documents and enrich these documents with new annotations. A wide variety of linguistic annotation types is supported through the FoLiA paradigm. It is a document-centric tool that fully preserves and visualises document structure.

FLAT is written in Python using the Django framework. The user interface is written using javascript with jquery. The FoLiA Document Server (https://github.com/proycon/foliadocserve), the back-end of the system, is written in Python with CherryPy and is used as a RESTful web service.

In CLARIAH WP3, work on FLAT was started ahead of schedule in April 2016. Throughout the past year, most additions to FLAT were due to user requests. Notable additions were the following. In May 2016, support for right-to-left scripts was added. In June, FLAT was selected as the annotation tool of choice by the PARSEME COST Action (http://typo.uni-konstanz.de/parseme/). In FLAT 0.5, released in September 2016, input options were added or improved: a better color scheme, a slider for setting confidence values manually. In October 2016, FLAT 0.6 was released with improved documentation, a new permission model with support for groups and group namespaces, and more user-friendly ways to edit metadata. In FLAT 0.7 (January 2017), the forms in which spans can be annotated have been extended to

[24] https://github.com/LanguageMachines/frog/raw/master/docs/frogmanual.pdf

cover dependencies and co-reference links. Phonetic content is now supported. Various visualisations are added (e.g. nested morphology and syntax trees).

For more information on releases, see https://github.com/proycon/flat/releases .

For a video demonstrating the current functionality of FLAT, see
https://www.youtube.com/watch?v=tYF6grtldVQ

5.2.3.3.5 T26 (TiCCL)

Github : https://github.com/martinreynaert/TICCL

TICCL, for Text-induced Corpus Clean-up, is the stand-alone command line version of the spelling correction and OCR post-correction system we have been developing since about 2008. TICCLops is TICCL as an 'online processing system'. It is a fully-fledged web application and web service based on CLAM (see above). Work on TICCL is performed by MR and KvdS, with contributions from MvG.

TICCL was published on GitHub in June 2016. Over time, KvdS and MR have worked together on profiling and improving parts of TICCL scripts by rewriting them from Perl scripts into C++ programs. In the fall of 2016 TICCL was extended from unigram-based OCR error correction to n-gram-based correction. Production-level versions of TICCL remain functioning at the unigram level until the n-gram-based fork is evaluated; this is ongoing work.

5.2.3.3.6 T55 (Ucto)

Github: https://languagemachines.github.io/ucto/

Ucto tokenizes text files: it separates words from punctuation, and splits sentences. It offers several other basic pre-processing steps such as changing case that you can all use to make your text suited for further processing such as indexing, part-of-speech tagging, or machine translation.

Ucto comes with tokenisation rules for several languages and can be easily extended to suit other languages. It has been incorporated for tokenizing Dutch text in Frog, our Dutch morpho-syntactic processor.

Development of Ucto is mostly based on bug reports. Tokenization is a task that has to operate on extremely large documents, and has to be robust towards deficiencies of many types (e.g. lacking newlines, deviating character sets, etc.). In October 2016, Ucto was made ‘language aware’ optionally: using language identification, Ucto can guess the language of a text, or even a sentence, and can pass this information onto Frog which can e.g. be told to ignore sentences that are not in the language of choice.

For more information on releases, see https://github.com/LanguageMachines/ucto/releases

5.2.3.3.7 T63 (Radboud lead)

AvdB continues to be the Radboud lead for CLARIAH WP3 after the start of his directorship at the Meertens Institute on January 1, 2017; AvdB stays on as professor at Radboud for two days per week until mid-2018 (after that date, one day per week). AvdB is also CLARIAH board member.

5.2.3.3.8 T71 (FoLiA)

Github: http://proycon.github.io/folia/

FoLiA, developed by Maarten van Gompel, is an XML-based annotation format, suitable for the representation of linguistically annotated language resources. FoLiA's intended use is as a format for storing and/or exchanging language resources, including corpora. Our aim is to introduce a single rich format that can accommodate a wide variety of linguistic annotation types through a single generalised paradigm. We do not commit to any label set, language or linguistic theory. This is always left to the developer of the language resource, and provides maximum flexibility.

XML is an inherently hierarchic format. FoLiA does justice to this by maximally utilising a hierarchic, inline, setup. Our aim with FoLiA is not to introduce yet another format, but to build a rich and practical infrastructure around this format. This includes tools, programming libraries, converters, visualisations and annotation environments.

With the VU group a collaborative effort was started in WP3 to develop conversion tools between FoLiA and NAF. A conversion tool for FoLiA to RDF was also developed.

FoLiA has seen many improvements. Just mentioning a small number of additions: FoLiA was extended to represent comments, various observations on texts, sentiment analysis, groups of semantic roles, and co-reference chains.

For a full overview of all releases, see https://github.com/proycon/folia/releases

5.2.3.4 RU Groningen (PaQu)

PaQu ('Parse and Query') is a web-based search tool for syntactically annotated corpora. It supports two search interfaces. The first one is a simple interface to query relations between words (this includes lemma and part-of speech). This interface allows for instance to list the frequency of all direct object nouns of a particular verb. The second one is an interface with a very powerful query language, which fully supports XPath2 and in addition has support for piped queries and macros.

PaQu includes support for the standard CGN and Lassy corpora. In addition, users can upload their own corpora. Those corpora will be parsed off-line, and then will also be available in both search interfaces.

In CLARIAH, two extensions to PaQu were foreseen. The first extension consists of support for more input formats, in particular support for FoLiA and TEI. The second extension is support for meta-data.

Both extensions of PaQu have been fully achieved.

PaQu now supports FoLiA and TEI input. In addition, PaQu also has support for compressed input formats (both zip and gzipped tar).

PaQu supports metadata. The standard corpora CGN and Lassy have been extended with meta data. Statistical information about the meta data of a corpus can be obtained, and, moreover, the results of queries can be broken down with respect to metadata attributes. The Dutch part of the CHILDES corpus has been added to PaQu as well, including various metadata attributes. Similarly, the VU DNC corpus including metadata has been added to PaQu.

Meta data in the input is supported as well. This includes support for meta data for corpora that use the FoLiA format, and corpora for which meta data is available in CMDI format.

Changes In the course of the project, a number of further extensions and improvements to PaQu have been implemented. These are listed here as follows:

- PaQu is available as a Docker file, for easier installation off-site, including installation at CLARIN centers.
- Improved support for federated login
- Code is now thread-safe
- Check labels of sentences for duplicates
- Real time syntax checker for XPath2 queries
- Improved visualization of dependency structures
- Support for alternative interpretation of co-indexed nodes

Status The work on PaQu has been completed

5.2.3.5 Taalmonsters

Taalmonsters has worked on WhiteLab, i.e. the frontend for the OpenSoNaR corpus search application.[25] It closely cooperated with INT, which worked on BlackLab, i.e. the backend of this search engine.

WhiteLab versions 2.1.0 and 2.1.1 have both been released late January 2017. Version 2.1.0 contains fixes for existing bugs and major updates to the interface as were requested for CLARIAH-CORE WP3. These include an interface for random sampling, detailed definition of PoS features in Extended and Advanced Search, import and export of queries in a straightforward XML format, and support for queries with gaps in Expert Search and N-grams Explore. Furthermore, sorting has been reimplemented on all result views, relative hit counts are reported wherever possible, case sensitive grouping has been added for grouping on token level features, and an option has been added to more precisely define the grouping context. Version 2.1.1 contains minor bug fixes on the described functionality.

A test version is currently online at http://opensonar.taalmonsters.nl[26]. Some of the new functionality, such as the gap queries, depends on functionality that has yet to be built into BlackLab Server and may therefore require additional adjustments after BlackLab Server development has been completed. However, developers on both teams were in constant contact during the WhiteLab development phase,

[25] https://github.com/Taalmonsters/WhiteLab2.0

[26] The test server runs on limited resources, which may affect the user experience..

so we expect these adjustments to be minimal, if they are needed at all. WhiteLab currently displays a message to the user when accessing features that are not supported by the backend.

The site tour and manuals[27] have been updated to include all new and existing functionality. The manuals also describe the complete installation procedure, which should be sufficient for anyone with basic Linux Server experience to set up the software.

Status The work on the OpenSoNaR+ frontend has been completed.

5.2.3.6 Utrecht University

Utrecht University is involved in the following WP3 tasks: Metadata Curation (T83), Metadata for Tools (T19), GrETEL upload (T31) and metadata search and analysis (T40) extensions, Upgrading the Typological Database System (TDS), and Management (T82),

We will briefly discuss each of them in a separate subsection. We also added a section on the cooperation with Leuven on GrETEL.

5.2.3.6.1 Metadata Curation

For metadata curation (T83) the aim is to make concrete proposals for

- A list of essential facets to be used in the VLO (Virtual Language Observatory) or other metadata search and browse application
- curation of values for facets that are currently used in the VLO
- extension of values for new and existing facets on the basis of implicitly available information (e.g. derivable from the metadata profile or collection).

A document describing the strategy proposed for data curation was written (Odijk 2015c). There is close cooperation with CLARIAH Austria on these matters. The Austria team made facilities to apply the proposals we (and others) make to the actual metadata used in the VLO, in part based on specifications derived for earlier work by Utrecht (Odijk 2014, Odijk 2015c). One of these is that it must be possible to map an existing facet value to a combination of values for multiple facets. We use local snapshots of the VLO metadata to experiment locally with different choices of facet value conversion.

An initial concrete proposal for a minimal set of facets has been proposed (contained in Odijk 2015d). Alternative proposals have been formulated as well (in particular dynamic facets that adapt to the current selection made), but it looks as if these are computationally intensive and might lead to confusing user interfaces. It is therefore unlikely that these will be implemented. It has also been investigated whether the selection of facets should be made dependent on the value of the resource type facet (Odijk 2015d). The conclusion was negative except for the distinction between software and data.

[27] http://whitelab.taalmonsters.nl/doc/manual/index

Concrete proposals for the curation of two facets (resource type and modality) as they occur in the VLO have been formulated (Odijk 2015a, Odijk2015b) and discussed with members of the CLARIN Metadata Curation Task Force. A start has been made for the curation of other facets, but this work has been delayed by the work for GrETEL updates (T31) and by work for CLARIAH WP1 by JO. This work will be picked up again in the second half of 2017 and is targeted to finish by Mid 2018.

Status Delayed and therefore re-planned. Some work by Meertens is partially dependent on this.

5.2.3.6.2 Metadata for Tools

Concerning Metadata for Tools (T19) the goals were as follows:

1. adapt the existing CMDI metadata profile for tools[28] to accommodate the information contained on the Services page of the CLARIN-NL portal (http://portal.clarin.nl/clarin-resource-list-fs)
2. transfer the information on the CLARIN-NL portal services page to CMDI metadata descriptions in accordance with the metadata profile mentioned in the first bullet
3. Gather additional information on the tools, especially technical information and locations of the source code from the original developers
4. Integrate the information obtained from the developers in the metadata descriptions
5. Publish the metadata in the VLO
6. Experiment with deriving the CLARIN-NL portal pages from the CMDI metadata records.

The bullets (1) through (3) have been carried out. The bullets (4) through (6) have been delayed because of personnel problems. These tasks will be picked up again in the second half of 2017, with new employees (still to be hired). Target finish date: Mid 2018.

Status: Delayed and therefore re-planned. No other tasks are dependent on this, so the impact is minimal

5.2.3.6.3 GrETEL upload facility

GrETEL is an application for searching in syntactically annotated corpora (treebanks). It was originally developed by KU Leuven in the context of the cooperation between the Netherlands and Flanders on CLARIN. Its major feature is example-based search, i.e. the user does not have to write a query but can have the desired query generated by the system on the basis of an example of the construction that the user is interested in.

GrETEL originally offered search facilities for LASSY-Small and the Spoken Dutch Corpus, it currently supports SoNaR-500 as well.

In task T31, the goal was to make it possible for a user to upload the user's own text corpus. This corpus can consist of text only, in which case the sentences of the text are automatically parsed by Alpino, or of text with syntactic structures for each sentence in the text. This upload functionality has been created. After such a corpus is uploaded, the resulting treebank is indexed by the system (using BaseX) and becomes searchable for the user. This facility requires login, which was also created (for the time being

[28] clarin.eu:cr1:p_1342181139640

Utrecht University login only; CLARIN-compatible login will be created when GrETEL is hosted by a CLARIN centre). This upload functionality has been demonstrated at the CLARIAH Toogdag 2016 (van der Klis & Odijk 2016). The uploaded corpus must be in zip format and consists of plain text files (extension .txt) or files parsed by Alpino (extension .xml).

The upload facility also enables one to upload metadata together with the data, in the PaQu Metadata format.[29] We made a cleaner and converter for corpora in the CHILDES CHAT format[30]: it removes annotations in the text and converts the data in CHAT format into the format that enables upload of the data and metadata. This converter and cleaner is available and has been tested but must still be integrated into GrETEL.

Support for uploading data and metadata in other formats (in particular, TEI, FoLiA) still has to be worked on and is planned for the second half of 2017.

Status Mostly finished, some small extensions for TEI and FoLiA to be added.

5.2.3.6.4 GrETEL Metadata Search and Analysis

In task T40 we have extended GrETEL so that one can use the metadata in analyzing the search results. First, the person who uploads the data can define which of the metadata are made visible in a faceted search interface, and how (by which GUI element) they are made visible.[31] Second, the user can filter search results by means of faceted search based on metadata elements and combinations of such elements. This functionality has been demonstrated at various occasions, inter alia the Utrecht Language Science Day, CLIN 27 in Leuven, and the CLARIAH Toogdag 2017 in Amsterdam (Van der Klis & Odijk 2017a,b,c). A full analysis component enabling full analysis of the data in combination with the metadata is targeted for June 1st, 2017.

Status On schedule

5.2.3.6.5 Cooperation with Leuven

The development of GrETEL is also on-going in Leuven, and we have closely collaborated with the Leuven researchers to ensure compatibility of the Utrecht and Leuven developments. Leuven has released Version 3 of GrETEL and we are currently in the process of integrating the Utrecht version with GrETEL Version 3. This has partially been finished but is targeted to be finalized by June 1, 2017.

Status: on schedule

The Utrecht GrETEL version is available via the URLs http://gretel.hum.uu.nl/ and http://gretel.hum.uu.nl/gretel3/ (for the integration with GrETEL3), but these are temporary

[29] See http://zardoz.service.rug.nl:8067/info.html#cormeta for a description of this format
[30] http://talkbank.org/manuals/CHAT.pdf
[31] Some metadata, e.g. transcriber, character encoding are essential as metadata but not useful for analyzing the search results.

development URLs. The fully integrated version will be available via the Leuven URL[32] and in a later phase by the CLARIN Centre INT (see Task 31). The source code is available at GitHub.[33]

The main GrETEL-developer at Utrecht will leave us as of June 1st. We are in the process of finding a replacement. Any further extensions of GrETEL (revised analysis functionality, support for more input formats) are targeted for Mid 2018.

5.2.3.6.6 Typological Database System

The Typological Database System (TDS) is a web-based service that provides integrated access to a collection of independently developed typological databases. In task T18 the goal is to add functionality to the TDS, in particular to make it possible to add new information to this database in an easy manner. Meetings have been held to clarify the requirements with users (Bylinina 2016), to specify the approach to be taken (Windhouwer 2016) and an initial work plan for the desired functionality has been created (Dimitriadis 2016). There is close collaboration with Menzo Windhouwer, one of the main original TDS developers, who will give (limited) assistance in the development. The work is to start around March 2017 and is expected to take several months in duration. An updated version may be expected by Q3 2017.

Status: On schedule

5.2.3.6.7 Management (T82)

This has been described in section 1.1.2.1.

[32] http://nederbooms.ccl.kuleuven.be/eng/gretel
[33] https://github.com/UiL-OTS-labs/GrETEL-upload

5.2.3.7 VU

We adapted the task specifications to the activities within the institutional settings. Within WP3, interoperability is thus defined along the lines of linguistic data categories in lexical resources, corpora and annotation systems. We carried out mappings and standardisation for NLP representations and lexical data. Interoperability was further defined at a WP exceeding level targeting semantic annotations and representations of textual interpretations. Semantic interoperability especially connects WP3 with WP2, WP4, and WP5 in which data in other modalities than text is modeled. We therefore decided to broaden the definition of the task and specify more specific subtasks within the originally planned tasks. Originally only task T11 "Shared vocabularies" was specified. We chose to subdivide this task into four tasks. Tasks T11.1 and T11.2 have been differentiated to logically divide the work within the VU and across other participants of CLARIAH. Tasks T11.3 and T11.4 are added for the semantic annotations tasks that are not covered by the Radboud University within the annotation task but are relevant for establishing semantic interoperability and annotation for use cases in the other Work Packages.

5.2.3.7.1 T06 CLARIAH internal formats and conversion scripts

NAF and FoLiA are two formats that are used in encoding linguistic annotations in texts that were previously not compatible. A first version of the NAFFoLiA specification and conversion scripts was developed and made available that contains the most important elements of both formats. This task started later than originally planned (30 September 2016), and will thus also finish later (30 September 2017 instead of 30 September 2016). This task is carried out in close collaboration with Radboud University Nijmegen.

Status This task is still ongoing. Currently, we are working on completing the conversion scripts and adding NIF conversions (NIF is an RDF format for linguistic information). We collaborated with Peter Bourgonje from DFKI, Saarbrücken, who developed a first conversion script from NAF to NIF. Finally, we participated in the ISO-Working group ISO/TC 37/SC 4/WG 5 to investigate the implications for NLP standards to be useful in practical applications.

5.2.3.7.2 T11.1 Definition of shared vocabularies

We considered various formats for modeling lexical data: contemporary, diachronic, and regional vocabularies. We adopted the LEMON standard with extensions for diachronic and regional usage as well as a SKOS representation for semantic linking and publishing vocabularies as linked open data. We tested the model for a range of different lexicons and developed conversion scripts.

In June 2016, a first version of the diachronic lexicons (DICOLOD) was presented at DHBenelux. This consisted of a conversion of 4 different lexicons, covering the Dutch language from 1600 onwards, to a shared format and conceptual model. For the conceptual model we used Lemon,[34] a W3C vocabulary for lexicons. The task started a bit later and will therefore finish a bit later too (30 September 2017 instead of 30 July 2017).

Changes

[34] http://lemon-model.net/

- The plan did not include the LEMON model, but as we came into contact with some of its creators and it is compatible with the formats used in WP4 and WP5 we decided to include this. Our extension of LEMON is now considered by research groups in Austria and Poland for representing their lexical data.
- Two more datasets are converted in accordance with the requirements of the model and added to the DICOLOD repository.
- New target: We planned to generate a diachronic emotion lexicon by linking the DICOLOD resources and enriching them with an existing emotion classification (i.e. WnAffect). This subproject has been started in Jan 2017 and will also be finished by September 2017.

Status The task is still ongoing. First versions of data, scripts and documentation can be found here.[35]

5.2.3.7.3 T11.2 Shared vocabularies entities: entity linking

We have created a fine-grained entity typing for Dutch module which can distinguish between 59 or 269 types (depending on the setting chosen). We are working on an entity linking module that utilises the entire linked open data cloud instead of only selected resources as most linkers do.

Changes While the task is described as "Specification, conversion and linking of entity repositories & publication as LOD", we found that there are many entities mentioned in humanities texts that are not present in any entity repository or the entity repository does not contain the desired information. As this is a pressing problem, we decided to start working on discovering information about these 'dark entities' such that factoids about these can still be linked to the text.

Status Slight delay due to a later start and the inclusion of dark entities but with some cross-project collaboration (Spinoza) we should be able to finish on time (30 March 2017).

5.2.3.7.4 T11.2 Shared vocabularies entities: occupations tagger

For social historians, the occupations of entities mentioned in text are an important piece of information. For most males, these can be found in structured data such as marriage certificates, but for women this information is often lacking. VU developed an occupations tagger[36] that can detect an occupation linked to an entity mention in text and link it to a resource such as HISCO.[37] This work is done in collaboration with Work Package 4 and will finish on 30 March (on time).

Changes While the task is described as "Specification, conversion and linking of entity repositories & publication as LOD", we found that there are many entities mentioned in humanities texts that are not present in any entity repository or the entity repository does not contain the desired

[35] https://github.com/cltl/clariah-vocab-conversion
[36] https://github.com/cltl/SimpleTagger
[37] http://historyofwork.iisg.nl/

information. As this is a pressing problem, we decided to start working on discovering information about these 'dark entities' such that factoids about these can still be linked to the text.

Status Improvements on the tagger will be carried out as part of the Pilot project HHuCap: The History of Human Capital.

5.2.3.7.5 T11.3.1 Event Detection

This task is scheduled for Q2-3 of 2017. We have code available from another use case that will be adapted for CLARIAH. The planned end-date is 30 September 2017.

5.2.3.7.6 T11.3.2 Emotion Detection

This task is scheduled for Q2-3 of 2017. We have lexicons and code available from other projects that will be put together for CLARIAH. . The planned end-date is 30 September 2017.

5.2.3.7.7 OTHER (dissemination)

- VU was co-organizer of the Linked Data Workshop (12 September, Amsterdam) http://www.clariah.nl/en/new/blogs/549-clariah-linked-data-workshop (collaboration with WP4 & 5)
- VU organized the WP3 Interoperability Workshop (16 September, Amsterdam) (related to task T0.6)
- VU participated in the Royal Library Hackalod and won the audience prize http://www.clariah.nl/en/new/blogs/575-team-clariah-wins-audience-award-at-hackalod-2016 (collaboration with WP4)

5.3 Detailed Overview WP4 Social Economic History

5.3.1 Overall Assessment

WP4 is progressing as planned. In the first year we delivered an experimental HUB, that provided tooling to Quiz (query Linked Data), Link (create Linked Data from csv or excel files) and Take (download Linked Data). In the second (past) year, we have built a more stable version of the HUB within the International Institute of Social History (IISG), that will host the HUB in the second half of CLARIAH and at least 5 years beyond.

We have also transposed and finalized a number of datasets into Linked (Open) Data as well as created various tools for transposing data and querying data. Finally, we are engaged in various cross-WP initiatives.

We are collaborating with WP2 on integration with Anansi as well as front-end design of tools. With WP3 we're engaged in two initiatives. One on extracting information on human interaction with buildings in order to show 'life courses' of buildings. In another project, we're augmenting occupations derived by WP3 from biographies with information on stature of occupations. By doing so we're able to augment the biographies with career trajectories. Finally, with WP3 and WP5 we're looking into how Linked Data helps to reconstruct.

In the remaining years we will connect the tooling to the new HUB, focus on usability with WP2, and complete transposing the remainder of the datasets to Linked Data.

Obviously, there have been some hurdles along the way. One of those was and is to find qualified personnel, especially for the knowledge representation part. In the first year we 'lost' a postdoc already, who was a perfect intermediary between the engineers and historians working on the project. The basic problem here is that you need an expert for the design of the infra-structure and tooling, but you don't want to bother this expert with the basics of knowledge representation.

Another issue with the project is the learning curve and coordination needed to transpose datasets into Linked Data in a *meaningful* way. The hard part is not so much to learn SKOS or specific vocabularies, as it is to learn how to apply it to actual data. Here the tension lies between the practical and proper representation of a dataset and the intended use of vocabularies.

The way we have dealt with this issue is to hire extra hands specifically taking care of the infra-structure part of the project. This person is responsible for the creation of the HUB and communication with tools and data. That relieves our engineers so they are able to focus on tool development and able to provide some support applying ontologies and vocabularies to historical datasets. By doing so, we have bent any negative influences on the timetable.

5.3.2 Assessment per task/sub goal

1. Deliver a 'structured data hub' that forms a single point of entry to a live repository of interconnected and (partially) harmonized datasets pertaining to the field of socio economic history (SEH).

   In year 1 we built a prototype HUB and evaluated it. In year 2 we are building an improved version of that HUB at the IISH, enhancing stability and scalability. This is on schedule as it will be in June 2016. Like the prototype version, the new HUB also provides various socio economic datasets.

2. By providing interlinked, integrated datasets, the hub should facilitate formulating and answering **cross-dataset research questions**.

   We have transposed various datasets into Linked Data. On the CLARIAH Toogdag 2017 we have demonstrated how we were able pull data from 4 different international datasets to answer a research question on the influence of religion on someone's social-economic position, controlled for a country's economic development. With the increasing number of datasets we'll transpose opportunities for cross-querying datasets will be increasing over the next two years.

3. The hub should grow to become a research infrastructure that is a **major resource** in the field of SEH.

   Having a more stable version of the HUB available we are now advocating the HUB in nationally and internationally. Nationally, since the start of the project, we are organizing sounding 'board' meetings whereby the 'board' consists of anybody in a historical discipline (e.g. economic history, historical demography). In these meetings we show are latest developments and ask for feedback on them. We are also organizing workshops, training master students, PhD students and Post-docs with –their data- and 'our' tools. We have also been very active in supporting historians with their CLARIAH call proposals, in order to retrieve maximum feedback on our tools, resulting in four financed projects with which we will work and receive feedback from.

   Internationally we're now focusing on advocating the HUB on social and economic conferences. For example, we in April we have a special slot in expert-workshop on linking Swedish Historical Person data and in November we'll be present at multiple session during the Social Science History Association conference, while for next year we're trying to obtain a session on the World Economic History Conference.

4. The hub should demonstrate its **value** (and the value of its data) to non-experts in the domain.

This is an important aim, but something we are planned to work on in year 3. The reason is that first we need to secure a stable product and then ask non-experts to work with it. We are already preparing for this item by closely collaborating with WP2 who have lots of expertise in creating front-ends and know about usability. We have dedicated an engineer 1 day a week to WP2 for exchange related issues and these front-end issues.

5. The project will deliver a **critical mass** of data made accessible through the hub. The datasets in the hub have to meet a number of criteria pertaining to relevance, importance and quality. This is to maximize the impact of the data hub on current research questions, and attract individual researchers that want to **contribute** and/or **study** the data it hosts.

   We are well underway with transposing datasets into Linked Open Data. For various datasets we had some minor issues, such as retrieving the actual owner of the datasets and ask for rights. For a number of datasets that are restricted we now also have restricted Linked Data and are able to demonstrate this to the respective parties and see whether they would be interested to share their data this way.

6. The hub should be **integrated** with the central CLARIAH infrastructure delivered within work package 2 of the project. This is safeguarded by membership of key WP4 personnel in WP2 fora, and physical proximity at regular intervals.

   To ascertain we would achieve this goal we have dedicated an engineer for one day a weak to WP2. This has resulted in the fact that at the moment ANANSI is able to extract all Linked Open Data from our HUB. The communication is however still experimental and will be fully fledged in the coming year.

7. The hub should be **hosted** at and **integrated** with the infrastructure of the IISG.

   In the initial plan we aimed to first build a fully functional prototype HUB at the VU and then rebuild it in the final year at the IISH. However, when inventorying the IISH environment, it seemed more appropriate to develop the HUB from within the IISH environment. Therefore, we already built the HUB from inside the IISH environment already allowing for perfect integration. At the moment the HUB allows for most data-related services, but tooling is still working from the VU-environment. In the coming year, we will focus on moving the tooling to the IISH as well.

8. The hub should integrate and **adopt lessons learned** from its predecessors. Most notably CEDAR, HSN and ClioInfra. This involves migration of CEDAR, HSN and ClioInfra data and tools to the CLARIAH infrastructure, to the extent that this contributes to CLARIAH goals.

The hub should be up to date and reflect the latest version of datasets as much as possible. It should therefore be **resilient to changes** to any datasets that it depends on.

- CEDAR: CEDAR is available from our HUB. In addition to migrating the data and endpoint we had to align domain names.
- HSN: a sample of the HSN has been transposed to Linked Data. Because of privacy, the data are non-open and require registration with HSN.
- ClioInfra: the data in ClioInfra are transposed to Linked Open Data.

We are still working on a workflow to properly adopt changes in files. We are able to pull data from Dataverse and then transpose it to Linked Data. This could even be automated with event hooks. (Thus recreate the Linked Data once a new version of the dataset appears in Dataverse). The question however is: what to do with the Linked Data of the old version? Obviously you want to preserve it, but you also don't want multiple versions of the same data in the graph, as it reduces speed.

5.4 Detailed Overview WP5 Media Studies

5.4.1 Overall Assessment

Aims: Work Package 5 (WP5) is engaged in setting up an infrastructure for access to audiovisual sources and related contextual material (program guides, RTV ratings, photographs, etc.). It aims to consolidate five existing tools for exploratory and targeted, contextual media research (CoMeRDa, AVResearcher, Trove, Dive and Oral History Today, developed as prototypes in the context of NWO CATCH, CLARIN-NL and CLARIAH-Seed), to continue their development and to train scholars in using them.

Approach: Most audiovisual sources are inaccessible due to copyright restrictions (broadcasting materials, films) or privacy issues (interview collections). Therefore, we have had to build a closed, authenticated environment in which the data can be made accessible to scholars and students. It also means that, since most of the audiovisual data cannot be exported, we have to bring the tools to the data, rather than the other way around. For this purpose, we have designed the Media Suite: a virtual research environment in which scholars find the data and tools for their research and which offers them the opportunity to build collections, bookmark and enrich them.

Making the original tools available in the Media Suite in practice meant we have had to rebuild each of their components and then reassemble these in the configuration of the original tool interface. In 2016 we have decided to exchange the original, tool-oriented approach for a modular one, whereby each component of the original tools is rebuilt and offered as a stand-alone mini-tool. In addition, we provide the original tools in the form of 'recipes' that combine the functionalities of the mini-tools into one more comprehensive interface, that allows for more complex research (for example, combining faceted search over multiple collections with visualizations of search results in the form of snippets, a word cloud and a timeline, as in AVResearcherXL). We also provide software libraries for the tools which allow developers to build their own, customized 'recipe' and embed them in a new interface. This modular approach facilitates maintenance (and thus is more sustainable) and provides researchers with a better insight into the workings of the tools (exposing rather than black-boxing them, so allowing researchers to take a tool-critical perspective). Finally, the Media Suite provides APIs for more advanced users to directly query the raw data.

The data collections that are made accessible within the Media Suite are registered in the CLARIAH WP5 Data Registry (a CKAN instance).

Organization: The efforts have been concentrated on providing an infrastructure at the Netherlands Institute for Sound and Vision (NISV), the CLARIAH data center for audiovisual data, as well as sustaining and further developing the five existing tools for audiovisual data retrieval and analysis. Initially the work package was structured in projects around 1. Infrastructure; 2. Tools; and 3. User studies/dissemination, whereby NISV was responsible for the infrastructure and the original development teams (humanities scholars, computer scientists and programmers) were responsible for the tools. In 2016 we have reorganized the work package in five new task forces, that better reflect the new, modular approach and that mix NISV staff and the researchers and developers involved in the original tools:

- TF1: Collections (=data)

- TF2: Tools
- TF3: MediaSuite
- TF4: User Studies
- TF5: Outreach & Dissemination

Since the building of the tools is closely related to the MediaSuite, in practice task forces 2 and 3 operate as one. We also implemented a new planning, indicating which components and collections will be made accessible in each version of the Media Suite.

Evaluation: The start of the project was marked by a long stage of inventory and design in 2015 and early 2016. In monthly meetings and various additional design sessions and user studies (aimed at defining and refining requirements for the Media Suite) we have defined our commonly shared goal. In 2016 we started building the foundations and components of the first version of the Media Suite. The initial plan was to make the tools and collections listed for version 1 available to researchers from Q1 2017. This has been postponed with three months; version 1 will be launched on 4 April 2017 so is available from Q2 2017 – still in time for usage in the Research Pilots.

The delay was mainly caused by the departure of one of the programmers at NISV in September 2016 and the difficulty to find a suitable replacement. The task of the original programmer has now been divided in two separate tasks: one related to translating user-needs into requirements and communicating between programmers and researchers (taken up by postdoctoral researcher Liliana Melgar for 0,2 fte at NISV, in addition to her work at UvA (0,6 fte) on coordinating the user studies and outreach and dissemination activities) and one focused on the actual programming tasks (taken up by a new programmer to be appointed shortly at NISV). In order to further extend the programming capacity required for the next versions of the Media Suite we have identified separate tasks that can be outsourced to external developers (such as interface design and customized tool-building, for which we collaborate with Frontwise, Dispectu and Spinque).

5.4.2 Assessment per task / sub goal

The following table describes for different phases in a research project what functionality will be offered (column *description)* and in which version of the Media Suite this functionality will be available (column *available in version*). The versions to be released and their release dates are listed below.

Description	Available in version*	
Phase in which you select which sources are relevant for your research, and determine how the quality of the source(s) influences your results		
Overview of available collections including collection descriptions	1	
Collection and metadata quality analysis, visualizing missing data	1	
Select documents from different collections to build a personal corpus	2	

Phase in which you query the selected collection(s) (e.g., via simple keyword search or more advanced queries) in order to find relevant results		
Allow users to select which collections to use in search/analysis functionalities and recipes	1	
For metadata records and transcripts. Includes full-text, field-restricted, faceted, date-restricted search, similar document search, cross-collection search	1	
Allow users to search with multiple queries in a comparative search results view	1	
Phase in which you analyse the corpus you selected in the search phase		
Playout of audio (e.g. radio) with segment selection and annotation (comments, tags and/or links)	1	
Playout of video with segment selection and annotation (comments, tags and/or links)	1	
Semantic browsing and exploration through AV collections, with AV playout	1	
Construct and annotate narrative paths	2	
For metadata, annotations and transcripts. E.g. named entity recognition and extraction, topic modelling, sentiment analysis	3	
Speaker identification, face recognition, shot detection, emotion recognition, key frame extraction, colour analysis	3	
Phase in which you present the findings obtained in the analysis phase in visual form. Alternatively, these functionalities may be useful to support the search and/or analysis phase, i.e., heuristically, rather than for presentation purposes.		
Shows search results or collection metadata on a timeline	1	
Word/tag/entity cloud (based on entire collection or user query)	1	
Shows search results (user query) in thumbnail format	1	
Phase in which you organize, manage and enrich your corpus; supports all of the above phases in the research process		
work space for organizing and analyzing search interactions, annotations, bookmarks	1	
platform for crowd/niche-sourcing annotations	1	
Automatic Speech Recognition service (for AV documents in core collections and own AV documents)	2	
Planned Releases		
2017-04-04	Version 1	
2017-07-01	Version 2	

2018-01-01	Version 3	
2018-07-01	Version 4	

6 Appendix: Acronyms

Acronym / Term (URL)	Type	Expansion / Explanation	Dutch Expansion
ADAH Call	Project Call	Accelerating Scientific Discovery in the Arts and Humanities	
AISB	conference	Society for the Study of Artificial Intelligence and the Simulation of Behaviour	
Alpino	Software	Parser for Dutch	
ALTO	format	Analyzed Layout and Text Object	
ANANSI	CLARIAH sub project	project for construction of the overarching components of the CLARIAH infrastructure	
API	term	Application Programmer Interface	
ATHENA	CLARIAH sub project	project to develop a data portal that will hold information on historical context of human – nature relationships	
AutoSearch	software	web application for token-based searching in one's own corpus	
AV	term	Audio-Visual	
AVResearcherXL	software	tool for exploring radio and television programme descriptions, television subtitles and general newspaper articles	
B2SHARE	software	software to store and publish data	
BaseX	software	XML database system	
BlackLab	software	Backend search engine	
CATCH	project	NWO Programme Continuous Access To Cultural Heritage	
CATCHPLUS	project		
CC	project	CLARIAH-CORE	

CDI	term	Collaborative Data Infrastructure	
CEDAR	project	Census Data Open Linked	
CELEX	data	Dutch Centre for Lexical Information	
CELEXweb	Software	web application around CELEX	
CEO	term	chief executive officer	
CGN	data	Spoken Dutch Corpus	Corpus Gesproken Nederlands
CHAT	organisation	Centre for Humanities and Technology	
CherryPy	software	Python Web Framework	
CHILDES	infrastructure	Child Language Data Exchange System	
CKAN	Software	data management system	
CLAM	software	Computational Linguistics Application Mediator	
CLARIAH	infrastructure	Common Lab Infrastructure for the Humanities	
CLARIAH-CORE	Infrastructure project	Dutch main CLARIAH project	
CLARIAH-PLUS	project proposal		
CLARIAH-SEED	infrastructure project	CLARIAH Seed Capital project	
CLARIN	infrastructure	The CLARIN infrastructure	
CLARIN ERIC	organisation	CLARIN ERIC	
CLARIN-NL	infrastructure project	The Netherlands national project for CLARIN	
CLARIN-PLUS	project		
CLAVAS	software	CLARIN Vocabulary Service	
CLEVER	CLARIAH sub project	project for evaluating the current situation with regard to availability and performance, and to make an estimate of future human and financial resources, and functional requirements	
CLIN	conference	Computational Linguistics in the Netherlands (local conference)	

Clio Infra	infrastructure	infrastructure for research into long-term development of worldwide economic growth and inequality.	
CLiPS	organisation	Computational Linguistics & Psycholinguistics Research Center	
CMD2RDF	project	CMD to RDF	
CMDI	software	Component MetaData Infrastructure	
CMDI2RDF	CLARIN-NL subproject	Project to convert CMDI to RDF linked data	
COMERDA	software	Contextualizing Media Research Data	
COST	funding programme	EU funding programme for pan-European networking of nationally funded research activities	
CQL	query language	Contextual Query Language	
DANS	organisation	Data Archiving and Networking Services	
DARIAH	infrastructure	Digital Research Infrastructure for the Arts and Humanities	
DFKI	organisation	Deutsche Forschungszentrum für Künstliche Intelligenz	
DH	term	Digital Humanities	
DH-Lab	organisation	Utrecht University Digital Humanities Lab	
DicoLOD	lexicon collection	DIacronous COnceptual lexicons through Linked Open Data	
DiDDD	data	Diversity in Dutch DP Design	
DIVE	software		
DODH	repository	Dutch Overview Digital Humanities	
Edisyn	wiki	Wiki on dialect syntax	

eHumanities.NL	platform	national platform that brings together expertise and research in the development and use of digital technologies in the humanities and the social sciences	
eLex	lexicon		
ERIC	organisation	European Research Infrastructure Consortium	
EU	organisation	European Union	
EUDAT	project	European Data Infrastructure	
EUR	organisation	Erasmus University Rotterdam	
FLAT	software	FoLiA Linguistic Annotation Tool	
FLAT	software	Fedora Language Archiving Technology	
FOLiA	Data format	Format for Linguistic Annotation	
FROG	software	software to assigne linguistc properties to text such as part of speech tagger, lemmatizer, and more	
FROGgen	software	software to generate a FROG given a set of training data	
GPL	license	GNU General Public License	
GrETEL	software	**Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics	
GUI	term	Graphical User Interface	
Hack-a-LOD 2016	event	Hackathon with LOD	
HERA	organisation	Humanities in the European Research Area	
HI	organisation	Huygens ING Institute	
HISCO	standard	Historical International Standard Classification of Occupations	
HLT Agency	organisation	Organisation to maintain and distribute Dutch language language resources	

H-PEP	CLARIAH sub project	project for creating a data model for for persons based on the *Biografisch Portaal* collection and transform this dataset into RDF.	
HSN	infrastructure	Historical Sample of the Netherlands	
HSS	term	Humanities and Social Sciences	
HUB	Software	a 'structured data hub' that forms a single point of entry to a live repository of interconnected and (partially) harmonized datasets pertaining to the field of socio-economic history	
Huygens ING	organisation	Huygens ING Institute	
IAP	term	International Advisory Panel	
IBM	company	International Business Machines	
ICT	term	Information and Communication Technology	
ID	term	Identifier	
IISG	organisation	International Institute for Social History	Internationaal Instituut voor Sociale Geschiedenis
IISH	organisation	International Institute for Social History	Internationaal Instituut voor Sociale Geschiedenis
ILK	organisation	Induction of Linguistic Knowledge Research Group (Tilburg University)	
INL	organisation	Institute for Dutch Lexicology	Instituut voor Nederlandse Lexicologie
INT	institute	Institute for the Dutch Language	Instituut voor de Nederlandse Taal
IPR	term	Intellectual Property Rights	
ISO	organisation	International Standards Organisation	

ISO TC37/SC4	standard	ISO Subcommittee focusing on Language Resource Management	
IT	term	Information Technology	
IVDNT	organisation	Institute for the Dutch Language	
KB	organisation	National Library of the Netherlands	
KNAW	organisation	Royal Netherlands Academy of Arts and Sciences	
KNHG	organisation	Royal Dutch Historical Society	Koninklijk Nederlands Historisch Genootschap
KU Leuven	university	Catholic University Leuven (Belgium)	Katholieke Universiteit Leuven (België)
KWIC	term	KeyWord In Context	
LaMachine	software	to warp softweare in a virtual machine	
LASSY	project	Large Scale Syntactic Annotation of written Dutch	
LimeSurvey	software	Open Source Survey software	
LOD	term	Linked Open Data	
LREC	conference	Language Resources and Evaluation Conference	
MAND	data	Morphological Atlas of the Dutch Dialects	
MBLEM	software	Memory-based lemmatization	
MBMA	software	Memory-based morphological analysis	
MI	organisation	Meertens Institute	
MIMORE	CLARIN-NL subproject	MIcrocomparative MOrphosyntax REsearch Tool	
MPI	organisation	Max Planck Institute for Psycholinguistics	
Mtas	software	Multi-tiered Aggregated Search	
MySQL	software	database system	
NAF	format	NLP Annotation Format	

NDE	network	Network Digital Herittage	Netwerk Digitaal Erfgoed
Nederlab	project		
NIF	format	NLP Interchange Format	
NIOD	organisation	Netherlands Institute for War Documentation	Nederlands Instituut voor Oorlogsdocumentatie
NISV	organisation	Netherlands Institute for Sound and Vision	Instituut voor Beeld en Geluid
NL	country	the Netherlands	
NL eScience Centre	organisation	Netherlands eScience Centre	
NLP	term	Natural Language Processing	
NWO	organisation	Netherlands Organisation for Scientific Research	
OAI	organisation	Open Archives Initiative	
OAI-RS	software framework	OAI Resource Sync	
OCR	term	Optical Character Recognition	
OPENSKOS	CLARIN-NL subproject	To create the CCR	
OpenSONAR	CLARIN-NL subproject	Online Personal Exploration and Navigation of SoNaR	
Oral History Today	software		
PaQu	CLARIN-NL subproject	Parse and Query	
PARSEME	project	PARSing and Multi-word Expressions	
PI	term	Principal Investigator	
PICCL	CLARIAH sub project	Philosophical Integrator of Computational and Corpus Libraries	
Qber	software	Crowd Based Coding and Harmonization using Linked Data	
QSODA	CLARIAH sub project	project on Documentation, Data & Software Sustainability	
RCE	organisation	Cultural Heritage Agency	
RDF	standard	Resource Description Framework	
RDM	term	Research Data Management	
REST	standard	Representational State Transfer	

RKD	organisation	Netherlands Institute for Art History	
RTV	term	Radio and TV	
RU	organisation	Radboud University Nijmegen	
RUG	organisation	Groningen University	Rijksuniversiteit Groningen
RUN	organisation	Radboud University Nijmegen	
SALAD	workshop	Services and Applications Workshop	
SAND	data	Syntactic Atlas of the Dutch Dialects	Syntactische Atlas van de Nederlandse Dialecten
SEH	term	Social Economic History	
SKOS	standard	Simple Knowledge Organization System	
SOLR	software	Lucene based search platform	
STEVIN	project	Dutch-Flemish programme for realizing the BLARK and HLT research for Dutch	
SURFSara	organisation	Collaborative organisation for ICT in Dutch education and research	
SWORD	protocol	Simple Web-service Offering Repository Deposit	
SWOT	term	Strength Weakness Opportunity Threat Analysis	
TDS	software	Typological Database System	
TEI	Standard text format	Text Encoding Initiative	
THATCamp	event	The Humanities and Technology Camp	
TICCL	software	Text-Induced Corpus Clean-up	
Timbl	software	software for memory-based learning algorithms	
TLA	organisation	The Language Archive	
Trove	software	Transmedia Observatory	

TTNWW	CLARIN-NL subproject	LRT Tools for Dutch as web services in Work flows	TST Tools voor het Nederlands als Web services in Work flows
Ucto	software	Unicode aware Tokenizer	
UCU	organisation	University College Utrecht	
UL	term	User Interface	
UM	organisation	University of Maastricht	
UNIX	operating system		
URL	term	Universal Resource Locator	
UU	organisation	Utrecht University	
VCC	organisation	Virtual Competency Centre	
VLO	software	Virtual Language Observatory	
VU-DNC	CLARIN-NL subproject	VU Diachronic Newspaper Corpus	
WHISE	workshop	Workshop on Humanities in the Semantic Web	
WhiteLab	software	Search Engine front end	
WP	term	Work Package	
XML	Standard data format	eXtensible Mark-up Language	
XSL	language	XML Style Sheet Language	
XSLT	programming language	XSL Transformations	