CLARIN-*NL* Final Report

Editor: Jan Odijk (Utrecht University)
Authors:
Hans Bennis (Meertens Institute)
Daan Broeder (MPI)
Arjan van Hessen (Utrecht University)
Marc Kemps-Snijders (Meertens Institute)
Steven Krauwer (Utrecht University)
Nelleke Oostdijk (Radboud University Nijmegen)
Ineke Schuurman (Utrecht University)

Date: 2016-06-08

FINAL

1 Introduction

This document is the Final Report of the CLARIN-NL project.

The CLARIN-NL project started on April 1, 2009, and finished on December 31st, 2015

The structure of this document is as follows. We first briefly highlight the major achievements of the project as well as the major points of attention for successor projects (section 2). After a summary (section 3), section 4 lists the success criteria for the various activities planned in the CLARIN-NL project as defined in the long term working plan. Sections 5 through 12 describe the results of each of these activities against the success criteria defined for this activity.

Section 13 describes some other developments, not originally covered in the long term working plan, and section 14 contains the major conclusion of this report.

Throughout the text, reference will be made to the CLARIN-NL Fact Book, which contains all factual information on the CLARIN-NL project. This CLARIN-NL Fact Book can be found on the CLARIN-NL website.

This report has been written by different authors, and though an attempt has been made to harmonize the texts, one will still encounter shifts in style. We have often used hyperlinks to refer to external resources. Expansions of the acronyms can be found in a separate document and here .

2 Highlights

We summarize the major achievements of the CLARIN-NL project:

- CLARIN-NL created the Netherlands part of the CLARIN infrastructure with 5 centres, 4 of which are certified CLARIN centres;
- CLARIN-NL has incorporated a wide range of data and dedicated software applications into the CLARIN infrastructure, enabling their use by a much larger community than before CLARIN-NL;
- CLARIN-NL has raised wide awareness of the existence and importance of the CLARIN infrastructure within the humanities researcher community in the Netherlands;
- The CLARIN infrastructure and the data and software applications contained in it are actually used in research, and its use is increasing;
- CLARIN-NL has a clear focus on language but covers a large spectrum within the Humanities;
- Big steps have been taken in improving interoperability, both on the syntactic and the semantic level;
- We managed to get significant amounts of funding for successor projects (CLARIAH-SEED, CLARIAH-CORE);
- Through CLARIN-NL, the Netherlands have played a leading role in CLARIN at the European level and promoted international cooperation.

Of course, there is still room for significant improvement. We list the major issues:

- There is as yet no business model that makes the CLARIN infrastructure sustainable, i.e. so that it can continue to exist also without occasional funding through the National ESFRI Roadmap funds;
- Interoperability of software and data still requires a lot of improvements, not only in the Netherlands but also in the whole CLARIN infrastructure;
- Visibility of the resources (e.g. via the CLARIN Virtual Language Observatory) must be significantly improved;
- The creation of common CMDI metadata must be made much simpler;
- More sophisticated options for searching through distributed content must be created.

Fortunately, all these topics (and others) are worked on in the CLARIAH-CORE successor project, taking into account the experience gained in CLARIN-NL.

3 Summary

The CLARIN-NL project ran from 2009 through 2015 and has set up the Netherlands part of the CLARIN infrastructure. It has formed a network of 5 candidate CLARIN centres that cooperated to create generic infrastructure services, centre-specific infrastructure services such as resource repositories, metadata harvesting, single sign on authentication and authorization, and resolution of persistent identifiers to enable stable access to resources. A metadata registry has been set up and filled with metadata profiles and components created on the basis of data and software from the Netherlands, editors to make new components and profiles, and editors to make or adapt metadata. The centres also have cooperated to create metadata search facilities and a first prototype of federated content search. All five centres have the Data Seal of Approval (one of the requirements to become a certified CLARIN centre) and 4 of them have become certified CLARIN centres. By their cooperation relevant knowledge and expertise has been built up in multiple centres so that a robustly supported infrastructure emerges. . Its robustness was severely but successfully put to the test when MPI decided to reduce its role in CLARIN. CLARIN-NL was also successful in connecting national data centres such as the National Library (including DBNL), NIBG and Utrecht University Library.

Some 51 data curation and demonstrator projects covering a broad spectrum of humanities disciplines were carried out on the basis of the needs of the targeted users of the CLARIN infrastructure, humanities researchers. In these projects existing data and software were curated to make them CLARIN-compatible, i.e. stored in a CLARIN-supported format, with CLARIN-compatible metadata (CMDI), stably referable by the use of persistent identifiers (PIDs), and with links of its data categories to CLARIN-supported data category registries to ensure semantic interoperability. These data and software, and their metadata, are hosted on CLARIN centres ensuring permanent visibility via the Virtual Language Observatory, the metadata search engine , accessibility via resource-specific and (limited) federated content search, and long term preservation. The curation of these data and software has provided tests of a whole range of infrastructural aspects, such as the proposed CLARIN standards, metadata creation, semantic interoperability via ISOCAT and later CCR, and hardware and software requirements for computationally and/or memory intensive services. A whole range of demonstrators has been created, which can be used not only for demonstration and educational

purposes but also for supporting actual humanities research. These include a workflow system that incorporates language and speech technology as web services, which was created in cooperation with CLARIN Flanders.

The user needs have been inventoried in a dedicated user survey and are otherwise ensured by the required active user involvement in data curation and demonstrator projects.

On the organisational level, CLARIN ERIC has been established in February 2012, hosted by Utrecht University. It ensures international coordination and cooperation among the CLARIN partners in the National Coordinators' Forum and in a range of committees on technical and legal aspects. On the national level an excellently functioning governance structure and National Coordination Point have been set up. The national CLARIN activities have been shaped up as a programme rather than as a project and this approach has been very successful in that it offered much opportunity to react to emerging problems, bring in more partners, and to react to ideas and proposals coming from our prospective users. The Netherlands plays a leading role in CLARIN because of the early start and the relatively high budget of its national project, it's hosting of CLARIN ERIC, and the excellent technical expertise it offers to the CLARIN community.

A wide range of knowledge sharing activities have been organized and supported, including national and international tutorials and workshops, seasonal school courses, information sessions, a Helpdesk including FAQ sections on the website, and support for travel of researchers to relevant events. Several educational packages were developed and are, together with other educational material, available via the CLARIN-NL Portal. Several courses in digital humanities in which the CLARIN infrastructure is prominently present have become part of the regular humanities curriculum of some universities.

Dissemination and PR has been implemented through the CLARIN-NL website, newsletters and newsflashes, active participation in the *eData & Research* magazine, messages via humanities mailing lists, presentations, publications, meetings with individual researchers or research groups, leaflets, and a range of movies and screen captures. Through all these activities, CLARIN-NL is well known among humanities researchers in the Netherlands. A book on CLARIN in the Low Countries with over 26 contributions is being prepared and is expected to appear by the end of 2016.

Two centres of expertise have been set up, one for data curation and one for data and tools for the longitudinal study of language and culture.

CLARIN-NL makes its data and services visible and available via a portal through which all services are visible and can be accessed.

It can be stated with confidence that the "general ambition of the CLARIN-NL project to be a major contributor both organisationally and technically to the design, specification, construction and exploitation of a European-wide CLARIN infrastructure" has been achieved. The Netherlands part of the infrastructure is "actually used by its intended users (humanities scholars)". Research that crucially used the CLARIN infrastructure is now appearing in the *Lingua* Special Issue on the CLARIN Infrastructure with Jan Odijk as guest editor. We are approached by many researchers when they are

preparing new project proposals (and some of them were awarded funding). CLARIN is also involved in a project proposal for the NWO *Gravitation* programme that is being prepared by a large consortium of linguists united in LOT. The first PhD in which crucial use is made of CLARIN has been successfully defended by Liesbeth Augustinus in Leuven. The CLARIN work for this was made possible in part through the close collaboration of Flanders with CLARIN-NL.

The major objectives of the project have been achieved, though a range of problems remain and will have to be addressed in successor projects. Fortunately, CLARIN-NL and DARIAH-NL have joined forces and managed to secure funding for the infrastructure projects CLARIAH-SEED (2012-2014) and CLARIAH-CORE (2015-2018). In addition, there has been close cooperation with NWO-Groot projects such as Nederlab and Taalportaal.

Making use of the CLARIN infrastructure by its intended users (humanities scholars) is becoming "a normal 'modus operandi' for them as reflected by the incorporation of training and education in the regular curricula of humanities studies".

4 Success Criteria

This section lists the success criteria for the CLARIN-NL project in general. They have been copied from the long term working plan. Here we assess to what extent the criteria have been met, and what the prospects are for meeting them by the end of the project.

The general ambition of the CLARIN-NL project is to be a major contributor both organisationally and technically to the design, specification, construction and exploitation of a European-wide CLARIN infrastructure that is actually used by its intended users (humanities scholars) and where its use has become a normal 'modus operandi' for them as reflected by the incorporation of training and education in the curricula of humanities studies.

This general ambition can be translated into more specific success criteria that are linked to the actions defined in the long term working plan (section 4), and that are subdivided into an EU-Line and an NL-Line (as was also done for the actions):

- EU-Line
    - Has the governance structure as recommended by the CLARIN preparatory project been implemented, and is it functioning successfully?
    - Has the main CLARIN office for administrative and logistic support for the governance bodies been set up in the Netherlands?
    - Has the main European CLARIN Technical Centre to build and maintain the technical infrastructure been set up?
    - Has the central CLARIN coordination point been set up for development and maintenance of standards, harmonization of IPR issues, and education, dissemination and promotion?
    - Has the Netherlands set up an international example infrastructure with one or two other leading partners?

        - Is this example infrastructure nationally and internationally recognized as exemplary?
        - Is this example infrastructure successfully used by the intended users (linguistic and humanities researchers)?
        - Is this example infrastructure sufficiently known among the intended users?
- NL-Line
    - Technical Infrastructure
        - Has the intended technical infrastructure actually been constructed?
        - Is the distributed nature of technical infrastructure indeed invisible to users?
        - Is the performance of the technical infrastructure sufficiently good?
    - Data Infrastructure
        - Are the data identified as “essential” included in the technical infrastructure?
        - Can they be easily found and accessed by users, and are they properly documented?
        - Have NL-specific IPR issues been adequately dealt with? Have procedures been defined to deal with IPR issues for new data?
        - Are there guidelines and procedures in place as well as supporting tools to easily incorporate new data into the infrastructure?
    - Language technology infrastructure
        - Have the tools and services identified as “essential” been included into the technical infrastructure?
        - Can they be easily found and accessed by users, and is there proper documentation to assess their appropriateness for a given task?
        - Has actual interoperability among tools/services and between tools and data been achieved in the technical infrastructure?
    - User Needs
        - Are the data and tools identified as “essential” indeed the data and tools the intended users need most?
        - Is the technical infrastructure indeed used by the intended users?
            - Are the intended users happy in using the technical infrastructure?
            - Do the interfaces promote and stimulate working with it or do they pose obstacles?
            - Are the interfaces user-friendly and self-explanatory where possible?
            - How many researchers actually use the infrastructure on a regular basis or occasionally? (Target: 40% of the intended users regularly use it; 70% use it at least occasionally)
    - Centres of Expertise
        - Have one or two centres of expertise been created and are they operating successfully?
        - Are they recognized as centres of expertise, both nationally and internationally?
    - Dissemination, education, awareness
        - Is the existence of the technical infrastructure known to the intended users (target: 80% know of its existence)

- Have all relevant players been sufficiently informed to be able to participate in designing and constructing the infrastructure?
- Has enough training and education on using the infrastructure been given? Are still ample opportunities offered to get such trainings?
- Has training in the use and actual use of the infrastructure been incorporated in the regular curricula of linguistics and humanities studies, or are there concrete plans to do so? (Target: incorporated in 60% of the curricula after 6 years)

o National Coordination Point
  - Has a national coordination point been set up and is it functioning successfully?
  - Has a business model been developed that guarantees the long term sustainability of the CLARIN infrastructure?

We will dedicate a separate section for a discussion of the evaluation of success criteria for each of the long term working plan actions. Each success criterion will be listed in the following font

Success Criterion

after which it is discussed.

5 EU-Line

Has the governance structure as recommended by the CLARIN preparatory project been implemented, and is it functioning successfully?

The CLARIN preparatory project recommended organizing the CLARIN governance structure as an ERIC. CLARIN ERIC was established in February 2012, and this joyful fact was celebrated on several occasions (at the Ministry of Education, Culture and Sciences; at the opening of the Academic Year 2012-2013 in Utrecht, and on a special occasion set-up by the European Commission).

CLARIN ERIC started with 8 countries and one intergovernmental organisation: Austria, Bulgaria, the Czech Republic, Denmark, the Dutch Language Union, Estonia, Germany, the Netherlands, and Poland. It now also counts Finland, Greece, Italy, Lithuania, Norway, Poland, Portugal, Slovenia, and Sweden among its members, with the UK as observer. Several other countries are working towards joining CLARIN ERIC.

Alice Dijkstra (NWO) is the representative for the Netherlands in the CLARIN ERIC General Assembly, with Jan Odijk as expert. Steven Krauwer was CLARIN ERIC executive director until September 2015, after which Franciska de Jong took over. The Netherlands is represented in the National Coordinator's Forum, attends the monthly (virtual) meetings, and is active in it. (Jan Odijk/Daan Broeder).

CLARIN-NL concluded an agreement with CLARIN ERIC on the contributions of the Netherlands to the CLARIN infrastructure in July 2012.

Has the main CLARIN office for administrative and logistic support for the governance bodies been set up in the Netherlands?

The main office for administrative and logistic support for CLARIN ERIC has indeed been set up in the Netherlands. CLARIN ERIC is hosted by the Netherlands, more specifically Utrecht University.

Has the main European CLARIN Technical Centre to build and maintain the technical infrastructure been set up?

According to the CLARIN Technical and Scientific Description[1] the CLARIN technical infrastructure is shaped as a federation of existing centres, providing access to data and tools as web services. Every CLARIN ERIC member has to provide at least one of those centres. No central hub is foreseen, but some centres will provide more services than others. Interoperability between centres is ensured by the Standing Committee for Technical CLARIN Centres, consisting of the heads of the main centres in each country. Quality is ensured by the Centre Assessment Committee that formulates quality criteria and checks whether centres and candidate centres satisfy these criteria.

Has the central CLARIN coordination point been set up for development and maintenance of standards, harmonization of IPR issues, and education, dissemination and promotion?

The CLARIN ERIC Board of Directors has been operational from the very start of CLARIN ERIC. It is responsible for the overall coordination, supported by the CLARIN ERIC Office in Utrecht and several committees, such as the Standards Committee, the Centre Assessment Committee, and the Legal Issues Committee, which have been set up in the course of 2012. Education, dissemination and promotion are the main activities in the so-called Knowledge Sharing Infrastructure, consisting of a network of Knowledge Centres (K-Centres). The Web Editorial Board is responsible for the creation and maintenance of the central CLARIN Website.

There is active participation of CLARIN-NL members in CLARIN ERIC committees, e.g.

- Marc Kemps-Snijders, Member Standing Committee for CLARIN Technical Centres[2]
- Daan Broeder member Centre Assessment Committee
- Daan Broeder, member CLARIN ERIC standards committee
- Menzo Windhouwer, member CLARIN ERIC standards committee
- Arjan van Hessen, Web Editorial Board
- Arjan van Hessen, Arwin van der Zwan, Hetty Winkel in the CLARIN ERIC Office

And in an earlier stage:

- Eelco Ferwerda, member CLARIN ERIC Legal Issues Committee
- Remco van Veenendaal, member CLARIN ERIC Legal Issues Committee

[1] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-76

[2] Dieter Van Uytvanck, is Chair Standing Committee for CLARIN Technical Centres (on behalf of Germany)

CLARIN-NL has been very active towards the European enterprise to create a CLARIN infrastructure under the guidance of CLARIN ERIC, and it has been very active in other CLARIN or infrastructure-related projects and initiatives. This can be illustrated with many examples:

- CLARIN-NL employees have organized several LREC tutorials and workshops on CLARIN-related topics (metadata, semantic interoperability)
- Other tutorials and workshops that attract international researchers have been organized
- Large delegations from CLARIN-NL attended the CLARIN Annual meetings since 2012 and many of them had contributions in the form of presentations and/or showing demonstrators.
- Many CLARIN-NL people and/or organizations are or were involved in important European projects and initiatives that are directly relevant to infrastructures in general and the CLARIN infrastructure in particular, e.g.
    - DASISH, which brings together **all 5 ESFRI research infrastructure initiatives** in the social sciences and humanities
    - EUDAT, which brings together data management service providers and research communities who are directly involved with the design of data services. The consortium includes key representatives from research community infrastructures from diverse disciplines.
    - RDA (Research Data Alliance), which is an organisation that aims to accelerate and facilitate standardizing and harmonizing research data sharing and exchange.
    - CLARA is training a new generation of researchers who will be able to cooperate across national boundaries on the establishment of a common language resources infrastructure and its exploitation for the construction of the next generation of language models with wide theoretical and applied significance
    - CLARIN PLUS: CLARIN-PLUS is dedicated to enhancing CLARIN. Following the recommendations of the 2013 ESFRI Assessment Expert Group, CLARIN-PLUS proposes to accelerate the implementation and to strengthen and consolidate CLARIN in the following areas: the central (technical) hub, the central office, partnerships with other infrastructures, outreach, and governance
    - PARTHENOS: The PARTHENOS project empowers digital research in the fields of History, Language Studies, Cultural Heritage, Archaeology, and related fields across the (Digital) Humanities, through a thematic cluster of European Research Infrastructures, integrating initiatives and world-class infrastructures, and building bridges between different interrelated fields. PARTHENOS will achieve this objective through the definition and support of common standards, the coordination of joint activities, the harmonization of policy definition and implementation, and the development of pooled services and of shared solutions to the same problems
    - EUDAT2020: EUDAT20202 is a Horizon 2020 project that brings together a unique consortium of e-infrastructure providers, research infrastructure operators, and researchers from a wide range of scientific disciplines under several of the ESFRI themes, working together to address the new data challenge: Data, and a fortiori Big Data, is a cross-cutting issue touching all research infrastructures. EUDAT2020's vision is to enable European researchers and practitioners from any research

discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres.

Has the Netherlands set up an international example infrastructure with one or two other leading partners?

With respect to connecting and building on the foundations laid in the CLARIN preparatory phase such as the Joint metadata domain using CMDI metadata, the AAI[3] solution using Federated Identity Management (FIM) and the use of Persistent Identifiers (PIDs) we can report the following.

CLARIN NL has continued building on and contributed considerably to extending the CMDI metadata infrastructure, both to the technology (IIP project) and the content side in the form of new profiles (metadata project and via resource curation) and made various contributions for the ISOcat content manager. CLARIN NL has discussed CMDI related issues with representatives from many other national CLARIN projects and is coordinating the CMDI ISO standardisation work.

All CLARIN centres are now members of the CLARIN Service Provider Federation (SPF). The Dutch situation does pose some special challenges to get our potential user group connected to the CLARIN SPF; this is referred to elsewhere in this report.

The use of PIDs is following the EU pattern: most centres are using PIDs based on the Handle System either by running their own Handle Server installation or making use of external parties. A small minority uses URNs for PIDs, which is the preference in the publication oriented Library domain, though all certified CLARIN centres in the Netherlands and one candidate CLARIN centre use PIDs based on the Handle System (which is now obligatory in CLARIN).

In general it can be said that CLARIN NL has made efficient use of the preparatory phase work and very much contributed to extending the laid foundations.

We furthermore realised international cooperation with Flanders: a project on language technology for the shared Dutch language. Its most important part for CLARIN-NL is the TTNWW project. It will be described in more detail below.

In addition, there has been close international cooperation on a variety of topics. For example, intense international cooperation on federated search, initiated in the CLARIN preparatory project, has been continued. The CLAVAS tool developed in CLARIN-NL is already in use outside of the Netherlands, inter alia in Austria, and will be deployed in other countries as well. Close cooperation with Austria on metadata curation and faceted search in the Virtual Language Observatory has been initiated in CLARIN-NL and is continued in CLARIAH-CORE.

[3] Authentication and Authorisation Infrastructure. See http://www.clarin.eu/glossary#AAI

At the level of end user oriented services for the humanities it turned out to be quite difficult to set up an international and fully integrated example infrastructure with one or two other leading partners. The reason is that in most countries the national projects are completely planned and there is little room in terms of human and financial resources for unplanned cooperation and integration activities with other countries. The Netherlands is thus far the only CLARIN ERIC country that has shaped its national CLARIN activities as a programme rather than as a project with fully fixed deadlines and deliverables throughout its duration of the project. This approach has been very successful in that it offered much opportunity to react to emerging problems, bring in more partners, and to react to ideas and proposals coming from our prospective users.

We have initiated such cooperation with Germany and the Czech Republic, but this has not yet led to cooperation at the operational level for the reasons indicated. With the CLARIN ERIC in place since 2012, coordination of international cooperation through independently financed projects (e.g. via Horizon2020) can be organized in a much more systematic manner, and this already has been successful (see the projects listed above) . International cooperation in the context of the CLARIN ERIC is already functioning well (e.g. in the committees referred to above), and this also offers opportunities for increased international cooperation on specific shared interests with CLARIN ERIC partners.

Cooperation between the Netherlands and Flanders CLARIN-NL cooperates with CLARIN Flanders in a cooperation project. The overall project consists of two parts:

- **Part I (1356k Euro, NL: 904k Euro, VL: 452k Euro)**
  - In this part a specific cooperation project has been formulated in which the Netherlands and Flanders aim to adapt various existing components developed in the CGN project and especially in the STEVIN programme and incorporate them into a work flow system for web services developed in CLARIN. The whole system will run on servers of recognized CLARIN-centres and aims to offer facilities to researchers from the humanities with little or no technical expertise. These facilities must 1) allow them to address their research questions in a better and easier manner, and 2) create opportunities for formulating completely new types of research questions
  - In the Netherlands this part consists on the one hand of the infrastructure implementation plan described in section 6, page 15, in which the basis for the technical infrastructure will be designed and implemented (with a budget of 504k euro). On the other hand 400kEuro will be invested for the further realization of this part (in the TTNWW project). Flanders contributes in this part 452k Euro in accordance with the usual NL: 2/3-VL: 1/3 proportion (all in the TTNWW project).
- **Part II (920k Euro; NL: 680k Euro; VL: 340k Euro)**
  - The contribution from the Netherlands consists of 680k Euro spent on the first open call for demonstrator and resource curation projects.
  - In Flanders the research community represented in CLIF has formulated project proposals for a total amount of 340k Euro.

The project started in March 2010 and ran until Dec 31, 2012. A keynote speech on TTNWW has been held at CLIN 2013 (Jan 18, 2013) in Enschede. A workshop on the results of this cooperation project was held in September 2013 in Leuven, Flanders. The TTNWW project and its results are discussed in more detail in section 8.

Is this example infrastructure nationally and internationally recognized as exemplary?

With respect to connecting and building on the foundations laid in the CLARIN preparatory phase as the Joint metadata domain using CMDI metadata, the AAI solution using Federated Identity Management (FIM) and the use of Persistent Identifiers (PIDs), it can be claimed with confidence that CLARIN is at the forefront internationally. This holds even though the technologies mentioned are not specific to humanities or linguistics but quite generic infrastructure functionality that is also useful in infrastructure for completely different domains.

For the results of the cooperation project between the Netherlands and Flanders, especially the TTNWW project, the system is available and has been used. The system and its design, as well as systems developed in Flanders in the context of the Netherlands-Flanders cooperation have been presented at various national and international conferences, and were well received there. External factors (e.g. hardware and software changes at SURFSara, where it runs) have made the system going down regularly. CLARIAH-CORE is working towards resolving these problems.

Is this example infrastructure successfully used by the intended users (linguistic and humanities researchers)?

The resources brought in by researchers and the participating CLARIN centres are used a lot by the researchers, but at this moment we have no means of quantifying this objectively. However, there are other indicators that clearly show that the CLARIN infrastructure is increasingly often used in actual research and in education. Research that crucially used the CLARIN infrastructure is now appearing in the *Lingua* Special Issue on the CLARIN Infrastructure with Jan Odijk as guest editor. We are approached by many researchers when they are preparing new project proposals. The research project proposal by Van Koppen and Dietz for the NWO Vrije Competitie programme, in which we contributed to the preparation and in which some of us participate actively was awarded funding. CLARIN (JO) is involved in a project proposal for the NWO Gravitation programme that is being prepared by a large consortium of linguists united in LOT, and, if it is awarded, the researchers will make heavy use of the CLARIN Infrastructure and also contribute to it. The first PhD in which crucial use is made of CLARIN has been successfully defended by Liesbeth Augustinus in Leuven (18 October 2015). Though this is a PhD in Flanders, the work on CLARIN in Flanders was made possible by the close collaboration of Flanders with CLARIN-NL.

Is this example infrastructure sufficiently known among the intended users?

For the results of the cooperation project between the Netherlands and Flanders, it can be said that several presentations on the system have been given both with the intended users as audience

([Kemps-Snijders 2009], Workshop in Leuven held on Sep 20, 2013) and with potential new contributors as audience ([Schuurman & Kemps-Snijders 2013], keynote speech at CLIN 2013). Two courses on CLARIN were given at LOT seasonal schools (July 2014 and January 2015) and both included the results of the cooperation project between the Netherlands and Flanders. Many other lectures and tutorials were given on specific subparts of this. The results are also clearly visible and accessible through the CLARIN-NL portal.

We conclude that its existence is well-known among language and speech technology researchers, and knowledge about its existence and opportunities is expected to continue to grow among other humanities researchers.

6 Technical Infrastructure

Has the intended technical infrastructure actually been constructed?

A lot of activities have been carried out to design and construct the technical infrastructure. We introduce them here and describe at the same time the structure of this section. Hyperlinks in this paragraph refer to the relevant subsection.

First of all, a number of organisations have expressed the ambition to become certified CLARIN Centres.

These candidate CLARIN centres form the backbone of the CLARIN infrastructure and worked together, in two projects, to construct the technical infrastructure (Infrastructure Implementation Project, IIP) and metadata and federated content search facilities (Search&Develop, S&D). We have approached libraries and data archives to provide their data in a CLARIN-compatible manner. Some of them worked on that and became CLARIN Data Providers.

In addition, a number of smaller projects have been carried out to construct specific parts of the technical infrastructure: the metadata project, CLAVAS, the metadata for tools project, and the Meertens Deposition Facility.

Finally, we have tested aspects of the technical infrastructure with a range of data curation and demonstrator projects.

CLARIN Centres The core of CLARIN infrastructure is formed by the CLARIN centres. In the Netherlands, there are 4 certified CLARIN Centres: Max Planck Institute for Psycholinguistics (MPI), Meertens Institute (Meertens), Institute for Dutch Lexicology (INL), and Huygens ING (HI), and one candidate centre: Data Archiving and Networked Services (DANS). All centres have the Data Seal of Approval, which is one of the conditions for becoming a certified CLARIN Centre. The certified centres were formally certified as CLARIN Centre by the CLARIN Centre Assessment Committee.

At the CLARIN Centre meetings not only the various technical issues of being a CLARIN centre but also the issues concerning the sustainability of running CLARIN services were discussed. This

discussion led in 2012 to a re-assessment by the candidate centres of their ambitions as a CLARIN centre.

The major issues are (1) how can the costs be covered after the CLARIN-NL-project ends; and (2) many centres are unsure whether they can host software services in a sustainable way[4], and are therefore reluctant to make strong commitments. (3) Some centres are research institutes that are used to deliver services to their own research groups and maybe a select external group. Delivering services to a potentially very large community (CLARIN EU) is considered risky. Clear agreements will have to be made here to determine the exact character and role of each centre, and to guarantee that the research data and software provided by the humanities researchers will find a centre as a host. Discussing these issues with the centres we have been able to convince several of them to keep or start offering services, provided they can do so on a best effort basis. Although we see that such guarantees could be better described, we feel that for the moment this is the maximum achievable. This issue has been taken up again in the CLARIAH-CORE project.

In order to address the problem of hosting software services, a new centre type (B+), was defined, In addition to the known types A and B. A B+-centre is a centre that explicitly aims to host software and data management services (in addition to hosting data) for the whole CLARIN community.

DANS has made clear that it aims to be a B-centre (and has made a plan towards it). INL, HI, and Meertens aim to be B+-centres. All centres are working towards certification as a CLARIN-centre. A dedicated workshop on this for whole CLARIN has been held in April 2013.[5] Though MPI / TLA reconfirmed its ambition to become an A/B+ centre in 2012, a change in policy in 2014 made them reduce their activities significantly. They reduced their role in the CLARIN infrastructure, basically only guaranteeing sustained preservation of data deposited with them but making no further commitments. This was a severe test for the stability of the CLARIN infrastructure, because many generic infrastructure services (so-called A-services) had to be transferred to other centres or to CLARIN ERIC. In addition, easy accessibility of data deposited at MPI/TLA became uncertain. A plan was set up to distribute the A-services over other CLARIN Centres throughout Europe and CLARIN ERIC. The transition from ISOCAT to the CLARIN Concept Registry (CCR) as the major registry for ensuring semantic interoperability, which was planned already, had to be accelerated, and this was taken up by the Netherlands (Meertens Institute). In addition, work was initiated on a new deposition system (FLAT), in part taken up by CLARIN-NL (MDF subproject), to ensure that copies of the data residing at MPI /TLA would be easily accessible for researchers in the long run and to accommodate new data that will be produced in the future.

In the course of 2012, the Dutch Language Union (DLU) decided to extract the HLT-Agency, until then running as a project within INL, from INL. The HLT Agency is hosted at the DLU as of January 2013. The DLU has brought in the data and tools it owns and manages via the HLT-Agency as its

[4] With many CLARIN services, research groups outside the centres, deliver the technology and it is not certain that the required expertise will remain available to them. Centers can only make commitments for availability of storage and processing resources.

[5] http://www.clarin.eu/node/3622

contribution to CLARIN. A complication in this situation is that the DLU is a separate member in the CLARIN ERIC and that thus the HLT agency formally falls not under the scrutiny of CLARIN-NL. We therefore consider the DLU responsible for these data and tools and their integration into the CLARIN infrastructure. We closely monitored the developments and closely collaborated with the HLT Agency because they manage a large set of crucial data and tools for the Dutch language (e.g. the Spoken Dutch Corpus, and data and tools created in the STEVIN programme). Fortunately, CMDI metadata were made available for these data and they are visible in the CLARIN VLO since 2014.

The separation of the HLT Agency and independent developments at INL caused a reassessment of the character and organisation of INL, leading to the decision to turn INL into an Instituut voor de Nederlandse Taal (INT, Institute for the Dutch Language)). Fortunately, this appears to incorporate an even stronger role for INT as a data and application provider, and a return of the HLT Agency data and applications to INT.

Infrastructure Implementation Project (IIP) The basic functionality of the CLARIN infrastructure was created in the IIP project, in which four CLARIN centres participated (MPI, Meertens, INL, DANS). In the course of 2012, the Huygens Institute for Dutch History (Huygens ING) also expressed the ambition to become a CLARIN Centre, submitted a plan for this, and joined the CLARIN centre meetings organised within the IIP project. The Huygens ING plan was accepted, started in 2013, and Huygens ING is now also a certified CLARIN Centre.

In the IIP project plan, two main goals were specified: (a) transferring infrastructure expertise to those candidate centres that need this, so they could set-up, connect and maintain the required CLARIN services such as publishing CLARIN metadata for their resources, (b) (further) develop and maintain the required central CLARIN services and tools: the CMDI, Schema, Relation, and ISOcat registries and the Virtual Language Observatory (VLO) and ARBIL tools. These services and tools are also contributions to the general CLARIN infrastructure.

With respect to the transfer of knowledge and the implementation of CLARIN technical requirements there was overall good progress thanks to the centres being able to hire competent staff and the CLARIN training efforts. Work on enabling Federated Identity Management (FIM) and connecting to the CLARIN Service Provider Federation (SPF) required expert knowledge that was originally not available in all centres. Work on this topic sometimes also required coordinating with different parties such as IT service providers and SURFnet. The fact that the Dutch Identity Federation (SURFfed) was transitioning to their new SURFConext platform has been a complication. Not all CLARIN-NL (funded & promoted) web-applications are accessible by CLARIN users by FIM. Priority was given to keep essential CLARIN infrastructure tools as ISOcat and CMDI registries FIM-accessible.

A general problem with enabling FIM access to CLARIN (web) tools is the complexity of connecting CLARIN users to SURFfed. An official request of the user's home organization IT department is required for that and many users are not capable to effectuate that. Therefore the CLARIN EB took action and approached the Universities and research institutes to allow their staff access to the CLARIN services & tools. This action was quite successful.

Although there is a need for refinement and additional functionality, all registries and tools with exception of the Relation Registry are operational.

The status of the centres with regard to the CLARIN requirements at the beginning of the CLARIN-NL project was characterized as follows in the IIP plan:[6]

Item	MPI	INL	Meertens	DANS
Proper repository system	x	X	planned	x
Explicit and harvestable metadata	x	Partly	in progress	x
Support of PIDs incl. citation options	x	Partly	planned	x
Long-term archiving solution in place	x	X	in progress	x
Resources in open resource formats	most	Most	relational DB	most
Membership in Dutch IdF	x	X	planned	planned
Membership of Dutch CLARIN SPF	easy	Easy	planned	planned
Quality self-assessment via DANS DSA[7]	ongoing	Planned	-	ongoing
Open for deposits from other researchers	x	X	-	x
Structured metadata and content search option	x	Planned	partly	x

Some of the items play a role in the certification of a centre as an A or B centre.

The status of the implementation of CLARIN requirements by the centres by the end of the CLARIN-NL project can be seen in the following table. We marked the cells for which progress has been made in green.

Item	**MPI**	**INL**	**Meertens**	**DANS**	**HI**
Proper repository system	Y	Y	Y.	Y	Y
Harvestable CMDI metadata	Y	Y	Y	Y	Y
Support of PIDs incl. citation options	Y	Y	Y	Y	Y
Long-term archiving solution in place	Y	Y	Y	Y	Y
Resources in open resource formats	Y	Y	Y	Y	Y
Membership in Dutch IdF	Y	Y	Y	Y	Y
Membership of CLARIN SPF	Y	Y	Y	In progress	Y
Quality self-assessment via DANS DSA	Y	Y	Y	Y	Y
Open for deposits from other	Y	Y	Y	Y	N

[6] The original table also contained two additional rows: one for 'accessibility of resources via web apps': its meaning is unclear and in its most obvious interpretation it holds for all centres (though not necessarily for all data in a centre); and one for 'open for digitization of audio/video formats'. This has never played an important role in CLARIN because it involves the creation of new digital data, and this was less in the focus of CLARIN. In addition, the item is of no relevance for certification as a CLARIN-centre.

[7] Data Seal of Approval: http://datasealofapproval.org/

researchers					
Structured metadata and content search	Y	partly	Y	partly	partly
Federated Search connection	Y	Y	Y	Y	N

Compared with the situation at the start of CLARIN, significant progress has been made. For example, all centres have adopted the use of PIDs. All use Handles, and DANS, which used to use only PIDs based on URNs now also uses DOIs. As stated, all centres except DANS are certified CLARIN-Centres. DANS not being a certified CLARIN centre is problematic since DANS is important as a centre of choice for many data-types. DANS only has to make the step to become a member of the CLARIN Service Provider Federation (SPF) to enter the certification procedure.

The various components of the CLARIN infrastructure that have been created separately in a variety of projects (infrastructure functionality, search functionality, curated data, curated tools, demonstrators, web applications, etc. etc.) are brought together and are made accessible via a single portal to the Netherlands part of the CLARIN infrastructure. This portal offers

- An inventory of CLARIN-NL tools and services that can be selected via facets that reflect important aspects of the tools and services
- An inventory of CLARIN-NL data that can be selected via facets that reflect important aspects of the data
- A summary of Dutch CLARIN centres
- A selection of tools to find relevant resources in CLARIN using metadata and content search
- A series of pages with recipes to do the most often used language data manipulation and processing actions, as well as educational packages to teach or learn working with the relevant services and data.
- A system to ask advice from colleagues about CLARIN tools and services issues.

The portal support CLARIN-compatible federated login.

Search & Develop (S&D)

General Overview The Search & Develop (S&D) project aimed to create a network of centres in the Netherlands, focusing around text content search services. The S&D projects aimed to create search facilities for search in metadata stored in a centralized database (itself filled and regularly updated via metadata harvesting), combined with federated search facilities in the actual data. A central metadata search engine capable of handling CLARIN metadata CMDI files has been delivered. A first prototype S&D federated content search demonstrator with limited functionality (term search only) has been created demonstrating the principles of federated content search based on the CLARIN SRU/CQL specification involving resources/content search engines at all participating CLARIN-NL centres. At the Sofia CLARIN conference (Oct 2012), it became clear that there are different views (and implementations) on federated search. After further discussions, it was decided that the approach proposed by German researchers would be adopted also in the Netherlands.

We now list the individual goals of the S&D project and discuss their status in more detail

S&D Goal 1: Generic search engine

Introduction At the start of the project there appeared to be a fragmentation of search interfaces. Lots of projects were in the process of building such interfaces for one or a few resources only; in fact the situation was that almost each corpus came with its own interface. This led to the unwelcome situation that the infrastructure to be realized had a large number of unrelated or partly overlapping search engines. Although the specifics of each research domain supports the development of tools that are specific to related research questions, many aspects of these search tools are of a more generic nature. In this project, we approached the problem of search engines from the general perspective of the infrastructure, rather than from the specific point of view of a research community. Aspects that were of primary concern here were metadata search and federated content search and the relation between the two. Although progress has been made in these areas it is fair to say that not all possible challenges have been dealt with. A metadata search engine has been realized that takes the ISOCAT concept references, as expressed in the metadata records, into account. Federated content search has been realized on selected resources of the project's participants but the level of integration remains relatively shallow with respect to the proposed use cases. This is related to the fact that there is currently still no sound technological basis for handling more complex annotation searches at all levels of granularity, let alone harmonization of these efforts into the SRU/CQL CLARIN specification. However, during the course of the project we have seen technological developments and adoption of new technologies that provide a perspective towards this in the future.

Original Goal A generic search engine. The S&D projects aims to create search facilities for search in metadata stored in a centralized database (itself filled and regularly updated via metadata harvesting), as well as federated search facilities in the actual data.

Adapted Goal The original project proposal specified three phases where each of the subsequent phases would tackle more complex metadata/content search functionality. The most complex cases, structured distributed content search in combination with metadata search could not be reached within this project. This is largely due to the lack of a widely accepted technological basis for addressing these complex annotation search issues, and the concomitant selection of resources where complex multitier annotations were only represented to a limited degree. While some progress has been made at the level of technological components, the S&D team considered it premature to try to harmonize and consolidate these efforts in the SRU/CQL specification.

Current status A metadata search engine was created at the Meertens Institute capable of ingesting diverse CMDI metadata records and harmonizing the metadata fields through the ISOcat Data Category Registry. All CLARIN-NL centres have furthermore provided a SRU/CQL implementation for selected resources that have been incorporated into this metadata search engine to provide the possibility of a federated metadata/ content search. The metadata search engine is also being used as part of the Nederlab project, but it has become clear that in particular the usability of the interface remains problematic due to the perceived complexity of the interface. As a result the

usability of the user interface has been addressed in the last stages of the project, in collaboration with the Nederlab project, to provide users with a more intuitive method of navigation. Also, the connection between the metadata and distributed content search was to be re-established here. However, the solution adopted in Nederlab focused completely on aggregated search, so no more work on federated search has been done. The issue will be taken up again in CLARIAH-CORE.

Towards achieving the (adapted) goal A metadata search engine has been created taking into account the diversity of the CMDI metadata descriptions and descriptive metadata elements. Harmonization of the metadata fields using ISOcat concepts has proven to be possible and an automated ingest procedure for CMDI metadata files has been realized. The set-up has been tested against all available CMDI profiles from the CLARIN EU community. While the desired level of integration at the content engine level was originally intended to go further than is currently the case, the implementations and experiences of the participating CLARIN centres provide a basis for further elaboration.

Other Considerations The German CLARIN D-SPIN project has taken on a similar project with the exception that this focuses on the federated content search integration and does not focus on the central metadata aggregation. The basis of this distributed content search approach is however the same, i.e. SRU/CQL with standard CLARIN EU extensions.

S&D Goal 2: National Centre Structure

Introduction At the start of the project the MPI was functioning as the main source of knowledge and expertise about various aspects of infrastructure development. In order to change this situation and stimulate dissemination of knowledge and expertise across all CLARIN-NL centres, the Search & Develop project was seen as a means to work together on a concrete set of tasks and create a more stable centre network. During the course of the project it was indeed observed that other CLARIN centres have gained a stronger position in this area. The INL, for example, have launched their BlackLab solution and the Meertens Institute is hosting a flexible metadata search engine capable of handling CLARIN CMDI files. Experiences with working on the files according to CLARIN specifications have been shared and a more common basis for the CLARIN principles has been established. The experiences within the Search & Develop project have also led to the spread of search technology developed at different centres thus helping to overcome the fragmented state of content search engines at various organizations.

Original Goal A national centre structure. The use cases of the Search & Develop project were also to be used to build up knowledge. At the start of the project the knowledge and expertise about many aspects of the development of a humanities infrastructure was hosted at MPI. Given the CLARIN-NL-centre structure where Meertens Institute, INL, MPI and DANS have the intention and ambition to position themselves as recognized CLARIN-NL centres, this situation should change, in particular with regard to the required redundancy (100% availability). The plan intended to indicate how a more balanced situation can be achieved. Given the state at the time, it was obvious that (a) this process required some time, (b) such a process could only be successful when the partners were working on concrete issues, (c) each of the collaborating centres had to be willing to build up long-term

knowledge and (d) MPI core people with years of experience had to be willing to share their knowledge.

Adapted Goal: no adaptations
Current status: The collaborations between the proposed CLARIN-NL centres through the Search & Develop project have indeed resulted in an increased degree of knowledge transfer. All centres have gained an increased awareness of the technologies and approaches taken by other centres and have even adopted some of the technologies developed at other centres into their own project. The BlackLab technological framework, developed at the INL for example, has been successfully deployed by the Meertens Institute in their CLARIN-NL FESLI project and was tested as part of the Nederlab project.

Towards achieving the (adapted) goal In terms of collaboration and knowledge dissemination the project has been successful in raising the awareness of different technological approaches and the willingness to adopt technological solutions from other centres as a basis within other projects. The proposed shift of knowledge and expertise from the MPI towards other CLARIN-NL centres has provided a more stable distribution and has enabled centres to achieve an independent position in the search technology landscape

Other considerations none

S&D Goal 3: A Leading Position in Europe

Introduction It was the ambition at the start of the project to be recognized as one of the leading countries in the field of federated metadata content search within the European CLARIN community. Though there were differences in approaches between the national initiatives, and the German approach towards federated search was adopted within CLARIN, the members of CLARIN-NL have gained a leading position in the European network.

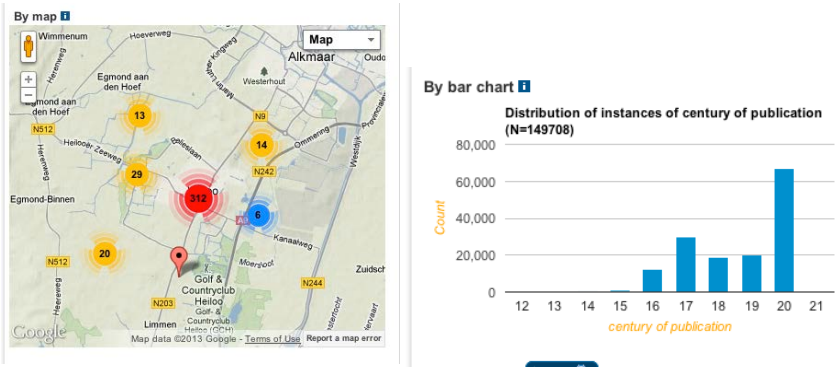
Original Goal A leading position in Europe in the field of federated metadata content search

Adapted Goal no adaptations

Current status

- The CLARIN-NL project has continuously made a contribution towards the specification of the acceptance of a single CLARIN standard (SRU/CQL) for content search integration. This specification serves as the basis for collaboration and implementation at the European level.
- The Search & Develop project has been able to produce a viable alternative to the Virtual Language Observatory (VLO) at the level of metadata search[8] and is the only one of its kind that is able to produce alternative representations from the metadata such as geographical or diachronic distribution.

[8] http://www.meertens.knaw.nl/cmdi/search



From the European workshops on Federated Content Search it has become clear the Dutch CLARIN project and the German D-SPIN project constitute the front runners of the CLARIN community in this area. Within the German CLARIN D-SPIN project a similar search and develop project has been launched. The focus of this project is more geared towards distributed content search engine integration and does not take metadata search into account. The German approach was adopted within CLARIN. See http://weblicht.sfs.uni-tuebingen.de/Aggregator/

This concludes the discussion of the S&D project.

CLARIN Data Providers CLARIN NL aims at disclosing large amounts of language resources to the CLARIN community. To further this end we have contacted organizations storing such resources and being able (with some help) to provide access to them in a CLARIN compatible manner. A range of independent data providers, more specifically *het Nederlands Instituut voor Beeld & Geluid* (NIBG, Institute for Sound and Vision), de *Koninklijke Bibliotheek* (KB, National Library), the *Digitale Bibliotheek van de Nederlandse Literatuur* (DBNL, Digital Library of Dutch Literature), and *Utrecht University Library* (UBU) have expressed their willingness to make their data available in a CLARIN-compatible manner and have carries out projects to realise this. The *Nationaal Archief* (NA, the National Archive), originally also planned to submit a plan for this, but it has informed us that it did not have the capacity to do this in the short term.

CLARIN-NL was successful in connecting the national data centres mentioned. The work by DBNL was fully incorporated in Nederlab, and all other organisations provide CLARIN compatible metadata and make them available for harvesting so that they are visible via the CLARIN VLO. For the National Library Metadata this created a problem, since the amount of metadata they provide overwhelmed all other metadata, and revealed the problem of granularity of the metadata. The National Library metadata are therefore temporarily not visible in the VLO until a more principled approach to granularity and collections of resources have been designed and implemented.

Metadata Project Work on the CLARIN infrastructure has started with the *Metadata project*. It builds on work done in the CLARIN preparatory phase, which has delivered a CMDI registry (further developed in the CLARIN-NL IIP project). This registry allows one to search for existing metadata components and profiles, or create new ones if needed. The project has created several metadata components and profiles for data that are available in the Netherlands. For the metadata elements contained in the metadata components, links have been made to the ISOCAT data category registry,

and where needed this registry has been extended with new data categories, in order to ensure semantic interoperability. The work done in this project formed a good basis for the creation of metadata for additional data in data curation projects (see section 7, page 25).

CLAVAS Because of scaling and subsidiarity reasons, it has become clear in the IIP project that ISOCAT cannot function as the only concept registry in the CLARIN infrastructure. For certain concepts ISOCAT is not particularly suited, and for other concepts independently maintained registries exist and duplication is to be avoided. For this reason, a separate project, CLAVAS (CLARIN Vocabulary Service), was started up to build an interface to a variety of already existing concept and data category registries, so that CLARIN users can transparently use concepts and data categories from each of them. One part of the project, automatic normalization of organisation names, was not successful. It was decided to have the relevant work done manually by the Data Curation Service. In other respects, the CLAVAS functionality is available and was also used by CLARIN in Austria. The CLARIN Vocabulary Service is a running instance of the OpenSKOS exchange and publication platform for SKOS vocabularies. This OpenSKOS instance currently publishes SKOS versions of three vocabularies:

- ISO-639-3 language codes, as published by SIL.
- Closed and simple Data Categories from the ISOcat metadata profile.
- A manually constructed and curated list of Organizations, based on the CLARIN VLO

CCR From the start of CLARIN, the ISOCAT registry for data categories was used as the major instrument for achieving semantic interoperability. During the course of the project it became clear that ISOCAT had many disadvantages, and alternatives were being considered in several workshops. One of the alternatives was to turn to a concept registry based on SKOS, building forth on the work done for CLAVAS. When MPI/TLA decided to stop hosting ISOCAT as part of its reduction of work related to CLARIN A-services, an alternative was urgently needed. This alternative, the CLARIN Concept Registry was created in a separate project by Meertens Institute and was financed by CLARIN-NL. It is now fully functional and will be further used and developed in CLARIAH-CORE.

Metadata for Tools A small project to create a CMDI *metadata scheme for tools* has been carried out. Initial results were provided and tested against a small number of tools. The developments of the CLARIN Portal made it necessary to make further adaptations, for which there were no resources in the CLARIN-NL project. This topic has been taken up again in CLARIAH-CORE.

Testing the infrastructure against real data and tools Calls have been issued for data curation projects (see section 7, page 25) and for demonstrator projects (see section 8, page 30). In the approved projects, data and tools developed and in use in the Netherlands were curated and a series of demonstrators were developed on the basis of existing technologies and applications to attempt to make them CLARIN-compatible. A major aim with these projects is to test the standards proposed within CLARIN against these data and tools, and to be able to provide evidence-based feedback on the proposed standards and thus to influence their character and role in the CLARIN context.

In each project metadata in accordance with CMDI has to be created and tested, making use of the experience gained and the tools developed in the Metadata project described above (page 21). Each demonstrator project will create a demonstrator web application, in which a clear separation between user interface modules and core modules is made. An API has to be defined and documented for the core module. The user interface will make calls to the core module via this API. This will make it possible later to easily turn the core module into a web service. Some functionality was created directly as a web service.

All projects have to deliver a document describing the requirements the infrastructure should meet and or desirable features it should offer, and form input for the requirements and specifications of the CLARIN infrastructure.

In this way, the infrastructure under construction is constantly tested against data and tools that actually will be used by the researchers, and its design can be adapted to meet the relevant requirements and desirable features. These documents have provided us with a wide range of requirements, desiderata and recommendations. An analysis of these documents from the projects from the first two calls has been made and concomitant actions taken, as described in [Odijk 2011, Requirements and Desiderata]. Requirements and desiderata were formulated for technical subjects but also for IPR and ethical issues. Technical subjects addressed include ISOCAT, Web Services, Metadata, and to a lesser degree, standards, PIDs and search. The results of an analysis of the requirements and desiderata for projects from Call 3 and later has been taken into account in developing the proposal for CLARIAH-CORE.

Is the distributed nature of technical infrastructure indeed invisible to users?

The Netherlands part of the CLARIN infrastructure is made visible to its targeted users through the CLARIN-NL Portal. This offers a single entry point, so that users do not have to worry where data are located or where software service are running. Faceted search leads the users to the desired data and applications, and hyperlinks bring them to the actual data and applications. The CLARIN-NL portal also brings users to the whole CLARIN infrastructure, e.g., to the Virtual Language Observatory (VLO), which makes the distributed nature of the infrastructure at the European level invisible. The VLO, which was originally created in the CLARIN preparatory project but has had several upgrades since is in a sense “too successful”. It currently hosts over 880k metadata records but the faceted search offers only limited ways for searching in them. This is in part caused by lacking or badly represented metadata elements. The problems that VLO poses for actually finding data that one is interested in were known among the experts and were clearly documented in [Odijk 2014]. CLARIN ERUC has since worked on several improvements of VLO, and the Netherlands (in the CLARIAH-CORE project) and Austria are closely collaborating on further improvements, so that the VLO can properly serve the goal it is intended for.

Is the performance of the technical infrastructure sufficiently good?

Initial versions of all ingredients for the technical infrastructure are available (tools, web services, web applications, infrastructure services) and are findable and accessible via the CLARIN-NL portal. Actual use is made of the services, in most cases satisfactorily, though some problems were

encountered with a small number of applications. For example, the TTNWW application was often down, and changes at SURFSara have made things worse. Performance issues arose with the search application OpenSONAR when twenty students worked on it at the same time and with particular types of queries (types that linguists, however, need very often). Many suggestions from actual users for additional functionality and improvements of the user interface were gathered (esp. for applications for search in linguistically annotated corpora). These issues have been taken up by CLARIAH-CORE, which is working (inter alia) on improved versions of PaQu, GrETEL and OpenSONAR.

Many ingredients work well (e.g. various demonstrators, Arbil tools, CMDI registry, ISOCAT ,the VLO, metadata search) though for some improvements of the interface have been suggested and metadata harmonisation will provide better results when browsing in the VLO or searching in metadata. Federated content search has been made available, but it so far has severe limitations, so that usage of resource-specific engines remains necessary. Federated Search will be taken up again in CLARIAH-CORE.

AAI has been implemented, but it was difficult to get every research organisation in the Netherlands to allow its employees and students access to the CLARIN Service Federation, not for technical reasons but for reasons of security and privacy policy.

Metadata for tools & services is still to be provided in a more systematic way. A CMDI core profile for web-services has been developed in collaboration with other European CLARIN partners, however more refinement is needed. Incorporation of the metadata for tools in the VLO, with dedicated facets, is highly desirable and has been taken up by CLARIAH-CORE.

Summarizing, a lot of functionality has been created, is available, and is actually used (in varying degrees). Most problems and suggestions for improvement are being addressed by CLARIAH-CORE.

7 Data Infrastructure

Are the data identified as “essential” included in the technical infrastructure?

We have actually not identified specific data as “essential” at the beginning of the project, because we believed we were unable to do so. However, we have set up a number of activities that lead to inclusion of data that are actually needed and used by humanities researchers, which is of course an important criterion for identifying data as “essential”. In addition, the data must meet the requirements stated in the long-term work plan and additional criteria set by the CLARIN-NL Board (e.g. genericity of the data, size of the user base, etc.).

We describe these activities here in more detail.

For the Data Infrastructure, a whole range of activities has been carried out.

First, a metadata infrastructure (CMDI), initiated in the CLARIN preparatory project, has been finalized and is in use. See the description of the Metadata project above (page 21).

Second, a whole range of data curation projects have been started up, based on project proposals for data curation.[9] 21 pure data curation project proposals were submitted, 12 were approved and have been carried out or are currently running. In addition, many projects included a data curation component: 30 of such mixed projects were submitted and 16 were approved and have been carried out or are currently running. Open Calls were organized in which proposals for such projects could be submitted. The projects were relatively small (max 120k euro), and had a maximum lead-time of 1 year. The projects were required to involve a targeted user (a humanities researcher), and the projects had actually to be coordinated by a humanities researcher. These requirements ensure that data are curated that a humanities researcher is really interested in for his/her research, thus providing natural priorities in the selection of data to be curated. We refer to the fact book for a complete overview of these projects.

The goal was to cover language resources from various humanities disciplines, with an emphasis on linguistics. The following table provides an overview of the projects by discipline:[10]

Count of Project-Name	Column Labels		
	yes		
Row Labels	Data Curation	Mixed	Grand Total
History	1	5	6
Linguistics	14	16	30
Literary studies	2	3	5
Religion studies		2	2
Grand Total	17	26	43

Certain disciplines (e.g. philosophy, media studies, and culture studies) are lacking here but most are covered by demonstrator projects (see section 8).

The following table specifies the subdisciplines covered for history, linguistics and literary studies.

Count of Project-Name	Column Labels		
Row Labels	Data Curation	Mixed	Grand Total
History	1	5	6
History of WW2		1	1
Mediaeval studies		1	1
Naval history		1	1
Oral history	1		1
Parliamentary history		2	2
Linguistics	14	16	30
Dialect studies		1	1

[9] An overview of these projects can be found in the CLARIN-NL Fact Book.

[10] A single project can deal with data that are relevant for multiple disciplines, and that has been taken into account in the numbers given.

Discourse studies	2		2
Historical linguistics	2	1	3
Language acquisition	1	2	3
Language attrition		1	1
Language Documentation	2		2
Language typology		1	1
Lexicography		2	2
Lexicology		2	2
Morphology	1	2	3
Phonetics	1		1
Second Language Acquisition	1		1
Semantics		1	1
Sign language	2		2
Syntax	2	3	5
Literary studies	2	3	5
Arthur novels		1	1
Emblem Studies	1		1
Literary reception		1	1
Mediaeval studies		1	1
Song Studies	1		1
Grand Total	17	24	41

One can see that many subdisciplines of these three disciplines are covered.

The following table gives the count of the projects by medium/modality:

Count of Project-Name	Column Labels		
Row Labels	Data Curation	Mixed	Grand Total
Yes	12	16	28
Audio-visual	2		2
Audio-visual + text	1	1	2
Speech	2		2
Text	7	15	22
Grand Total	12	16	28

Not surprisingly, text is dominant but other media/modalities are covered as well and there are some multimodal projects.

Third, the Data Curation Service (DCS) has been set up. The DCS has created an inventory of projects that generated resource collections that are potentially eligible for curation. This has yielded a list of some 150 resources from around 350 projects. Some of these have been prioritized for curation, for

others more information must be collected before it can be decided whether curation is indeed desirable and feasible. We refer to section 10 for an overview of the curation work carried out by the DCS.

All in all, around 35 resources out of the 150 identified by the DCS will have been curated by the end of 2013, which is a good start but also makes clear that we are still far off from curating all resources identified by the DCS. In addition, the DCS list is not complete, and many more resources exist.

Fortunately, other projects make their contributions too, in particular the *Nederlab* project that focuses on historical Dutch data, especially text corpora. And an increasing number of research projects create their data directly in a CLARIN-compatible manner.

Resource descriptions can differ dramatically in granularity and, concomitantly, number. For example, the C-DSD project has made available some 243k CMDI metadata objects for the *Song Database*, while other projects just provide a single metadata object for a big resource.

Fourth, as mentioned above (page 21), a range of independent data providers have, upon our invitation, expressed their willingness to make their data available in a CLARIN-compatible manner. They made available a significant range of resources:

Data Provider	#	Unit
National Library (KB)	2 million (and 16 million in the pipeline)	Metadata objects
NISV (NIBG)	46,156	Metadata objects
DBNL	>3,000	Editions
UBU	64,641	Metadata objects

Fifth, the CLARIN centres are working on making all their existing resources and resource descriptions available in a CLARIN-compatible manner, as part of their activities for certification as a CLARIN-centre. For example, MPI has converted all its IMDI metadata records to CMDI. The Virtual Language Observatory (VLO) contains at least 377k[11] metadata objects for resources created in the CLARIN-NL project.[12]

Fortunately, for most of the data from the HLT Agency (which moved from INL to DLU and now back again to INL/INT), among which many data from the STEVIN programme have, metadata that are visible in VLO.

Summarizing, many data that have proven important to research hers and are actually used in research have been made visible and available in the CLARIN infrastructure. Many other data are still lacking. Some of them will be addressed by CLARIAH-CORE, and by research projects that need these

[11] See http://catalog.clarin.eu/vlo/?fq=nationalProject:CLARIN-NL measurement date: 2016-04-27

[12] This number is most probably an underestimation, since not all metadata are properly marked as coming from the CLARIN-NL project and important metadata sets (e.g. from the National Library) are not yet included in VLO.

data. The data that have been included or are at the point of being included arguably contain the *essential* data, since the selection has in part been driven by researchers, and in part consists of data available via data providers already operating before the start of the CLARIN-NL project.

Can they be easily found and accessed by users, and are they properly documented?

The data are currently visible and accessible to the users in two ways:

- Via the *Virtual Language Observatory*, with inter alia a faceted browser for language resources in which one can search via a variety of facets, including language, collection, resource type, genre, national project, and others; geographical access to the "Virtual Language World" via Google Maps; and access to the CLARIN LRT Inventory[13], an earlier data and tools overview which will slowly be deprecated in favour of the VLO.
- Via the metadata search engine created at the Meertens Institute (see page 18) capable of ingesting diverse CMDI metadata records and harmonizing the metadata fields through the ISOcat Data Category Registry.

In this way, both browsing and searching in metadata records is offered. Actual usage of the VLO has shown, as described above, that the overwhelming number of metadata records, the highly varying quality of the metadata, the highly varying granularity of the metadata, and the limited number of facets one can use for search do not make it easy to find the data one is interested in, especially of these data are not known yet to the user and must (so to speak) be 'discovered' [Odijk 2014]. These issues are well-known and CLARIN ERIC, CLARIAH-CORE and CLARIAH Austria are actively addressing these issues.

All resources are well-documented in the sense that they have been assigned CMDI metadata, which is a prerequisite for being visible via the VLO and the metadata search engine. In data curation projects, further documentation has been brought to a proper level where this was needed. For most legacy data that were already in CLARIN-supported standard formats, the legacy documentation is available. For most of the data, the associated documentation is of a high level since the data were produced in dedicated projects.

Have NL-specific IPR issues been adequately dealt with? Have procedures been defined to deal with IPR issues for new data?

We have initially focused on data and tools for which there were no IPR-issues. Settlement of IPR-issues was a pre-requisite for data curation and demonstrator projects, and we have rejected some projects because they did not meet this requirement. For some projects, new agreements have been concluded with participants in the recordings to ensure reuse of the data by the CLARIN research community.

We intentionally focused on data and tools without IPR-issues for a number of reasons. First, our initial focus was on technical issues around making data and tools CLARIN-compliant, and we wanted

[13] http://lrt.clarin.eu/inventory

to invest effort only in data and tools of which we were sure in advance that they could be made available to the CLARIN research committee. Second, when we started, the CLARIN preparatory project had not finished yet, and we first wanted to await the conclusions and recommendations of the IPR work package of this project. Third, for many data and tools for the Dutch language IPR issues had been settled in the STEVIN programme, so the problem was less urgent.

In this way, we have so far (almost) avoided having to deal with IPR issues. Nevertheless, CLARIN-NL has formulated its policy with regard to IPR, and promotes Open Access, with specific provisions to cater for the interests of individual researchers (e.g. allowing delayed publication of data).[14]

However, this policy does not solve problems with legacy data. Some examples of this came up during the project. NIBG makes its data available under the "Academia"-license, which is a paid license between NIBG and academic organisations in the Netherlands. We also allowed that NIBG does this in the CLARIN-context. Not all academic organisations have such a license, and access so far is restricted to organisations in the Netherlands. So these data are not available to the whole CLARIN research community, a situation that is not desirable. This problem, and others are coming up in CLARIAH-CORE as well and are being addressed there. Other concrete examples exist in the context of the data curation service (see section 10, page 40), and in the Cornetto-LMF-RDF project. The latter deals with a lexico-semantic resource that includes data from a commercial party (Van Dale). In principle, IPR issues were properly taken care of here, but Van Dale has decided to stop the licensing contract after the Cornetto-LMF-RDF project had already started. For this case, a project has been started up (financed with remaining STEVIN funding) to replace the Van Dale data by public data, so that the IPR-issues vanish and the dependency on unreliable and unstable commercial organisations is avoided. However, the coverage of the new resource (Open Dutch WordNet) is not as good as the original source which contains the Van Dale data.

Ethical issues have played a role with regard to certain data. It mainly involves the issue of the identifiability of participants in recordings. In metadata, the problem has been solved by having two sets of metadata: anonymized metadata that are public, and non-anonymized metadata that are accessible only after special permission. In the actual data, identifiability cannot be avoided, but researchers have to sign a specific license before they can get access to the data.

Are there guidelines and procedures in place as well as supporting tools to easily incorporate new data into the infrastructure?

All centres have clear guidelines and procedures to incorporate new data into the infrastructure, in particular MPI and DANS. These procedures are supported by (partially) automated ingestion software, LAMUS in the case of MPI/TLA and EASY in the case of DANS. After MPI decided to reduce its activities and support for LAMUS fell away, the organisations united in TLA started developing a light-weight alternative to LAMUS called FLAT (Fedora Language Archiving Tool). As part of this effort, Meertens has carried out a pre-study for this financed by CLARIN-NL (MDF subproject).

[14] http://www.clarin.nl/file/19039 (in Dutch)

8 Language technology infrastructure

Have the tools and services identified as "essential" been included into the technical infrastructure?

On the one hand we have identified a specific set of tools developed in the STEVIN programme as "essential", and set up a project (in cooperation with Flanders) to include them in the CLARIN infrastructure.

On the other hand we have adopted the same strategy as with the Data Infrastructure: we have set up calls for proposals for demonstrator projects from which we selected projects dealing with language technology that was proposed by a humanities researcher and that he/she actually needs or uses, and that meets requirements stated in the long term work plan and additional criteria set by the CLARIN-NL Board (e.g. genericity of the technology, size of the user base, etc.).

For setting up the language technology infrastructure, a range of activities have been carried out. We list them here and describe them in more detail below:

- Demonstrator projects
- CKCC project
- TTNWW project, in which existing language and speech technology components have been converted to web services in a workflow system. See also section 5.

We discuss these activities one by one in more detail below. Statistics on these projects will be presented after these individual descriptions, since the statistics have been calculated on the whole set of projects.

Demonstrator Projects We have issued several calls for demonstrator projects. In response to these calls, 29 project proposals were submitted for demonstrator projects, out of which 19 could be awarded funding. In addition, 30 mixed demonstrator and data curation projects were submitted, out of which 16 could be awarded funding. Open Calls were organized in which proposals for such projects could be submitted. Two projects were initiated directly by the CLARIN-NL executive board (OpenCGN and CorpusSearchWeb). The projects were relatively small (max 120k euro), and had a maximum lead-time of 1 year. We refer to the fact book for a complete overview of these projects. One of these projects, the CKCC project, had a special character and is therefore described separately here.

CKCC-project The CKCC project (*Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic*, 'Geleerdenbrievenproject ') is an independently financed (NWO) project that is carried out by a consortium of Dutch institutes. It investigates, on the basis of a corpus of 20,000 letters of scientists from the 17th century in the Dutch Republic and using language technology, the research question of how knowledge circulated in the 17th century. It was selected in the CLARIN-EU call for humanities and social sciences projects as the project proposal that "[would] best demonstrate the use of LRT and would show the potential of a research infrastructure in the

humanities" (CLARIN Newsletter 6, p. 3).[15] It has received funding from CLARIN-NL to apply language technology in the project (esp. part of speech tagging) and to make their results CLARIN-compatible. The project is about to finish.

TTNWW project Components from the previous projects in the area of text and speech analysis have been converted to web services and combined in standard workflows using the Taverna workflow engine. A front end has been provided allowing users to upload text or speech files and invoke specific tasks from a temporary workspace. All intermediate and end results are stored back into the user's workspace for further inspection. The application can be found at: http://yago.meertens.knaw.nl/apache/TTNWW/. We discuss the goals of the TTNWW project and their status one by one.

TTNWW Goal 1 The original goal of the project was to include existing components from, amongst others, the CGN[16] and STEVIN[17] projects as web services into a workflow system. This should be running on servers of recognized CLARIN centres. The system should provide facilities for HSS researchers with little or no technical background. These facilities should enable them to answer research questions better or more easily and provide opportunities to propose new types of research questions, that is, research questions that could previously not or not effectively be answered.

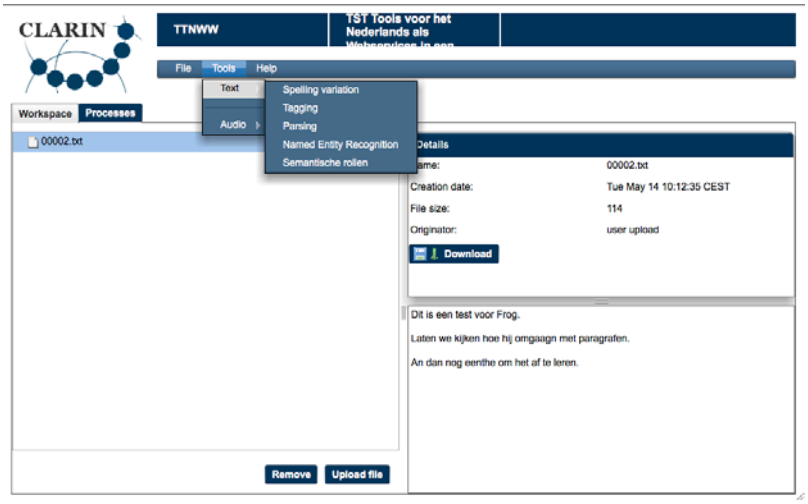
Adapted Goal Two problems arose during the project:

- During the course of the project it became clear that hosting of the web services in dedicated CLARIN centres became an expensive exercise that had not been taken into account into the original project proposal. To accommodate for this, a separate project was launched to investigate the possibilities of hosting the web services in a cloud environment provided through the BigGrid project. The goal of this was twofold: first to gain experience with operating the workflow system and web services from within a cloud environment and secondly to test out the initial principles of dynamic service deployment. The latter goal was inspired by the realization that these types of web services would not need to be operational 24/7, and would only periodically be used. Dynamic service deployment, where a server is only instantiated when requested could provide a lower footprint of the necessary server capacity. While the principles of this approach were successfully tested the current set up of the project relies on continuously available servers, mostly from within a cloud environment.
- Conversion of previously created standalone applications to web services has proven to be a major step by some participants. To make this easier the Computational Linguistics Application Mediator (CLAM)[18] was promoted through the project allowing technology providers to more easily convert their applications to RESTful web services.

Status: All components from the project have successfully been converted to web services, either with the help of CLAM or otherwise. Most text analysis services are running in a cloud environment

[15] See the CLARIN Newsletter 6: http://www.clarin.eu/files/cnl06_web.pdf
[16] Corpus Gesproken Nederlands (Spoken Dutch Corpus) http://lands.let.ru.nl/c
[17] http://taalunieversum.org/archief/taal/technologie/stevin/
[18] http://proycon.github.io/clam/

under the control of the Meertens Institute. Speech recognition services are currently still under control of the respective technology providers and run on their servers, mostly to allow them to monitor the performance of these services.



Taverna has been used as a workflow system. The desktop version has been used for development of the workflows, connecting the web services. For the production environment the server version of Taverna is used. One drawback of this approach is that the Taverna server version does not provide advanced support for process monitoring and fail over options. This proves to be a disadvantage at times where requested services are temporarily unavailable.

The workflows that have been created all target specific tasks, such as orthographic normalisation, part of speech tagging, parsing, named entity recognition, semantic annotation and speech transcription. These tasks have been combined in an easy to use user interface where the user uploads his/her data and selects the workflow to perform from a pull down menu.

The main technical goal has been achieved. The application is available since 2013, has actually been used, but in part due to external circumstances it has often been down, and new hardware and software choices at SURFSara require a reimplementation of certain parts, which will be taken up in CLARIAH-CORE. It enables HSS researchers to enrich textual and speech data with rich linguistic annotations. But that is only a first step. The next thing a researcher would want to do is search in these data for these linguistic annotations, and that is functionality TTNWW does not offer. The existing applications for search in linguistically annotated corpora originally allowed search in pre-compiled corpora only. In order to solve this problem, new subprojects were started for extending search applications with functionality to upload one's own enriched corpus and search in that corpus. This has been done for the Groningen LASSY Word Relations Search Application, resulting in the PaQu application, and for the Corpus Hedendaags Nederlands application, resulting in the Autosearch application. It was requested for the GrETEL application, but no (human) resources were available to carry out the work. This functionality is being added in the CLARIAH-CORE project. The combination of TTNWW and these search applications that allow a researcher to upload one's own corpus enables linguistic research of various kinds, e.g. the type of research described by [Odijk

2015]. This functionality is also used in an UU-internal project called *Anncor* for enriching various language acquisition databases with syntactic annotations.

Other considerations Within the German CLARIN D-SPIN project the WebLicht environment has been developed to include web services in a workflow engine. The approach taken in the WebLicht project is however slightly different from the approach taken in this project. While the WebLicht project originally intended to provide users with the possibility to construct and execute their own workflows, different user groups, i.e. developers and end users, handle construction and execution steps in the TTNWW set up. The difference in approach was grounded in two main considerations; firstly it was not expected that the intended end users, HSS researchers with little or no technical background, would be willing or interested in constructing workflows themselves. Another consideration is that the workflow execution engine itself remains completely hidden from the end user. This makes it possible to replace the workflow execution engine in the architecture without affecting the end user experience, creating additional technical flexibility. Workflows themselves would also need to be redefined if a different workflow definition language is used but this also remains hidden from the end user's view.

TTNWW Goal 2 An important goal of the project was to promote commonly used, *de facto* standards for data formats, tools interfaces and protocols within the CLARIN community, and to compare these with proposed standards within CLARIN. Through this action the notion of CLARIN compatibility, under development within the CLARIN community, was to be reinforced and further extended.

Status During the project, *Folia* emerged as a *de facto* standard for linguistically annotated corpora. It was adopted for most annotation types in the annotated text corpora in the TTNWW project, and also used in various other CLARIN-NL subprojects (e.g. VU-DNC). However, for the annotation of syntactic parses of sentences the *Lassy* format has remained. In the CLARIAH project, FoliA is being extended so that it can also accommodate Lassy-style annotations and convert from Folia to the Lassy format. Folia has also been used in independent projects that created linguistically annotated data (e.g. Basilex). It is to be expected that Folia will continue to be the most important format for linguistically annotated corpora in the Netherlands and Flanders in the coming decades.

With respect to tool interaction interfaces, most participants have opted to use CLAM to wrap their tools. The interaction pattern associated with this wrapper provides predictable and easy to use interface behaviour for web service interaction as well as a standalone user interface option. Some problems appear to be associated with the processing of large (audio) files, but these remain under investigation.

Though full convergence onto a single CLARIN-NL linguistic annotation standard did not prove to be attainable, the project has seen a strong convergence of all participants onto the *Folia* and *Lassy* as *de facto* standards. However, the formats used are so far limited to linguistic annotation of Dutch language text corpora and in use in the Netherlands and Flanders, but not beyond.

Other considerations From an international perspective a number of linguistic annotation standard proposals have been submitted. Through ISO's TC37 the Linguistic Annotation Framework (LAF), Morphosyntactic Annotation Framework (MAF) and Semantic Annotation Framework(SemAF) a

raised awareness towards the need for standardization has been achieved. Within the CLARIN EU context the German D-SPIN project has positioned their Text Corpus Format(TCF) as a contribution towards this.

Statistics on the Language Technology projects

The distribution of the language technology projects over disciplines is as follows:

Count of Project-Name	Column Labels		
	yes		
Row Labels	Demonstrator	Mixed	Grand Total
Archaeology	1		1
General	4		4
History	6	5	11
Linguistics	18	16	34
Literary studies	2	3	5
Media Studies	1		1
Philosophy	1		1
Religion studies		2	2
Grand Total	31	26	59

This is a good distribution in the various humanities disciplines, with again, as with the data curation projects, linguistics and history represented best.

The following (language) technologies played a role in these projects:

Count of Project-Name	Column Labels		
Row Labels	Demonstrator	Mixed	Grand Total
yes	44	19	63
Annotation	1		1
Diarisation	1		1
Lemmatization	2		2
Machine translation	1		1
Multilingual interface	1		1
Multiword identification	1		1
Named entity recognition	2		2
Network analysis	1		1
Optical character recognition	1		1
Orthography normalization	2		2
Part of speech tagging	4	1	5
Phonetic distance technology	1		1
Referring to time-based resources	1		1
Search	7	15	22
Semantic annotation	1		1

Speech alignment	1		1
Speech recognition	2		2
Syntactic parsing	2	1	3
Tokenization	3		3
Topic modelling	1		1
Visualisation	10	2	12
Grand Total	44	19	65

Not surprisingly, quite generic technologies such search technology and visualisation technologies play a prominent role, but also most basic language and speech technologies are present.

The number of projects by medium/modality is represented in the following table:

Count of Project-Name	Column Labels		
Row Labels	Demonstrator	Mixed	Grand Total
Yes	19	16	35
Audio	1		1
Audio-visual	4		4
Audio-visual + text		1	1
Speech	1		1
Text	13	15	27
Text + speech	2		2
Grand Total	19	16	37

We see that the most common media/modalities are covered and that some projects deal with multiple media/modalities.

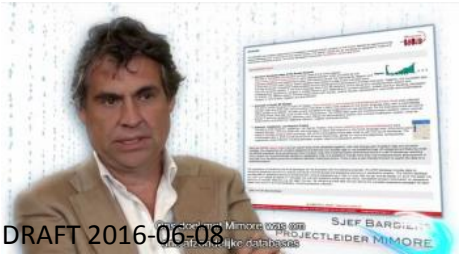
This concludes the statistics on the language technology related projects.

In conclusion, we can state that the technologies dealt with include the "essential" ones. We have systematically covered language and speech technology tools from the CGN and STEVIN programmes. As with data curation, the demonstrator projects always involved a humanities researcher as user, so that the selection of technologies was driven by the needs of the targeted users. And we supported an independently financed research project, again clearly addressing needs that are considered "essential" or at least highly desirable by the researchers involved.

Can they be easily found and accessed by users, and is there proper documentation to assess their appropriateness for a given task?

Visibility and accessibility of the tools is currently taken care of mainly via the CLARIN-NL portal. The section 'services' provides an overview of the services offered and faceted search to discover services that are interesting to the user. One can search for facets such as *research domain*, *language*, *tool task*, and others so that it is relatively easy to find services of which the user does not know the existence before the search.

However, the CLARIN-NL portal and the faceted search is made in an ad-hoc manner, and it covers services created in the Netherlands and Flanders only. The information contained in the portal should be encoded in CLARIN-compatible 'metadata' for these services, and then also become available via the VLO. A start with this has been made in CLARIN-NL and the work is currently continued in CLARIAH-CORE.



Concerning documentation, each demonstrator project was obliged to provide a *demonstration scenario* that illustrates the basic functionality of the demonstrator by using a concrete example, and by guiding the user through the

various steps that must be taken to use the demonstrator functionality. This surely can serve as a good instrument to assess the appropriateness of a tool for a given task. In addition, each project was obliged to deliver documentation, not only for the user but also for developers of the technology itself, and for application developers who want to use the software in other applications. Several projects have made movies or screen captures to illustrate their work or their demonstrators. See http://www.clarin.nl/node/185 for a list of such videos.

In addition, for certain services, educational packages have been made that can be used to teach students in the use of these services. They can be found on the CLARIN Educational Packages section of the CLARIN-NL Portal.

Has actual interoperability among tools/services and between tools and data been achieved in the technical infrastructure?

We have made serious attempts to achieve interoperability among tools and services, in particular in the TTNWW project, but also in demonstrator and data curation projects by requiring conformance to CLARIN-supported standards (for formal interoperability) and by the requirements on semantic interoperability (ISOCAT, later CCR). And we have achieved some partial successes, e.g. in the TTNWW functionality, by the fact that the Folia format is being used beyond the TTNWW project (e.g. in the VU-DNC project), and by the fact that an increasing number of new tools is being created that operate on the resources in Folia format. On the other hand, there are still many different formats, even for a restricted domain such as linguistically annotated text corpora: we have described that the *Lassy* and Folia formats could not be unified so far, in Europe many more different formats for linguistically annotated text corpora are in use, and with the Folia format we now encounter differences with earlier data formats created by more or less the same researchers: for example, there are now (small but real) differences between SONAR (the Dutch written corpus, in Folia format) and the Spoken Dutch Corpus (CGN), which makes searching across these resources difficult, and software that applies to data in CGN-format does not apply to data in Folia-format and vice versa. This problem has partially been addressed by the OpenCGN subproject in which the OpenSONAR search application has been extended to search in CGN as well.

We also started up a subproject (OpenConvert) to develop converters from formats often used in the real world into CLARIN standard formats. It created an initial set of converters, more specifically to convert from plain text, HTML, ePub, Word 1997-2003, Word 2010, and ALTO into TEI or FoLIA.

So it is clear that we still have a long way to go. The issue has been raised at the European level as well, in particular in the CLARIN Standards Committee, and some further initial steps have made and will be continued in the CLARIAH-CORE project.

9 User Needs

Are the data and tools identified as "essential" indeed the data and tools the intended users need most?

We have carried out a User Survey, which took the form of dialogues with selected researchers from a wide variety of disciplines. We chose for the dialogue form because we expected that most humanities researchers would have insufficient knowledge to assess the potential of the CLARIN infrastructure and the tools and services it offers without any explanation of this potential. And the CLARIN developers had insufficient insight into the actual research questions and data and tool needs of these researchers. In the dialogue, the researcher had the opportunity to explain his/her research topics, the data and tools that he/she uses and/or would like to use. In an interaction with the CLARIN experts opportunities for using functionality offered or potentially offered by CLARIN to address these research topics could be explored. This User Survey also resulted in an overview of the resources in use by the researchers and a brief description of these resources, and in a raised awareness of the CLARIN-NL project and the CLARIN infrastructure among these researchers. Nine of the (app. 36) interviewed researchers actually got involved in one or more CLARIN-NL subprojects.

As described above in section 7 and section 8, the selection of the data and technologies dealt with has to a large extent been driven by users, so their match with the user needs is more or less guaranteed in this way. In the project proposals in response to the open calls for data curation and demonstrator projects certain disciplines were not represented or underrepresented. We stimulated proposals from such underrepresented disciplines by changing the prioritized disciplines from call to call. In addition, for disciplines where even this had no or insufficient effect, we set up closed calls where specific researchers from underrepresented disciplines were invited to submit a project proposal with which they were not in competition with other researchers. Researchers from philosophy, media studies, (classical) philology, language attrition, and religion studies were invited for such closed call proposals. In this way we could ensure that data and/or tools for all humanities disciplines were covered in the CLARIN infrastructure

Is the technical infrastructure indeed used by the intended users?

Are the intended users happy in using the technical infrastructure?

Do the interfaces promote and stimulate working with it or do they pose obstacles?

Are the interfaces user-friendly and self-explanatory where possible?

How many researchers actually use the infrastructure on a regular basis or occasionally? (Target: 40% of the intended users regularly use it; 70% use it at least occasionally)

We treat these 5 questions together, for the following reason: though several parts of the infrastructure are being used by our targeted users, it is difficult to quantify this precisely. At this moment, there is also not (yet) a way of measuring whether and how often the CLARIN infrastructure

is used. Though CLARIN ERIC has started using PIWIK and proposed the CLARIN centres to use the same tool, this turned out not to be so easy, for both technical and policy-related reasons:

- Most centres already use a system for obtaining metrics, and in most cases it is not PIWIK (but OWA or Google Analytics).
- Even if they use PIWIK, the reports/data they produce are not necessarily the ones that CLARIN is interested in
- It is therefore better (and necessary anyhow) to propose clear guidelines on what has to be measured and how this has to be represented, independent of the tool to obtain this information. How the centres obtain this information is their concern.
- There should be clear statements on what CLARIN ERIC, or people working for CLARIN ERIC are allowed to do and are not allowed to do with the obtained metrics, and one should get the buy-in from the centres for this (otherwise they will not provide you with the information)

We are not aware of any other actions to systematically measure the actual usage.

Nevertheless, it is clear that the CLARIN infrastructure is increasingly often used in actual research and in education. Research that crucially used the CLARIN infrastructure is now appearing in the *Lingua* Special Issue on the CLARIN Infrastructure with Jan Odijk as guest editor. We are approached by many researchers when they are preparing new project proposals. They want to use CLARIN tools and applications, and they want to make sure that data or software resulting form their research projects is CLARIN-compatible. The research project proposal by Van Koppen and Dietz for the NWO *Vrije Competitie* programme, in which we contributed to the preparation and in which some of us participate actively was awarded funding. CLARIN (JO) is involved in a project proposal for the NWO *Gravitation* programme that is being prepared by a large consortium of linguists united in LOT, and, if it is awarded, the researchers will make heavy use of the CLARIN Infrastructure and also contribute to it. The first PhD in which crucial use is made of CLARIN has been successfully defended by Liesbeth Augustinus in Leuven (18 October 2015). Though this is a PhD in Flanders, the work on CLARIN in Flanders was made possible by the close collaboration of Flanders with CLARIN-NL. CLARIN-TCC is taking preparatory actions to make a European research proposal around parliamentary data in Europe.

There are some initial experiences with usage by the targeted users. For example, the metadata search interface was, upon first exposure, considered too difficult by some users in the Nederlab project, and this has led to a complete redesign of the interface that will be used in Nederlab.

It has also become clear, as stated before, that the functionality offered by federated search is too limited for many researchers. As a consequence, the CLARIN infrastructure continues to offer resource-specific search engines in the coming years, though CLARIAH-CORE addresses the issue again in slightly different manner.

Occasional users have pointed out small bugs in software, or pointed out that certain functionality is desirable but not offered yet. A concrete example is a way of searching for tokens in the MIMORE databases that are not inflected, which is currently not possible: it requires either the

implementation of a NOT operator, or the adaptation of the annotations so that absence of inflection is explicitly marked.

10 Centres of Expertise

Have one or two centres of expertise been created and are they operating successfully?

Two centres of expertise have been created, Nederlab, and the Data Curation Service.

Nederlab

The *Nederlab* project is funded independently and aims to provide data and tools for the longitudinal study of the Dutch language and culture. It has also been supported financially by CLARIN-NL and has been set up as the second (virtual) CLARIN-NL centre of expertise, more specifically on data and tools for the study of the Dutch language and culture across time. Within the Nederlab project large amounts of historical data are curated, and their metadata created or curated. A dedicated search application has been constructed to search in the data, their metadata and in annotations on these data on multiple tiers. A first version of the Nederlab application was demonstrated at the CLARIN-NL final event, and extended versions that incorporate parts of the multitier annotation search facilities at various workshops in 2015.

Data Curation Service

Goal of the DCS The Data Curation Service (DCS) was established as a centre of expertise. It has been charged with:

1. the curation of resources, especially those presently held by individual researchers or research groups;
2. assisting in the curation efforts of CLARIN centres (if and when such is desired);
3. advising researchers who wish to undertake the curation of their resources themselves.

The DCS directs its efforts primarily at the curation of language resources created, stored and/or used in the Netherlands. The focus is on resources that are at risk of being lost, especially as there is no other party that can be expected to see to their curation.

The DCS does not host the data itself; it curates the data so that the data can be hosted by a CLARIN Centre in such a way that it is visible (via CMDI metadata), accessible (via a CLARIN centre), uniquely and stably referable (by Persistent Identifiers), and suited for interoperability with tools and other data, both formally (via standardized data formats) and semantically (via links to data category registries).

Experiences in identifying and prioritizing data for curation by the DCS Using information from sources such as the *Nederlandse Onderzoeksdatabank* (NOD; national research database, NARCIS) and the user survey conducted (see section 9, page 38), a list was made of resources that might potentially qualify for curation. Criteria were formulated for establishing whether a resource qualifies for curation by the DCS and for prioritizing resources that do.

As it turns out, collecting reliable information as to the whereabouts and the status of a resource has proven to be a highly time-consuming task, especially with resources that were created in finished projects. In such cases finding researchers who want to be involved in the curation effort, if only by providing information as regards ownership, IPR, and such matters, has only been successful when the researchers themselves stood to benefit from the curation. In a number of cases the resource listed in the national research database could not be located, in other cases IPR problems form an obstacle. The notion of sharing resources with fellow researchers has not been embraced by everybody and some researchers or institutions appear unwilling or reluctant to contribute their data to the CLARIN infrastructure.

Though the relation between the DCS and the CLARIN-NL executive board with respect to the selection of data to be curated has been clearly defined[19], there have been misunderstandings about their respective roles showing up since there have different views as to what data is most desirable to curate: what is perceived of as answering to the user needs or various resources that help attract additional potential users of the infrastructure and show its potential.

Collaboration with the CLARIN Centres The DCS is collaborating with the various candidate CLARIN Centres. The collaboration takes on different forms. Thus, expert knowledge from the centres is brought to bear where PID-attribution is concerned or the structuring of the data collection. The centres store the resources once they have been curated and provide access to them. Each CLARIN Centre targets a specific area and so far deciding on what would be the right centre has been quite straightforward. Unfortunately, the DCS has the experience of encountering a candidate CLARIN Data Centre (DANS) which to date has not adapted to the requirements of such a centre. This delays e.g. the harvesting of metadata files that are created by the DCS in the process of data curation.

Results The following data have been curated:

- LESLLA
- IPNV Interviews with veterans
- DBD/TCULT
- Woordenboek Gelderse Dialecten, Rivierengebied
- Woordenboek Gelderse Dialecten, Veluwe part)
- Curation organisation names for CLAVAS / OpenSkos
- Six dialect dictionaries from Brabant
- WLD and WBD part III
- Roots of Etnolect (collection Linda van Meel)
- Traces of Contact (collection Kofi Yakbo)

[19] The agreement states: “In the DCS-project a list will be made of data to be curated, ranked by priority. For each dataset that is to be curated a detailed plan with estimated effort, lead time, and costs will be made and presented to the executive board for approval” (translated, the original is in Dutch).

During the curation of these resources we made a couple of relevant experiences:

- Researchers sometimes offer incomplete or obsolete databases (not matching the available documentation) and had to resubmit data or new data. After some iteration the set becomes complete and fixed but this takes time.

- Intensive consultation with researchers is needed to understand the nature of the data and the best way to curate them. The DCS often encounters unexpected inconsistencies at all levels of the resource (e.g. file formats, names, locations, directory structure)

- When third parties hear of the curation of a dataset this arouses their interest for the curation of similar data. This leads to interesting possibilities for cooperation. A good example of this is (4), where Nicoline van der Sijs pointed us to her initiative at the Meertens Institute to give dialect dictionaries a second digital life. This helps the DCS to develop metadata profiles and data models that exceed the level of a single resource and meet the needs of a variety of similar resources. This is valued as a great advantage but is accompanied by delays in the curation of the individual resource at hand. Similar cross fertilization took place between the metadata profiles for DBD and LESLLA.

The cooperation on dialect dictionaries led to a new subproject in 2015 (CARE) to continue the work on the curation of regional dictionaries.

Recommendations Based on the experiences with the DCS the following recommendations can be made:

- It would be advisable to have DCS curate a variety of resources that can serve as showcases for future curation so that in future researchers can benefit from examples and guidelines when undertaking the curation of their resources themselves.

- As the DCS experiences on a regular basis, there is an urgent need to address the arrangement of IPR, especially for already existing data collections. Currently no assistance is being offered for settling IPR or for arriving at a suitable arrangement in situations where there is a risk that informants' rights may be trespassed upon while making the data more widely available. As a result data remain with the original researcher and any curation effort remains unsuccessful. One might consider the development of some IPR templates or agreements for likely scenarios, such as for the researcher wants to share his data but not right now, but only after he has had time to publish his research; or: for collections where data have privacy issues, and researchers may use data but when publishing should avoid revealing privacy-sensitive info, etc.

Are they recognized as centres of expertise, both nationally and internationally?

The DCS is recognized as a centre of expertise nationally, and internationally. The DCS has presented about its work on various occasions both nationally [Oostdijk & van den Heuvel 2012, Dutch Society of Phonetic Sciences] and internationally ([Oostdijk & van den Heuvel 2012, LREC2012],[Oostdijk &

van den Heuvel, 2013], [Oostdijk & van den Heuvel 2014], [van den Heuvel et al. 2014]). The DCS also produced a document with guidelines on data curation for CLARIN-NL. This guideline is available here. The work on data curation and metadata curation has been continued in CLARIAH-CORE by the same core team.

11 Dissemination, education, awareness

Is the existence of the technical infrastructure known to the intended users (target: 80% know of its existence)

According to the NARCIS database, there are 2391 humanities researchers in the Netherlands (measured 2013-05-24). We have not yet been able to measure how many of them are aware of CLARIN.

We have of course taken several actions to create awareness of the CLARIN-NL project and the CLARIN infrastructure. These include:

- Setting up and maintaining the CLARIN-NL website
- Being present and visible at national and international conferences and workshops
- Organizing events ourselves
- Presentations
- Publications
- Leaflets
- Meetings with individual researchers or research groups
- Providing support for other events
- Providing support for travel of Dutch researchers to CLARIN-related events
- CLARIN-NL Newsflashes (36 in total) sent out to each registered CLARIN participant
- Messages in CLARIN EU Newsflashes
- Articles in the CLARIN EU Newsletter
- eData & Research: CLARIN-NL has joined the eData & Research Magazine in 2011. CLARIN-NL finds the existence of this magazine important, and aims to make itself more visible by actively participating in its board and editorial team and co-financing it. Several articles on CLARIN-NL have appeared in eData&Research (see the fact book for details). On behalf of CLARIN-NL, Erica Renckens participates in the editorial team and Jan Odijk in the board.
- Movies and Screen Captures: We have made 12 short movies to illustrate the work of certain projects. Some projects have made movies or screen captures themselves to illustrate their work or their demonstrators. See http://www.clarin.nl/node/403 for a list of such videos.
- Announcements of CLARIN events and calls via mailing lists of various research schools (LOT, mediaeval studies, literary studies, Huizinga), newsletters of various organisation (TST-newsletter, Neder-L, Utrecht Humanities newsletter, etc.).

We put some statistics on these actions in the following table

Year	Own Events	News-flashes	Supported Events	Attended Events	Press Articles	Publications	Presentations[20]
2009	2	4	2	11	0	0	23
2010	7	6	4	11	7	11	68
2011	4	5	4	23	10	8	60
2012	5	8	4	27	11	12	42
2013	7	2	2	21	8	21	48
2014	9	3	2	11	17	74	49
2015	6	5	5	14	6	15	45
2016	0	0	0	0	0	17	8

Though we have not been able to measure how well-known CLARIN is, we do have some indications.

First, CLARIN-NL started with 21 organisations as participants, and this has grown during the project to 33 organisations. These include the humanities faculties of all universities, all KNAW humanities institutes, several libraries and data centres, and independent research organisations. See http://www.clarin.nl/node/382 for an overview.

App. 320 persons are registered on the CLARIN-NL website. And over 280 persons have been active in the CLARIN-NL projects in some form.

Have all relevant players been sufficiently informed to be able to participate in designing and constructing the infrastructure?

First of all, all of the activities described under the preceding success criterion are relevant here. Second, we have issued 4 open calls for projects, have explicitly given higher priority to disciplines that were until then underrepresented, and have explicitly invited specific researchers from underrepresented disciplines for submitting project proposals without competition.

All these calls have been announced in a variety of ways (see the instruments under the preceding success criterion). For each call an information session has been announced and held. For each call a public kick-off meeting has been organized. A variety of other CLARIN events has been announced and organized. We have raised awareness for the CLARIN infrastructure and the CLARIN-NL project via the User Survey. We have had meetings with several researchers and research organisations to explore whether and how they can participate in the design and the construction of the CLARIN infrastructure. We have given each demonstration and data curation project the opportunity and in fact the obligation to provide us with information of what the researcher or the researcher's methods, data or tools require of the CLARIN infrastructure. We started with four candidate CLARIN-centres, but during the project the Huygens ING institute joined as a candidate CLARIN-centre and we made available budget for this institute to become a CLARIN –centre.

[20] Data for presentations are underestimates since not all researchers report each of their presentations.

In conclusion, it can be claimed that every humanities researcher and every humanities organisation has had enough opportunities to be informed about CLARIN-NL and to participate in the design and construction of the CLARIN infrastructure.

Has enough training and education on using the infrastructure been given? Are still ample opportunities offered to get such trainings?

We have set up a variety of instruments for training and education, and more generally for knowledge sharing. We list these below. On the basis of all the instruments set up and the activities organised, we submit that there has been ample opportunities for training and education on the infrastructure so far. Education and training will be continued in the CLARIAH-CORE project.

The instruments referred to above include:

- Organisation of a large number of **tutorials and workshops**, in particular on semantic interoperability (e.g. ISOCAT) and CMDI metadata (metadata registry, ARBIL tools, etc.). Most of these were in the Netherlands and aimed at CLARIN-NL participants, but also international tutorials and workshops have been held, both in the Netherlands and abroad (e.g. at LREC 2012, Istanbul)
- Systematic support for semantic interoperability. This will be discussed in more detail below
- **Systematic support for CLAM**. In the context of the TTNWW project, Maarten van Gompel (initially Tilburg University, currently Radboud University Nijmegen) has developed CLAM, software to facilitate turning existing tools into RESTful web services. CLAM is successful and is in use by most participants in the TTNWW project. For this reason, systematic support, maintenance and extension of the CLAM-functionality are required and were offered by CLARIN-NL since 2011. This has later been extended with support for the de facto standard Folia resource format.
- Giving various **courses at seasonal schools**, e.g. by Jan Odijk and Arjan van Hessen (LOT 2011 winter school, Amsterdam), by John Nerbonne on Gabmap at various occasions, by Jan Odijk and guest lecturers at LOT Summerschool 2014 (Nijmegen) and by Jan Odijk and Sjef Barbiers at the LOT Winterschool 2015 (Amsterdam).
- A variety of other **lectures and courses**, e.g. courses on Corpus Analysis, inter alia with the tools developed in CLARIN-NL (and CLARIN more generally) by Jan Odijk in 2013/2014 and 2014/2015, courses in the Utrecht Digital Humanities Minor (by Jan Odijk, Els Stronks, Franciska de Jong, and others), Tutorial of the use of SPARQL for querying the STCN (by Els Stronks), etc.
- Plenary lectures on CLARIN at CLIN25 and TIN 2015
- **THATCamps** Support has been given for the organisation of so-called THATCamps (The Humanities and Technology Camps) in 2014 (The Hague) and 2015 (Utrecht).
- Helpdesk. The **Helpdesk** has been set up in 2010 and has been running smoothly since then.
- **Educational Packages**. Six educational modules were developed and are available via the CLARIN-NL Portal. Several other pieces of educational material are available there, including a recorded lecture on searching for grammatical constructions with GrETEL

A special project (TPC) was also started up to connect the Taalportaal with search and analysis applications developed in CLARIN-NL. This project created links from Taalportaal to the front-end of search and analysis applications operating on grammatical corpora and databases, e.g. to illustrate syntactic constructions (via search in syntactically annotated corpora), morphological, and phonological phenomena (via links to the CELEX database). It thus enriches the Taalportaal, stimulates the use of corpora and lexical databases by linguists in their research and makes linguists gradually acquainted with such corpus search applications.

We discuss support for semantic interoperability in more detail:

Support for Semantic Interoperability (ISOCAT/ CCR)

We have required that each demonstrator and data curation project took measures to ensure interoperability. On the level of formal interoperability, CLARIN-supported formats should be used. On the level of semantic interoperability, use was made of the ISOCAT data category registry as a place where the meaning of data categories used in resources and their metadata could be defined and made available to other researchers. Semantic interoperability is in the short term relevant for metadata (the metadata search engine described in section 6, page 18 makes crucial use of ISOCAT to return search results about for specific meaning of a given input string). We had hoped it would also be relevant in the short term for initial versions of federated content search (FCS), but FCS turned out to be more difficult than expected.

Since working on semantic interoperability and working with ISOCAT was new to almost all researchers, tutorials and workshops on these topics were organized from the start of the demonstrator and resource curation projects. However, since the actual work happens in many different projects, coordination of these activities was needed to assist people, and also to identify overlaps, ensure consistency and avoid duplication of work. For these reasons, systematic co-ordination of ISOcat activities for CLARIN-NL was started in April 2011. The original task was to coordinate work concerning ISOcat from call 2 on, taking care of the content of the inserted data categories (DCs). However, it turned out that first new guidelines were to be developed, to be used in all ISOcat data for CLARIN NL projects. ISOcat in general is very open: everybody can insert data, hampered by very few rules. We thus developed rules for use within CLARIN NL and made some adaptations to the ISOcat registry itself, in order to ensure clear and consistent definitions of the data categories.

The data categories could originally not be found easily, due to the fact that ISOCAT is just a long flat list of data categories and many legacy data categories of dubious quality were included in it. Though, some further organisation of the data categories was introduced to make search easier, it only became easy to work with once a dedicated CLARIN-NL/VL view was set up. ISOCAT makes a distinction between private and public data, and normally private data are only visible to its author. The CLARIN-NL/VL view, however, enables CLARIN-NL/VL users to consult private data in this view. For ISOCAT users outside of CLARIN-NL and CLARIN-VL, part of our data is recognizable as ‘recommended by CLARIN NL/VL’, a new label that was added to ISOCAT.

Public data can be searched using the ISOcat search options. Documentation is available in the form of a glossary.

By the end of 2013, several alternatives to ISOCAT were considered, and in 2014 it was decided to change from ISOCAT to the SKOS-based CLARIN Concept registry (CCR). From that moment on, all educational and training effort were aimed at teaching researchers in the use of CCR.

Has training in the use and actual use of the infrastructure been incorporated in the regular curricula of linguistics and humanities studies, or are there concrete plans to do so? (Target: incorporated in 60% of the curricula after 6 years)

We have made several attempts to get education and training in the use of the CLARIN infrastructure into the regular curricula of linguistics and humanities studies. We have set up a task force to prepare a proposal for this, in which Jan Odijk participated. However, due to various reasons, this task force silently died.

However, independent initiatives to make digital humanities, eHumanities, or computational humanities a component of the regular curricula have been started up at several universities. Though these are independent initiatives, the very existence of CLARIN en CLARIN-NL and its targeted successor project CLARIAH has been influential. In Utrecht, the Humanities lectures were dedicated to digital Humanities in the 2012/2013 academic year, and Jan Odijk and Els Stronks gave lectures there. In addition, a Digital Humanities minor has been approved and was set up as of the academic year 2013/2014. CLARIN was prominently present there with involvement of the board members Martin Everaert and Els Stronks, and EB-member Jan Odijk. Several CLARIN-related lectures were given in the framework of this course, inter alia by Jan Odijk and Franciska de Jong. The course on Corpus Analysis (using the CLARIN infrastructure) was made part of the regular curriculum in Utrecht.

There are also courses in digital humanities at other universities (e.g. at the Free University Amsterdam). In the context of the CLARIAH-SEED project a Digital Humanities Course Registry has been set up for the Netherlands as part of the European registry, which provide a more general overview of Digital Humanities courses.

Six educational packages have been developed that can be used by teachers in their regular courses, covering the GrETEL Treebank search application, the MIMORE search application, the ePistolarium, the Arthurian fiction search application, the Dutch Song Database, and the Emblem data. The materials used for the LOT seasonal school courses on CLARIN have been made available, as well as two articles based on these courses. All this material and more is available via the CLARIN-NL Portal Educational Packages section

12 National Coordination Point

Has a national coordination point been set up and is it functioning successfully?

The National Coordination Point has been set up in 2009 and functioned well since then.

The National Coordination Point has consisted of the programme director and the project secretary. The project secretary function has been fulfilled by three different persons since 2009.

A lot of work is involved in organizing the calls for subprojects, organizing their evaluations, and organizing all administrative work when a subproject starts, is running and ends. Most projects had delays and requested extensions. Some projects were very slow in finalizing, e.g. submitting their deliverables, final reports and financial reports. It should be considered in the future to organize less but bigger calls, and to be stricter on the amount of delays or extensions that will be allowed.

The CLARIN-NL governance structure consists of an Executive Board (EB), Board, National Advisory Panel (NAP) and International Advisory Panel (IAP). Their tasks and responsibilities have been laid down in the consortium agreement. Their composition can be found on the CLARIN-NL website and in the fact book.

Executive Board The executive board has held regular meetings (once every two weeks) to formulate, prepare and implement policy procedures and concrete activities.

Board The Board consists of 8 members. A full list of its members can be found in the fact book. Some members have left the CLARIN Board (including its chair) in 2012 but excellent replacements have been found for them. Prof.dr. John Nerbonne (RUG) is the chair of the CLARIN Board as of September 1, 2012. The board has initially met 2 to 3 times per year, in the later phases of the project once a year.

NAP The National Advisory Panel consists of 16-17 members. At the end of 2011, some members left the NAP, but excellent replacements have been found for them in 2012, providing a good representation of the fields of the humanities, language and speech processing and infrastructure design and implementation. The NAP has met 2 to 3 times per year until 2013, but it has been impossible to get all NAP-members together in one meeting. The NAP met once in 2013, had no meeting in 2014 and held its final meeting in early 2015..

IAP The International Advisory Panel consisted of 7 members since 2010. It has been extended in 2012 with a number of renowned international experts to reduce the work load in evaluating subproject proposals. It currently consists of 10 members. The IAP has held 1 meeting per year from 2010 through 2013, did not meet in 2014 but (in part) attended the CLARIN-NL final event in 2015.

Consortium Agreement The consortium agreement, which was approved by the Board on Nov 2, 2009 and which was initially signed by 21 partners, has remained unchanged. The consortium is open to new partners, and several partners have joined the consortium, bringing the total number of partners to 33 (see http://www.clarin.nl/node/382). The consortium agreement can be found on the CLARIN-NL website: http://www.clarin.nl/node/72.

CLARIN-NL Long Term Working Plan The CLARIN-NL Long Term Working Plan was approved by the Board on August 19, 2009, and has remained unchanged.

Annual Reports, Annual Financial Reports and Working Plans are written every year and approved by the board . They can be found here, where also the Long Term Working plan can be found.

Overall, the governance structure as set up functions appropriately.

Has a business model been developed that guarantees the long term sustainability of the CLARIN infrastructure?

We have worked on developing a business model that guarantees the long term sustainability of the CLARIN infrastructure, but we certainly have not fully solved this problem.

One ingredient of the business model is that the backbone of the CLARIN infrastructure is formed by centres that have the management, maintenance, distribution and long term preservation of data and/or software in their mission independently of the CLARIN-NL project. This means that they already do work and have reserved money in their yearly budgets for such activities independently of the CLARIN-NL project. The CLARIN-NL project enables them to change their way of working so that they can do this work in a CLARIN-compatible manner. After they have made the transition, they still do the same work, for the same money, but now in a CLARIN-compatible manner.

However, it is to be expected that the amount of work for the management, maintenance, distribution and long term preservation of data and/or software will increase after the CLARIN-NL project has finished, since (if CLARIN-NL is a success) all humanities researchers want their data and software made available via a CLARIN centre. The centres have no independent budget for this increase in work.

One model that is being considered is that the research funding agencies (1) require that (language-related) research data and software are stored at CLARIN centres after the research project has finished[21] and (2) require that in each research project a part of the budget is reserved for management, maintenance and distribution of the resulting data and software after the research project finishes, as part of a data management plan. In this way, there will be funding for the CLARIN-centres for the increased work on management, maintenance, distribution and long term preservation of data and/or software, exactly in function of the amount of data and software offered to them. This approach of course requires the full support of the funding agencies, and we are discussing these matters with them.

Apart from financial problems in guaranteeing long term preservation, there are also problems of a more technical and organisational nature. For example, it is technically difficult to maintain software services (given the ever changing operating systems, interfaces, programming conventions, and hardware), and almost impossible for a centre to update or upgrade software developed by others. Maintenance of such software is therefore only feasible if there is an active developer community

[21] They already require this of certain types of projects that yield large amounts of data.

among the researchers that closely cooperates with the CLARIN centre that hosts the service. The lifecycle of software should therefore be made explicit and taken into account, so that work on software without a supporting developer community is reduced (or, if it is essential software for the infrastructure, the centre(s) should build up their own developer group for it).

This problem is not unique to CLARIN, and it makes sense to learn from what others have done in this regard. Several meetings on this topic have been organised by the eScience Centre, and CLARIN-NL has played an active role at these meetings. The CLARIAH-CORE project, which succeeds the CLARIN-NL project, provides us with some more time to come up with solutions for this problem, and it is an explicit part of this project. The matter is dealt with at the European level as well, inter alia in the CLARIN National Coordinators Forum and in the CLARIN ERIC Board.

In summary, we are still actively seeking to develop a business model to guarantee the long-term sustainability of the CLARIN infrastructure and have made considerable progress, but we have to do more work to come up with a concrete and workable approach, and the CLARIAH-CORE project offers us some more time to come up with such a model.

13 Other Activities

As a response to the invitation of the *Commissie voor de update van de Nationale Roadmap voor Grootschalige Onderzoeksfaciliteiten* (Committee for the update of the National Roadmap for Large Scale Research Facilities), CLARIN-NL and the Dutch players in DARIAH joined forces in 2011 and formulated a new proposal for a large scale national infrastructural project called *Common Lab Research Infrastructure for the Arts and the Humanities (CLARIAH)*, uniting all humanities organizations in the Netherlands. This proposal has been submitted and it has received excellent and good-excellent reviews by external experts. The consortium has written a rebuttal to the comments of the external experts, and it has presented and defended the proposal in an interview with the Committee on January 9th, 2012. The project has been put on the Roadmap and has received 1 million euro as 'seed money' from the deputy minister Halbe Zijlstra on a special event for this occasion in Utrecht (March 3, 2012), to maintain the dynamics of the consortium and to prepare a new, improved, proposal. This resulted in the CLARIAH-SEED project, which has finished by now. The major result of this project is a renewed improved project proposal, which was awarded funding on June 1, 2014, and which led to the CLARIAH-CORE project. This project started on Jan 1st, 2015 and will run until Dec 31st, 2018. Details about these projects can be found on the CLARIAH website (www.clariah.nl).

14 Conclusions

The CLARIN-NL project ran from 2009 through 2015 and has set up the Netherlands part of the CLARIN infrastructure. It has formed a network of 5 candidate CLARIN centres that cooperated to create generic infrastructure services, centre-specific infrastructure services such as resource repositories, metadata harvesting, single sign on authentication and authorization, and resolution of persistent identifiers to enable stable access to resources. A metadata registry has been set up and

filled with metadata profiles and components created on the basis of data and software from the Netherlands, editors to make new components and profiles, and editors to make or adapt metadata. The centres also have cooperated to create metadata search facilities and a first prototype of federated content search. All five centres have the Data Seal of Approval (one of the requirements to become a certified CLARIN centre) and 4 of them have become certified CLARIN centres. By their cooperation relevant knowledge and expertise has been built up in multiple centres so that a robustly supported infrastructure emerges. . Its robustness was severely but successfully put to the test when MPI decided to reduce its role in CLARIN. CLARIN-NL was also successful in connecting national data centres such as the National Library (including DBNL), NIBG and Utrecht University Library.

Some 51 data curation and demonstrator projects covering a broad spectrum of humanities disciplines were carried out on the basis of the needs of the targeted users of the CLARIN infrastructure, humanities researchers. In these projects existing data and software were curated to make them CLARIN-compatible, i.e. stored in a CLARIN-supported format, with CLARIN-compatible metadata (CMDI), stably referable by the use of persistent identifiers (PIDs), and with links of its data categories to CLARIN-supported data category registries to ensure semantic interoperability. These data and software, and their metadata, are hosted on CLARIN centres ensuring permanent visibility via the Virtual Language Observatory, the metadata search engine , accessibility via resource-specific and (limited) federated content search, and long term preservation. The curation of these data and software has provided tests of a whole range of infrastructural aspects, such as the proposed CLARIN standards, metadata creation, semantic interoperability via ISOCAT and later CCR, and hardware and software requirements for computationally and/or memory intensive services. A whole range of demonstrators has been created, which can be used not only for demonstration and educational purposes but also for supporting actual humanities research. These include a workflow system that incorporates language and speech technology as web services, which was created in cooperation with CLARIN Flanders.

The user needs have been inventoried in a dedicated user survey and are otherwise ensured by the required active user involvement in data curation and demonstrator projects.

On the organisational level, CLARIN ERIC has been established in February 2012, hosted by Utrecht University. It ensures international coordination and cooperation among the CLARIN partners in the National Coordinators' Forum and in a range of committees on technical and legal aspects. On the national level an excellently functioning governance structure and National Coordination Point have been set up. The national CLARIN activities have been shaped up as a programme rather than as a project and this approach has been very successful in that it offered much opportunity to react to emerging problems, bring in more partners, and to react to ideas and proposals coming from our prospective users. The Netherlands plays a leading role in CLARIN because of the early start and the relatively high budget of its national project, it's hosting of CLARIN ERIC, and the excellent technical expertise it offers to the CLARIN community.

A wide range of knowledge sharing activities have been organized and supported, including national and international tutorials and workshops, seasonal school courses, information sessions, a Helpdesk

including FAQ sections on the website, and support for travel of researchers to relevant events. Several educational packages were developed and are, together with other educational material, available via the CLARIN-NL Portal. Several courses in digital humanities in which the CLARIN infrastructure is prominently present have become part of the regular humanities curriculum of some universities.

Dissemination and PR has been implemented through the CLARIN-NL website, newsletters and newsflashes, active participation in the *eData & Research* magazine, messages via humanities mailing lists, presentations, publications, meetings with individual researchers or research groups, leaflets, and a range of movies and screen captures. Through all these activities, CLARIN-NL is well known among humanities researchers in the Netherlands. A book on CLARIN in the Low Countries with over 26 contributions is being prepared and is expected to appear by the end of 2016.

Two centres of expertise have been set up, one for data curation and one for data and tools for the longitudinal study of language and culture.

CLARIN-NL makes its data and services visible and available via a portal through which all services are visible and can be accessed.

It can be stated with confidence that the "general ambition of the CLARIN-NL project to be a major contributor both organisationally and technically to the design, specification, construction and exploitation of a European-wide CLARIN infrastructure" has been achieved. The Netherlands part of the infrastructure is "actually used by its intended users (humanities scholars)". Research that crucially used the CLARIN infrastructure is now appearing in the *Lingua* Special Issue on the CLARIN Infrastructure with Jan Odijk as guest editor. We are approached by many researchers when they are preparing new project proposals (and some of them were awarded funding). CLARIN is also involved in a project proposal for the NWO *Gravitation* programme that is being prepared by a large consortium of linguists united in LOT. The first PhD in which crucial use is made of CLARIN has been successfully defended by Liesbeth Augustinus in Leuven. The CLARIN work for this was made possible in part through the close collaboration of Flanders with CLARIN-NL.

The major objectives of the project have been achieved, though a range of problems remain and will have to be addressed in successor projects. Fortunately, CLARIN-NL and DARIAH-NL have joined forces and managed to secure funding for the infrastructure projects CLARIAH-SEED (2012-2014) and CLARIAH-CORE (2015-2018). In addition, there has been close cooperation with NWO-Groot projects such as Nederlab and Taalportaal.

Making use of the CLARIN infrastructure by its intended users (humanities scholars) is becoming "a normal 'modus operandi' for them as reflected by the incorporation of training and education in the regular curricula of humanities studies".