



CLARIN-NL

Long Term Programme

2009-2014

Author: *Jan Odijk*

Under the supervision of the CLARIN-NL preparatory committee

- Prof.dr. Hans Bennis, Meertens Institute, Amsterdam
- Prof.dr. Lou Boves, Radboud University, Nijmegen
- Prof.dr. Martin Everaert, Utrecht University
- Prof.dr. Franciska de Jong, Twente University
- Drs. Steven Krauwer, Utrecht University
- Prof.dr.ir. John Nerbonne, Groningen University
- Ir. Peter Wittenburg, Max-Planck Institute, Nijmegen

April 7, 2009.





Summary

CLARIN-NL aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent *eScience* working environment for researchers in the Humanities, and Linguistics in particular, who want to make use of language resources and the technology to use these resources for their research. This infrastructure will provide these researchers with a wide variety of resources and services, intelligent access methods for exploring the resources and innovative ways of combining different resources into virtual collections, so that information hidden in unstructured textual and multimedia documents can be disclosed. Inter-operability of independently developed resources and services will be key for a properly functioning infrastructure. The infrastructure will be easy to use for non-technical researchers. Targeted dissemination activities, educational programmes and training sessions will enable a whole generation of researchers and students to acquaint themselves with this new research methodology and the potential for groundbreaking research it offers, creating an advanced scientific environment in the Netherlands that will attract top-researchers and students from abroad.

CLARIN-NL forms the Netherlands national counterpart of the CLARIN enterprise on the European level (CLARIN-EU). It therefore resembles and complements the preparatory project that is currently being executed on the European level (CLARIN-prep). Many of the activities and sub-projects within CLARIN-NL implement activities in the Netherlands that in the programme of work for CLARIN-prep are envisaged to take place in every participating country and that will be funded through the national contributions to CLARIN. Such activities include (1) the design and implementation of the infrastructure technology; (2) application projects in which technology providers and the intended users integrate local repositories and set up local services for prototypical test installations as initial demonstrators, enabling evidence-based contributions to the discussion on standards and best practices for inter-operability, and to contribute to the survey of requirements for the infrastructure technology; (3) the preparation of an essential data collection and service set for the locally relevant languages (ideally on the basis of existing tools and data) that allows for testing and validation of proposed standards, services and tools in the experimental prototype; and (4) the integration of advanced infrastructure services. Since it is not possible to assign all these tasks to participants right from the start, CLARIN-NL has been set up as a mixture between a programme and a project.

CLARIN-NL, however, also contains a range of activities that aim to further strengthen the leading position the Netherlands currently has in CLARIN-EU (both the principal coordinator and the technical coordinator for infrastructure technology are based in the Netherlands). It has a separate line of activities aimed to position the Netherlands prominently in CLARIN-EU also beyond CLARIN-prep, and to extend its leading position further by initiating, in an early stage, projects with selected international partners to develop, in a multilingual setting, showcase demonstrators of the infrastructure and the services it offers, as well as by setting up at least two centres of expertise



The CLARIN-NL proposal covers a period of 6 years, partitioned in three phases of two years: the preparation phase, the construction phase, and the first two years of the exploitation phase. Though the infrastructure is primarily aimed at language and humanities researchers, it offers various opportunities for usage in other domains and by other users, both for commercial applications as well as for important developments in society.



Table of Contents

| | |
|--|----|
| Summary | 3 |
| Table of Contents | 5 |
| 1 Introduction | 6 |
| 2 CLARIN-NL Goals | 6 |
| 3 Background | 6 |
| 4 Actions to be undertaken in CLARIN-NL | 8 |
| 4.1 EU-Line | 8 |
| 4.2 NL-Line | 9 |
| 4.3 Phasing | 10 |
| 5 Finances | 11 |
| 6 Governance | 11 |
| 7 Criteria for evaluation of project proposals and task descriptions | 14 |
| 8 Network Creation, Knowledge Transfer, and Embedding | 15 |
| 9 Success Criteria | 17 |
| 10 References | 20 |
| Appendix A Budget | 21 |
| Appendix B Elaboration of the technical infrastructure related goals and tasks | 22 |
| Appendix C CLARIN-NL Initial Partners | 26 |



1 Introduction

This is the CLARIN-NL long term programme. It covers the whole period of the CLARIN-NL project (2009-2014). It describes

- The *CLARIN-NL Goals* (section 2)
- the *Background* of CLARIN-NL (section 3)
- the range of *Actions to be undertaken in CLARIN-NL*, both for the *EU-Line* and for the *NL-L*, including their *Phasing* (section 4)
- the *CLARIN-NL Finances*, more specifically the budget assigned to the various activities (section 5 and Appendix A)
- the *CLARIN-NL Governance* structure (section 6)
- *Criteria for evaluation of project proposals and task descriptions* and their assignment to research and development groups (section 7)
- *Network Creation, Knowledge Transfer, and Embedding* of CLARIN-NL in the Netherlands and cooperation with other relevant projects and programmes (section 8)
- Success Criteria (section 9)
- The CLARIN-NL Initial Partners (Appendix C)

2 CLARIN-NL Goals

CLARIN-NL aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent *eScience* working environment for researchers in the Humanities, and Linguistics in particular, who want to make use of language resources and the technology to use these resources for their research. This infrastructure will provide these researchers with a wide variety of resources and services, intelligent access methods for exploring the resources and innovative ways of combining different resources into virtual collections, so that information hidden in unstructured textual and multimedia documents can be disclosed. Inter-operability of independently developed resources and services will be key for a properly functioning infrastructure. The infrastructure will be easy to use for non-technical researchers. Targeted dissemination activities, educational programmes and training sessions will enable a whole generation of researchers and students to acquaint themselves with this new research methodology and the potential for groundbreaking research it offers, creating an advanced scientific environment in the Netherlands that will attract top-researchers and students from abroad.

3 Background

Following a request from the European Commission, the European Strategy Forum for Research Infrastructures (ESFRI) decided early 2005 to compile a list of opportunities in order to assist the Commission in the preparation of its proposal for the Seventh Framework programme (FP7).



Following a mandate of the Council of Ministers, ESFRI, in its work towards a European roadmap, set up, during summer 2005, different Roadmap Working Groups (RWGs) to analyse topical issues, including the *Social Sciences and Humanities Roadmap Working Group*.

This ESFRI Social Sciences and Humanities Roadmap Working Group decided to recommend a number of projects, among them CLARIN, for consideration for the ESFRI Roadmap, on May 15, 2006. This decision has been laid down in the report of this working group, published in summer 2006 [Henrichsen 2006],

On reception of the RWG reports, ESFRI started drafting the roadmap in March 2006. The Roadmap document was checked by a 'Review Group' and finally approved by ESFRI during its meeting held on 28-29 September 2006 [Wood 2006].

The projects on the roadmap could participate in the first call 2006/2007 under FP7 Capacities Specific Programme of the European Commission, to implement a *preparatory phase*. This preparatory phase aims at bringing the projects to the level of legal and financial maturity required to implement the project.

In this call the CLARIN preparatory project (CLARIN-prep) was awarded funding. Its major non-technical aim is to prepare a ready-to-sign agreement between the participating countries whereby they commit themselves to the joint construction and exploitation of the CLARIN Infrastructure.

In the Netherlands, a national roadmap commission for research infrastructures ('Commissie Nationale Roadmap Grootchalige Onderzoeksfaciliteiten', also known as the 'Commissie Van Velzen') was set up. It identified, in December 2007, eight projects that the Netherlands should support financially [Van Velzen 2008]. CLARIN was among them.

The CLARIN representative of the Netherlands (UIL OTS, Utrecht University) was requested, in June 2008, to work out the proposal for a project for the Netherlands contribution to the overall CLARIN infrastructure. This resulted in a proposal for the project CLARIN-NL, submitted in September 2008 [Odiijk 2008].

This proposal was evaluated by an NWO-SenterNovem advisory committee ('Commissie Van Duinen'). The committee recommended to award funding to five projects, among them CLARIN-NL. However, the funding awarded to CLARIN-NL was less than requested (9 million euro instead of 22.5 million euro) and the committee recommended to set the focus of CLARIN-NL on linguistics and the humanities outside of linguistics, and to postpone work for sound, vision, and the social sciences. [Van Duinen 2008].

The minister of Education, Culture and Science ('OCW') decided on November 28, 2008, to adopt the recommendations of the 'Commissie Van Duinen', and asked NWO to finance the selected projects.

4 Actions to be undertaken in CLARIN-NL

CLARIN-NL will undertake a range of actions, which to a large extent reflect and complement the actions taken in the CLARIN-EU work packages and working groups. However, given the advice of the evaluation committee and the budget assigned, CLARIN-NL will aim primarily at *linguistics* (broadly construed as the study of the structure of language and speech in different modalities, historical dimensions, social and cultural settings) and *the humanities outside of linguistics*, but pay no or less attention to the social sciences. In addition, activities in the area of sound and image will be limited and focused on their use in the study of language. The CLARIN-NL actions can be divided in two major classes: one set of activities ('European Line') is focused on Europe and aims to strengthen the position of the Netherlands in CLARIN-EU. A second set of activities focuses on the Netherlands itself ('NL line').

CLARIN-NL will take the lead in the start-up and the execution of the construction phase of CLARIN-EU, following the preparatory phase ending on January 1 2011. To this end, we will, in the course of 2010, set up and gradually expand the European CLARIN Headquarters. The implementation of the governance and support structures for CLARIN-EU will be set up following the recommendations from the preparatory phase.

As there is no funding available for these activities in the preparatory phase budget of CLARIN-EU these costs will have to be covered by the CLARIN-NL budget. This is reflected in the EU-line budget for the first two years.

4.1 EU-Line

Throughout the construction and exploitation phase the central management, support and coordination facilities for CLARIN-EU will be the responsibility of the Dutch CLARIN team. This is reflected in the EU-line of the budget for the years 2011-2014.

The aims of the European line can be summarized as follows:

- consolidate NL's leading position in CLARIN-EU beyond the preparatory phase
- ensure seamless transition from the preparatory phase to the construction phase
- make sure that the Netherlands becomes a main hub in the European infrastructure
- position CLARIN-NL also outside of Europe

From these aims a number of actions can be derived which are enumerated here:

Actions:

1. implement and host the governance structure as recommended by the preparation phase (CLARIN-prep)
2. set up and host a main CLARIN Office for administrative and logistic support for the governance bodies
3. set up and host a main European CLARIN Technical Centre to build and maintain the technical infrastructure



4. set up and host the central CLARIN Coordination point for
 - i. development and maintenance of standards
 - ii. harmonization of IPR issues
 - iii. education, dissemination and promotion
5. set up a local but international example infrastructure with one or two other leading partners, to be ahead of the others (see below); candidate partner are: Germany (MPG/MPI-link) and Flanders & South Africa (via the Dutch Language Union); the partners should include also users that are not LRT providers so that it will be possible to fully test and demonstrate this example infrastructure
6. maintain close connections with other relevant players (EU and non-EU)

4.2 NL-Line

The project plan for the CLARIN-EU preparatory phase contains a number of activities that are envisaged to take place in every participating country and that should be funded from the national contributions to CLARIN. This will consist of contributions to project work packages and execution of work packages defined at the national level (see p. 47 of the CLARIN-EU Description of Work). From this a number of aims for CLARIN-NL can be derived, but CLARIN-NL goes further since it wants to establish itself a leading group in the CLARIN enterprise.

The aims of the NL line can be summarized as follows:

- make sure that the CLARIN-prep specifications match the requirements of the national research community by
 - making an inventory and analyzing the needs of the national research community
 - broad participation in the definition of standards
 - conducting validation projects to establish the requirements for the infrastructure
- build and exploit the national part of the CLARIN-EU federation as a best practice example for the other CLARIN-EU partners
- act as a world class service centre in at least two specific application areas to be defined by the needs and ambitions of the Dutch research community
- set up a CLARIN-NL national coordination point including a support office

From these aims, a number of actions can be derived, which are also listed here:

Actions:

1. technical prototype infrastructure as specified by CLARIN-prep
 - a. building the grid-based structure
 - b. providing the generic services
 - c. operating and validating the prototype
2. data infrastructure as determined by CLARIN-prep and national priorities
 - a. surveying existing data and specifying the essential data set
 - b. arranging for NL-specific IPR issues, especially related to existing data with IPR restrictions



- c. agreeing on representation standards within CLARIN-prep
- d. constructing the essential set by means of conversion or (if needed) new digitization actions, and in parallel developing tools to facilitate this process for other resources of the same type.
- e. validating the prototype on the basis of concrete usage cases
3. language technology service infrastructure as determined by CLARIN-prep and national priorities
 - a. specifying a dynamically growing set of essential tools and services
 - b. agreeing on inter-operability standards within CLARIN-prep
 - c. constructing the essential set by means of encapsulation or (if needed) porting or building
4. establishing user needs in coordination with CLARIN-prep
 - a. surveying current practice in the Netherlands
 - b. carrying out pilots and demonstrators with HSS researchers, preferably in an international setting
 - c. establishing governance procedures for eliciting, prioritising and selecting extensions and improvements
5. creation and operation of two centres of expertise
6. creation and operation of dissemination, education and awareness facilities
7. setting up and operating national coordination point and develop a business model that guarantees the long term sustainability of the CLARIN infrastructure

We elaborate on some of the technical infrastructure in *Elaboration of the technical infrastructure related goals and tasks*

4.3 Phasing

As to the phasing of the project, we foresee three major phases over the period of the coming six years, each roughly taking up two years.

Phase 1: Preparation (2009-2010). This phase overlaps with the CLARIN-EU project so that the Dutch research groups and data centres have the opportunity to provide their requirements for the infrastructure and will have an influential role in specifying the infrastructure design. These requirements will be derived from actual usage cases reflecting the needs of HSS researchers and from technological requirements imposed by the existing data and technology underlying the envisaged services. Adherence to the user centred design principle implies that already in this phase we will implement service prototypes. Also, activities will start for developing a sustainable business model for access to digital information.

Phase 2: Construction & initial operation (2011-2012). In this phase the infrastructure will be constructed and subjected to extensive testing and evaluation by user groups. An initial version will result in which a growing number of eScience methods can already be applied. A strategy will be developed for making CLARIN-NL sustainable after the end of the project.

Phase 3: Operation & further development (2013-2014). In this phase the focus will be on exploitation of the infrastructure, though further developments will take place as well. Extending the infrastructure with data, tools and services will be a continuous activity, and it is expected to increasingly become a routine task, supported by



conversion and metadata enrichment tools developed in earlier phases. As soon as the infrastructure is in actual use there will be requests for modified or extended functionality. These will be surveyed and prioritized, and a selected subset will be implemented, next to normal maintenance activities that are required for the proper operation of the infrastructure.

5 Finances

The budget for the proposed CLARIN-NL programme is specified in some detail in Appendix A. Funds have been allocated to various actions over the three phases, viz. the preparation phase (2009-2010), the construction & initial operation phase (2011-2012) and the operation and further development phase (2013-2014). The budget over the six-year period for the EU-line is 1.35M€ for the NL-line 7.65M€ and 9M€ in total.

The budget is distributed over the years as follows:

- 2009: 1.35M€
- 2010: 1.35M€
- 2011: 1.90M€
- 2012: 1.90M€
- 2013: 1.25M€
- 2014: 1.25M€

The budget will be reviewed yearly and adapted if new developments internal to CLARIN-NL or external (e.g. developments in CLARIN-prep) would make this necessary or desirable.

6 Governance

CLARIN-NL has a mixed character: partially it is a project that requires efficient assignment and execution of a variety of tasks; partially it is more like a programme in which projects can be submitted to open calls and calls for tender.

The organizational structure will have to take a number of central objectives of the program into account simultaneously:

- a) to bring together expertise on a range of sub-disciplines in order to address the core issues involved in setting up a research infrastructure for specific strategic research questions;
- b) to assure the coherence of the overall activities program in terms of the conceptual issues at the core of the program;
- c) to develop and apply new data sources;
- d) to assure high quality both of the overall program and of the individual projects;
- e) to keep to the schedule proposed in the planning;
- f) to stimulate communication among the researchers working in the programme



The CLARIN-NL organizational structure consists of the following bodies:

- **Board**
 - Consists of senior researchers and other experts with expertise in the field and in governance
 - Max. 10 members
 - The Board elects a chair from its members
 - Typically meets twice a year
 - takes strategic decisions
 - is responsible to Funding Agencies
 - determines the Annual Report
 - determines the Financial report
 - determines the Long Term Work Plan
 - determines the Yearly Work plan
 - decides on appointing members in the governance bodies
 - monitors progress of the project as a whole
 - acts as mediator in case of conflicts
 - assigns, if needed, new tasks to governance bodies
 - determines the procedure for avoiding conflict of interest
 - decides on admission of new participants
 - decides on all matters not arranged specifically in the consortium agreement
 - Programme Director and CLARIN European coordinator attend the Board meetings as informant/observer
- **Executive Board (EB)**
 - 4 members
 - Frequent meetings (once every two weeks) and other contacts
 - The executive board reports to the Board
 - A special role is played in the EB by the programme director (Prof.dr. J. Odijk), who has the full and final responsibility for the CLARIN-NL project reaching its goals and the effective functioning of its infrastructure. He will be in charge of the daily coordination of activities, implementation of evaluation procedures, finances, and personnel. The programme director will have regular contacts with Board, NAP and IAP.
 - The other members have specific subtasks, e.g. technical coordination, dissemination, education, awareness, IPR, etc., as their assignment.
 - The EB elects a chair from its members.
 - The EB prepares the long term work plan including success criteria for evaluating the project
 - The EB translates the long term programme into yearly work plans
 - The EB monitors the progress of assigned sub projects
 - The EB organizes evaluations of the project
 - The EB develops a procedure for avoiding conflicts of interest
- **National Advisory Panel (NAP)**
 - 15-20 representatives from Linguistics and Humanities
 - The representation of intended users of the infrastructure should be dominant



- The NAP elects a chair from its members.
- Typically meets 5x per year
- Advises on a yearly work plan specifying steps to be taken with respect to the technical (prototype) infrastructure, the data infrastructure, the language technology service infrastructure, establishing user needs, creation and operation of centres of expertise; creation and operation of dissemination, education and awareness facilities
- Advises the EB on developments, trends and desiderata in the national research field
- Evaluates and ranks subproject proposals from calls for project proposals especially taken into account the research environment in the Netherlands.
- Evaluates finished subprojects and their results.
- Monitors progress of the yearly work plan
- **International Advisory Panel (IAP)**
 - Appr. 8 members
 - Typically meets once a year
 - The IAP is a body of prominent and experienced researchers outside the Netherlands who are actively involved in the creation or maintenance of research infrastructures in e-Linguistics/e-Humanities, more specifically experts in CLARIN, experts in linguistics, experts in the humanities minus linguistics, experts in technical infrastructures and experts in IPR matters.
 - Advises the executive board and the board in international matters, e.g. cooperation, coordination and harmonisations with other international initiatives and programmes
 - Evaluates and ranks subproject proposals on request of the board resulting in an advice to the board

The initial composition of these bodies is determined by the CLARIN-NL preparatory committee.

All members of these bodies are member 'in person', i.e. they do not act as representatives of their organizations, cannot be replaced by colleagues, and will stay member if they would leave their organization.

A list of candidates for these bodies has been made and will soon be finalized.

The executive board will develop a procedure to ensure that conflicts of interests are avoided.

Developments in the CLARIN-prep proposal at the European level may require a revised governance structure after the preparatory phase. The board decides on revisions of the governance structure.

CLARIN-NL has a number of participating (research) organizations as its institutional members. Only participating organizations are eligible for funding of activities or subprojects. The list of initial CLARIN-NL participating research organizations has

been provided in Appendix C. CLARIN-NL is open for new institutional participants. The Board decides on admission of new participants.

7 Criteria for evaluation of project proposals and task descriptions

CLARIN-NL will use several instruments to achieve its goals. It will use direct assignments of tasks to specific organizations or persons, but also use open calls for projects and calls for tender for specific targets to have its goals realized.

General criteria to evaluate project proposals and task descriptions are the following:

- **Quality**
 - Clarity and originality of the project proposal, in particular of the problem and the proposed approach
 - Suitability of the method and plan for the problem at hand
 - Feasibility of the project targets: can they be realized within the specified amount of time and with the instruments proposed?
 - Adequate balance between requested instruments and funds and proposed targets
 - Clearly specified and realistic work plan
 - Conformance to established standards and protocols as supported within CLARIN, or contribute to the development such standards and protocols.
- **Project Participants**
 - Competence of the participating partners (including their past performance);
 - Balanced cooperation and task assignments within the project. Justification of the composition of the team.
 - Availability of the infrastructure required for the project to be successful
 - Embedding of the work in other research programmes or projects, and/or additional funding from other funding sources is an advantage
- **User-orientation of the project**
 - Does the project address needs of the targeted infrastructure users (linguists and humanities researchers)?
 - Is there cooperation with or support from the targeted (future) infrastructure users?
 - Is the resulting tool / service user-friendly, i.e. will non-technical linguistic and humanities researchers be able to use the tool /service?
 - Is dissemination of the results to the targeted users and (where appropriate) training of them planned?
- **Contribution to CLARIN-NL as a whole**
 - Conformance to the goals of CLARIN-NL in particular and CLARIN in general and the priorities set within them



- Contribution to knowledge transfer and network creation. In particular, cooperation between the intended users (linguists and humanities researchers) and technology and service providers (researchers in language and speech technology, computer science, etc.) is an advantage.
- **Intellectual Property Rights and Synergy**
 - Each proposal must contain clear statements about the situation of the IPR of the data and tools/technologies used, and a detailed plan to resolve any open issues.
 - The project participants have the obligation and must therefore have the rights to incorporate the core data and tools used in a project into the CLARIN infrastructure (this is a sine qua non). There has to be a clear specification and justification of the use of any data or tools needed in the project that cannot be incorporated into the CLARIN infrastructure.
 - Each proposal must show that the submitters have adequate and up-to-date knowledge of data, tools and services that are already available, so that any duplication of effort can be avoided.
- **Formal compliance**
 - A proposal must meet the formal requirements imposed by the CLARIN-NL organization for proposals, such as
 - conformance to the prescribed format and proposal template
 - submission before the set deadline, using the means prescribed
 - conformance to the prescribed language of the proposal
 - etc.

Additional criteria, or more specific variants of the criteria listed above can be imposed by the Board for specific calls or actions depending on the specific circumstances, requirements or targets.

8 Network Creation, Knowledge Transfer, and Embedding

An important goal of CLARIN-NL, apart from the realization of the technical infrastructure, is strengthening the cooperation between the targeted users and the technology and services providers. These groups currently operate in largely different circles, use different scientific terminology, and sometimes the same terms but with a different meaning. The technology and services providers do not know exactly what the intended users need, and the intended users do not know exactly what the technology and service providers can offer. Therefore, network creation, information exchange, knowledge transfer, not only at the national level, but also at the international level is crucial for the success of CLARIN.

A range of potential activities and events in this area is planned for in CLARIN-NL. Here we provide an initial list of types of activities and some concrete examples. Each



year at the beginning of the year, a more detailed plan will be set up for the specific events to be organized that year.

- **Network Creation**
 - Organize a yearly CLARIN-NL day (national)
 - Organize a yearly CLARIN day (international)
 - Organize CLARIN brokerages (to stimulate cooperation between the targeted users and technology and services providers)
- **Knowledge Transfer**
 - Publication of articles in (scientific) journals
 - User meetings (to exchange knowledge, experiences)
 - Participation in conferences and workshops
 - Temporary relocation of researchers from the user group to a technology/service provider group, or vice versa.
 - Invite foreign top researchers / experts for a stay in the Netherlands
 - Promote CLARIN by making attractive demonstrators available and clearly visible
 - Create CLARIN-NL website
 - Set up and distribute newsletters and/or news flashes
 - Create a collaborative web-based environment for a CLARIN forum (e.g. Wiki)
 - Organize training sessions for users of the infrastructure, its (future) tools and services and its (future) data
 - Stimulate incorporation of training in the use of CLARIN in the regular linguistics and humanities educational programmes.
 - Description of the CLARIN-NL project and infrastructure in Wikipedia and similar electronic knowledge sources
- **Embedding**
 - Regular consultation with other related programmes and projects, such as CLARIN-prep, other national CLARIN projects (as currently already running in Flanders, Denmark, Germany and other countries), STEVIN, Alfalab, FIDLR-Start, FlaReNet, NWO Dyslexie, etc. (see the next paragraph for more examples).
 - Regular consultations with Flanders on aspects that involve the Dutch language, in order create synergy and avoid duplication of effort.

Concerning embedding and cooperation, CLARIN is widely supported in the Netherlands. In addition, CLARIN-NL will be embedded in the European-wide CLARIN network. Furthermore, the main driving forces of CLARIN-NL, Utrecht University and MPI are also the main forces behind CLARIN-EU. CLARIN-EU also aims at wider international cooperation. There are close relationships (and even overlapping participation) with other international networks such as FlareNet and comparable initiatives in the US (*Interop-SILT*), Japan, Korea, South America, Australia and South Africa. Several related research programmes that currently are running or have just finished strengthen the environment in which CLARIN-NL will run, and emphasize the need for the CLARIN infrastructure. Examples are IMIX (just finished), STEVIN, NWO Dyslexie <http://www.nwo.nl/dyslexie>, the ERC Advanced Research are Grant awarded to Pieter Muysken (RU), DoBeS, the Multilingualism



Project (MPI & RU), and the Sign Language Project (RU et al). CLARIN-NL will actively seek cooperation with these and similar future projects.

There are also some programmes running or in preparation that relate directly to CLARIN since they deal with aspects of infrastructures. Examples are Alfalab (starting up) and FIDLR-Start, which aims to prepare a proposal for “NWO-Groot” in 2009. CLARIN-NL has already established close relationships with these projects and the people involved in them, and will intensify this even further.

Looking at the Dutch language area, institutes in Flanders seek cooperation with institutes in the Netherlands on infrastructure-related matters. This is natural for aspects related to the Dutch language and in line with the recent (April 17, 2008) ministerial statement of intention on further strengthening the collaboration between the Netherlands and Flanders in economics, science, and innovation. The inclusion of INL, working under the governance of the intergovernmental Dutch Language Union, already guarantees to some extent that synergy is created and duplication of effort is avoided. We also have close contacts with the CLARIN coordinator in Flanders and have already discussed with the coordinator for CLARIN in Flanders what nature the cooperation could take. Minutes of this meeting can be found on the CLARIN-NL website. These contacts will be maintained and revised if the situation with regard to CLARIN in Flanders/Belgium would give occasion to do so.

9 Success Criteria

This section lists the success criteria for the CLARIN-NL project. In general, the criteria should be applied at the end or after the project, though occasional criteria can be evaluated earlier. The executive board will formulate specific targets and instruments to achieve these targets for each individual year in the project in yearly work plans.

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 and where its use has become a normal mode of operation for them as reflected by the incorporation of training and education in the curricula of linguistics and humanities studies.

This general ambition can be translated into more specific success criteria that are linked to the actions defined in 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)
- 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?



10 References

- [**Henrichsen 2006**] Bjørn Henrichsen (ed.), *Report of the Social Sciences and Humanities Roadmap Working Group*, version September 4, 2006.
ftp://ftp.cordis.europa.eu/pub/esfri/docs/ssh-rwg-roadmap-report-2006_en.pdf
- [**Odiijk 2008**] Jan Odiijk (ed.). *CLARIN-NL Project Proposal*. September 1, 2008.
http://www-sk.let.uu.nl/pc/clarin_nl/clarin-nl.pdf
- [**Plasterk 2008**] *Investeren in grote onderzoeksfaciliteiten*, November 28, 2008.
<http://www.minocw.nl/actueel/nieuws/35730/investeringingroteonderzoeksfaciliteiten.html>
- [**Van Duinen 2008**] R.J. Van Duinen et al., *Advisory report of NWO and SenterNovem to the Minister of Education, Culture and Science concerning the eight ESFRI proposals from the National Roadmap*, November 12, 2008.
http://www.minocw.nl/documenten/NWO_ESFRI_compleet.pdf
- [**Van Velzen 2008**] Commissie Nationale Roadmap Grootchalige Onderzoeksfaciliteiten, *Nederlandse Roadmap Grootchalige Onderzoeksfaciliteiten*, Amsterdam, oktober 2008, ISBN 978-90-6984-575-3.
<http://www.minocw.nl/documenten/Nederlandse%20Roadmap.pdf>
- [**Wood, 2006**] John Wood (ed.) *European Roadmap for Research Infrastructures: Report 2006*. Luxembourg: Office for Official Publications of the European Communities, 2006. ISBN 92-79-02694-1
ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap-report-26092006_en.pdf



Appendix A Budget

| CLARIN-NL 2009-2014 all amounts in M€ | preparation 2009_2010 | | | construction 2011_2012 | | | exploitation 2013_2014 | | | all periods 2009_2014 | | |
|--|--------------------------|--------------|--------------|---------------------------|--------------|--------------|---------------------------|--------------|--------------|--------------------------|--------------|--------------|
| | <i>labour</i> | <i>other</i> | <i>total</i> | <i>labour</i> | <i>other</i> | <i>total</i> | <i>labour</i> | <i>other</i> | <i>total</i> | <i>labour</i> | <i>other</i> | <i>total</i> |
| EU-level | | | | | | | | | | | | |
| Management and coordination | 0.02 | 0.03 | 0.05 | 0.11 | 0.05 | 0.16 | 0.24 | 0.06 | 0.30 | 0.37 | 0.14 | 0.51 |
| Technical coordination | 0.01 | 0.02 | 0.03 | 0.05 | 0.03 | 0.08 | 0.12 | 0.04 | 0.16 | 0.17 | 0.09 | 0.26 |
| Linguistic coordination | 0.01 | 0.01 | 0.01 | 0.05 | 0.03 | 0.08 | 0.12 | 0.04 | 0.16 | 0.17 | 0.08 | 0.25 |
| Outreach coordination | 0.01 | 0.01 | 0.02 | 0.05 | 0.03 | 0.08 | 0.12 | 0.04 | 0.16 | 0.17 | 0.08 | 0.26 |
| Internationalisation | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.04 | 0.08 |
| Total EU level in M€ | 0.04 | 0.08 | 0.12 | 0.26 | 0.15 | 0.41 | 0.62 | 0.20 | 0.82 | 0.93 | 0.43 | 1.35 |
| | | | | | | | | | | | | |
| NL-level | | | | | | | | | | | | |
| Technical construction | 0.30 | 0.05 | 0.35 | 0.48 | 0.07 | 0.55 | 0.19 | 0.02 | 0.21 | 0.97 | 0.14 | 1.11 |
| Data conversion | 0.48 | 0.06 | 0.54 | 0.54 | 0.10 | 0.64 | 0.29 | 0.05 | 0.34 | 1.31 | 0.21 | 1.52 |
| Tools and services | 0.54 | 0.07 | 0.61 | 0.60 | 0.08 | 0.68 | 0.17 | 0.04 | 0.21 | 1.31 | 0.19 | 1.50 |
| User needs and usage cases | 0.43 | 0.06 | 0.50 | 0.59 | 0.09 | 0.68 | 0.18 | 0.02 | 0.20 | 1.20 | 0.17 | 1.37 |
| Advanced LT services | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Expertise centres | 0.06 | 0.01 | 0.07 | 0.20 | 0.06 | 0.27 | 0.19 | 0.07 | 0.26 | 0.46 | 0.15 | 0.61 |
| Dissemination and training | 0.12 | 0.08 | 0.20 | 0.19 | 0.09 | 0.28 | 0.12 | 0.09 | 0.21 | 0.43 | 0.26 | 0.69 |
| Coordination & management | 0.23 | 0.08 | 0.30 | 0.22 | 0.07 | 0.29 | 0.20 | 0.06 | 0.26 | 0.65 | 0.20 | 0.85 |
| Total NL level in M€ | 2.16 | 0.41 | 2.57 | 2.82 | 0.57 | 3.39 | 1.34 | 0.34 | 1.69 | 6.32 | 1.33 | 7.65 |
| | | | | | | | | | | | | |
| Grand total in M€ | 2.20 | 0.49 | 2.69 | 3.08 | 0.72 | 3.80 | 1.97 | 0.54 | 2.51 | 7.25 | 1.75 | 9.00 |



Appendix B Elaboration of the technical infrastructure related goals and tasks

Technical Infrastructure

CLARIN-NL intends to create the technical infrastructure that is needed to provide a sustainable and persistent *eScience* working environment for researchers in the Humanities that want to make use of language resources and technology. This infrastructure will take the form of a Resource Provider Federation that will rely on a Service Oriented Architecture (SOA) and also Grid services (where necessary). The focus on SOA should be appealing, because this approach will avoid costly failures with monolithic ICT projects.

CLARIN-NL will set up and test three (interrelated) kinds of technologies, viz. Federation Technologies, Registry Services and Information Services. Where necessary it will develop missing components to establish an integrated infrastructure.

Federation Technologies will be realized and tested in close collaboration with the leading experts from TERENA, SURFnet, SARA and other agencies. The federation will start with a set of selected centres as service providers and a core group of universities and research institutes as identity providers, i.e. allowing their researchers seamless access to all CLARIN services. In the course of the project the federation will grow by accepting additional identity providers in parallel with the development of the Dutch national identity federation that is currently being constructed by SURFnet. The collaboration with the leading parties in the field of identity federation building will guarantee that eventually all Dutch research institutes that are eligible for NWO funding will be able to access the services provided by the CLARIN resource provider federation. The affiliated researchers, students and other members officially registered at one of the research institutes will be able to access the offered services with a single identity and single log-in.

In accordance with the general CLARIN rules the selected Dutch centres need to fulfil a number of requirements such as to associate their servers with accepted certificates, to install and integrate middleware components such as Shibboleth and to guarantee the quality and accessibility of their services for a certain number of years. With respect to the technical aspects CLARIN-NL will help selected centres to become a fully functional CLARIN resource providing federation. Based on simplified licence models it will be easy to come to an agreement with the Dutch identity federation about the terms of access at the national level. Due to the continuous interaction between TERENA and SURFnet for example, we can assume that decisions about the attributes with which users will be described in the Dutch identity federation will be compliant with the general trends in Europe.

The availability of efficient federation technology will have ramifications to the R&D aimed at the development of a sustainable business model for the use of language resources, with far-reaching implications for novel IPR agreements. Any advantage in this respect will be of high relevance for European initiatives such as the Alliance for Permanent Access for which the harmonization of licence conditions is one of the topics.



Registry Services for different sorts of information (researchers, metadata, PIDs, centres, concepts, relations etc) will have a very important role in the common European LRT market place. In CLARIN-NL these services will be developed under the guidance of the Dutch scientists and technologists who already have leading roles in the LRT community. Beyond the realization of distributed authentication mechanisms which is the core of federation technologies, MPI and INL have already implemented and tested distributed but integrated domains for metadata and persistent identifiers in the DAM-LR project. At the Dutch and European levels centres such as INL and MPI will offer appropriate registry mechanisms and portals. In addition, developing the CLARIN infrastructure will profit from recent standardization efforts in ISO TC37/SC4 (Subcommittee focusing on Language Resource Management) and TEI that are focusing on generic models for linguistic resources types such as lexicons and on frameworks that could help bridging gaps between different vocabularies used. In ISO TC37/SC4 Tilburg University and MPI play major roles, for example in defining the ISO Data Category Registry model, developing the *ISocat* (<http://www.isocat.org/>) data category registry software, and in defining metadata and semantic annotation categories. The latter is based on the experience of earlier initiatives such as SAMPA, EAGLES, IMDI, etc. Despite successful previous efforts, interoperability on the semantic level remains a challenge that needs additional efforts by CLARIN-NL to provide a simple and user friendly framework to create, manipulate and share ontologies that refer as much as possible to registered concepts. Due to the strong involvement of Dutch scientists and technologists in these efforts, it is the intention to make this one of the core efforts of CLARIN-NL and in doing so, maintain the strong position in the worldwide activities. With respect to inter-operability at syntactic level CLARIN-NL will rely as much as possible on generic models such as the Lexical Markup Framework as standardized by ISO, since it can capture a wide variety of lexical resources regardless of different structure and content. Based on 10 years of experience with IMDI and other metadata sets such as Dublin Core, OLAC etc. and a deep understanding of their limitations, CLARIN will develop a flexible infrastructure for component based metadata with the aim to allow describing and registering all language resources and tools. INL, MPI, DANS and Tilburg have the potential to play a major role in defining the new metadata framework, which will be crucial to organize the European and Dutch LRT market place. This new infrastructure will become part of the ISO standardization process.

Language resources and the attendant eScience research tools are by definition dynamic. New concepts, resources and tools will be introduced and existing LRT will be updated, extended and improved. Therefore, we will need to create workflows, procedures and tools for keeping the registries intact and consistent. To manage the variety of objects and versions Persistent Identifiers (PIDs) will play a crucial role in an open research infrastructure. Many details remain to be investigated and defined, which will be one of the tasks in CLARIN-NL. CLARIN-NL will offer such PID registration and resolution services based on the well-known Handle System providing access not only to LRT centres, but also to corresponding centres in other research disciplines.



The success of these registry services depends very much on their seamless integration into the CLARIN infrastructure and its acceptance by researchers. CLARIN-NL, therefore, will spend much effort creating consensus about procedures among a large group of researchers and data providers who often have different backgrounds and different aims. An increasing amount of researchers is recognizing the benefits of registry mechanisms. Therefore, we are confident that we can make important steps forward in motivating researchers to accept and adhere to the procedures and workflows that CLARIN-NL will develop as part of CLARIN-EU. The single most important means for motivating researchers to participate will be offering excellent services.

Information Services in CLARIN will be web services in a Service Oriented Architecture that offer LRT in encapsulated form so that not only humans, but also programs can access them. The encapsulation will guarantee that researchers can re-use existing resources and tools and combine them to new applications. So-called workflow frameworks which are already in use in other disciplines and in industry will not only allow technically versed users to create new powerful applications, but also users who lack the technical skills. CLARIN-NL will carry out the additional research and standardization efforts that are needed to come to a smoothly functioning SOA for LRT. Therefore, it will adopt a stepwise procedure and carry out model projects to study the interfacing problems in great detail. In parallel to web services CLARIN-NL will establish simple to use web sites to allow people to make use of the various services. MPI has extensive experience in building web services for LRT (LMF, DCR, metadata) and in turning existing resources and tools into web services in a SOA framework.

Part of the services will make computational requirements that are well beyond what can be handled by the common comparatively slow internet protocols. Evident cases are the processing of video recordings, part of which may be distributed over multiple sites. Therefore, CLARIN will need to develop an infrastructure for Grid computing. The know-how of SURFnet and the BIGGrid project will be used and the results of national developments will form input for the discussions in CLARIN-EU. Selected Dutch centres will participate in conducting the Grid computation tests. Although Grid computing remains challenging, past experience and performance guarantees that we will be able to develop excellent operational technology and workflows.

User-Oriented Approach While it will be possible to hide the technical and operational complexity of SOA services for the Humanities researchers, we are still facing the problem that future *eScience* services are inherently complex and that it is impossible to define and fix the functionality of those services once and forever from the very start. Inherently complex services require excellent user interfaces. Therefore, CLARIN-NL will pay due attention to user interface design and to real-world usability tests. In addition, CLARIN-NL will follow a user centred design approach in defining the services, by working closely together with the researchers who will use the services. Special attention will be devoted to novel ways in which the SOA approach can facilitate adding and changing services, tools and resources in the CLARIN infrastructure. Due to the limited funds for these aspects at the European level, CLARIN-NL will focus on these aspects in the construction phase.



IPR and Business Models form another essential aspect of CLARIN, both at the NL and EU level. Here, technical, legal and economic issues are at stake, and decisions at the legal or economic level will affect choices at the technical level. In CLARIN-NL we intend to develop sustainable IPR and Business Models that can be generalized to the European level.

For Humanities research it is essential that scientists have access to the full and raw versions of textual and multimedia documents (rather than to information derived from the documents, such as a list of words). However, full access to the documents raises fundamental problems with respect to property rights. This is not only the case for content producers (print and multimedia) but also for museums, cultural heritage institutes, etc. CLARIN-NL will launch projects aimed at solving these problems. In these projects we will collaborate with relevant organizations and projects at the European and global level, if only because organizations such as the Alliance for Permanent Access need to address the same problems. Eventually, these activities must result in a novel business model that provides access to LRT much in the same way as SURFnet provides the Dutch research community access to the ICT infrastructure.

Access to raw data may incur issues related to privacy protection, for example in the case of personal data collected in language pathology projects. Special access protection needs to be implemented for such ‘sensitive’ data, and it is a technical and user interface challenge to accomplish this without undue burden for the researchers who are allowed to access those data. Comparable constraints may apply to scientists who are not directly affiliated with a CLARIN institute and need to access some of the data managed by CLARIN. Here too, an integrated set of legal, technical and economic measures must be taken to find the best possible compromise between protecting investments and supporting research,

CLARIN-NL will also need to develop and test-drive the business infrastructure that makes essentially all digital language resources accessible for researchers in the Netherlands (and in the CLARIN-EU context also for researchers in Europe) with a single identity and a single and transparent licence and cost structure. For the development of the latter part we will need contributions from experts in information law and information business. Here too, we see no problems in attracting top-ranking scientists, because of the far-reaching implications of the CLARIN-NL outcomes for developments in the Publishing Industry where (partly) Dutch companies such as Wolters-Kluwer and Reed-Elsevier have high stakes.

Developers’ base At this moment the emerging LRT infrastructure is fully dependent on a small number of highly motivated researchers and software engineers. This makes the infrastructure quite vulnerable and brittle. CLARIN-NL aims to educate a much broader base of scientists and engineers who will be able to provide a sustainable and solid infrastructure.



Appendix C CLARIN-NL Initial Partners

| Participant | Unit | Acronym | Location | Project leader | Director |
|---|--|---------|-----------|----------------------------|---------------------------|
| Universiteit van Utrecht | UIL-OTS | UU | Utrecht | prof. dr. Jan Odijk | prof.dr Martin Everaert |
| Universiteit van Utrecht | Landelijke Onderzoeksschool Taalkunde | LOT | Utrecht | prof.dr Henriette de Swart | prof.dr Martin Everaert |
| Max-Planck-Institute for Psycholinguistics | | MPI | Nijmegen | Ir. Peter Wittenburg | prof dr. Ann Cutler |
| Digitale Bibliotheek voor de Nederlandse Letteren | | DBNL | Leiden | Cees Klapwijk | Cees Klapwijk |
| Instituut voor Nederlandse Lexicologie | | INL | Leiden | drs. Remco van Veenendaal | dr. Jeannine Beeken |
| Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) | Meertens Instituut | KNAW | Amsterdam | drs. Douwe Zeldenrust | prof.dr. Hans Bennis |
| Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) | Huygens Instituut | | Den Haag | Dr Karina van Dalen-Oskam | Dr Henk Wals |
| Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) | Data Archiving and Networked Services | DANS | Den Haag | Dr. Dirk Roorda | dr. Peter K. Doorn |
| Radboud Universiteit Nijmegen | Centre for Language and Speech Technology | CLST | Nijmegen | dr. Nelleke Oostdijk | prof.dr. Lou Boves |
| Radboud Universiteit Nijmegen | Centre for Language Studies | CLS | Nijmegen | prof.dr. Pieter Muysken | prof.dr. Ans van Kemenade |
| Universiteit van Amsterdam | Intelligent Systems Lab Amsterdam | ISLA | Amsterdam | Prof.dr. Maarten de Rijke | Prof.dr. P.M.A. Sloot |
| Universiteit van Groningen | Center for Language and Cognition | CLG | Groningen | dr. Erik Tjong Kim Sang | prof.dr.ir. John Nerbonne |
| Universiteit van Leiden | Centre for Linguistics | LUCL | Leiden | prof.dr. Johan Rooryck | Prof. dr. Jos Schaecken |



| Participant | Unit | Acronym | Location | Project leader | Director |
|---|--|----------|-----------------------|---|----------------------------|
| Universiteit van Tilburg | Tilburg Centre for Creative Computing. Department of Communication and Information Sciences | | Tilburg | Prof.dr. Antal van den Bosch | Drs. Lex Oostrom |
| Universiteit Twente | Human Media Interaction Group | HMI | Twente | Roeland Ordelman | prof.dr. Franciska de Jong |
| Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) | Frysk Akademie | FA | Leeuwarden | Dr. Arjen Versloot | Prof.dr. Reinier Salverda |
| Katholiek Documentatie Centrum | | KDC | Nijmegen | dr. Vefie Poels | dr. Lodewijk Winkeler |
| Veteraneninstituut | | | Doorn | Stef Scagliola | |
| Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) | Internationaal Instituut voor Sociale Geschiedenis | IISG | Amsterdam | Drs. T. van der Werf-Davelaar | Prof. Dr Erik-Jan Zürcher |
| Kennisinstituut sociale en psychische gevolgen van oorlog, vervolging en geweld | | COGIS | Utrecht | drs. Frederiek Eggink | drs. Trudy Prins |
| Internationaal Informatiecentrum en Archief voor de Vrouwenbeweging | | IIVAV | Amsterdam | Shan Swarts | Marjet Douze |
| Koninklijke Bibliotheek Vrije Universiteit | | KB VU | Den Haag Amsterdam | drs. Paul Doorenbosch Prof.dr. Piek Vossen | dr. Martin Bossenbroek |