European Strategy Forum on Research Infrastructures (ESFRI)

1. Leading partner, contact person (if not the same) and partners

Apart from Utrecht University (leading partner) and MPI (main technological leading partner) the partners have been listed alphabetically. The partners listed are legal entities. Several partners have multiple participating units (e.g. different institutes). These have been listed as well.
Appendix D contains a list of Acronyms and their (English) expansions, for ease of reference.

Leading partner	Name, title(s):	male / female
	Institute: Universiteit van Utrecht Unit: UIL-OTS Name, title(s): prof.dr. Jan Odijk Address: Janskerkhof 13, 5613 BL Utrecht Telephone number (s): +31 253 6006 E-mail: jan.odijk@let.uu.nl	male
	Unit: Landelijke Onderzoeksschool Taalkunde (LOT) Contact person(s): Henriette de Swart	female
Partner 2	Institute: Max-Planck-Institute for Psycholinguistics, (Nijmegen) Unit: Contact person(s): Peter Wittenburg	male
Partner 3	Institute: Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) Unit: Contact person(s): Cees Klapwijk	male
Partner 4	Institute: Instituut voor Nederlandse Lexicologie (INL, Leiden) Unit: Contact person(s): Remco van Veenendaal	male
Partner 5	Institute: Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) Unit: Meertens Instituut Contact person(s): Hans Bennis	male
	Unit: Huygens Instituut Contact person(s): Karina van Dalen-Oskam	female
	Unit: Data Archiving and Networked Services (DANS) Contact person(s): Dirk Roorda	male
Partner 6	Institute: Radboud Universiteit Nijmegen (RU) Unit: Centre for Language and Speech Technology Contact person(s): Lou Boves / Nelleke Oostdijk	male/female
	Unit: Centre for Language Studies Contact person(s): Pieter Muysken	male
Partner 7	Institute: Universiteit van Amsterdam (UvA) Unit: Intelligent Systems Lab Amsterdam (ISLA) Contact person(s): Maarten de Rijke	male

Partner 8	Institute: Universiteit van Groningen Unit: Center for Language and Cognition Contact person(s): John Nerbonne	male
Partner 9	Institute: Universiteit van Leiden Unit: Centre for Linguistics Contact person(s): Johan Rooryck	male
Partner 10	Institute: Universiteit van Tilburg Unit: Tilburg centre for Creative Computing, Department of Communication and Information Sciences Contact person(s): Antal van den Bosch	male
Partner 11	Institute: Universiteit Twente Unit: Human Media Interaction Group Contact person(s): Roeland Ordelman	male

2. Title infrastructure: CLARIN-NL

3. Summary

CLARIN-NL aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent *e*Science working environment for researchers in the Humanities and Social Sciences (HSS) who want to make use of language resources and technology. 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 already now, CLARIN-NL is set up more as a programme than as 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 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 HSS 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.

4. Keywords: language resource infrastructure, humanities and social sciences, human language technology, inter-operability, Service Oriented Architectures

5. Science Case

Recent international reviews have shown that the Netherlands hold a top ranking position in all disciplines in the broad field of language studies, including theoretical, descriptive and historical (Dunn et al. 2005) linguistics, psycholinguistics, language acquisition studies (Bree et al. 2006), computational linguistics and language and speech technology. Externals assessments of the linguistic research in the Netherlands (UvA, VU, Tilburg, Nijmegen) in 2006 and a year earlier in Utrecht have qualified the linguistic research in all cases as 'excellent'. In Groningen, the computational linguistics group of CLCG obtained the highest possible score over 1998-2004. MPI was recently evaluated very positively and its technical group was characterized as "an outstanding world-class provider of the latest technologies for analyzing and storing multimedia data". The Huygens institute's research was recently evaluated as very good. So was the linguistic research at the Meertens institute. Meertens's variation linguistics group was considered an "international player", and some of the formal linguistics research was judged to be "excellent".

The Landelijke Onderzoeksschool Taalkunde (Netherlands Graduate School of Linguistics, LOT) has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW) on three consecutive occasions, and was internationally evaluated highly positively, as being "without parallel" and "having a unique position in European linguistics".

Prestigious grants were awarded to researchers active in the field of language studies, including *Spinoza* grants to Ann Cutler (MPI) and Pieter Muysken (Nijmegen), two ESF EURYI grants (Nijmegen, Meertens), an ERC Starting Grant (Nijmegen), and an ERC Advanced Grant (Nijmegen). Six *Pionier* grants were awarded to researchers from Utrecht, Groningen en Nijmegen, and several *VIDI* and ten *VICI* grants were awarded to researchers from this field. Several researchers from the field are KNAW professors or KNAW members.

To a large degree this top ranking position could be obtained because Dutch researchers have played (and are still playing) a leading role in the development of language resources, the collections of textual data, lexical databases, spoken corpora and the tools that are needed for ground breaking research in the field. Examples of corpora and lexical resources include the CELEX Lexical Database developed by the MPI and Nijmegen University, the EuroWordNet initiated and coordinated by the University of Amsterdam, the SpeechDat corpora for almost all European languages modeled after the seminal Dutch Polyphone corpus (coordinated by the Speech EXpertise Centre SPEX in Nijmegen), the Spoken Dutch Corpus (a project with participation from all relevant universities and in collaboration with Flanders, managed by Nijmegen University), the TDS Typological Database System (multiple universities, managed by Utrecht), the MPI and DoBeS Archives held at MPI for endangered languages, and other MPI collections pertaining to gesture, multilingualism, child language and Dutch Sign Language. In addition, many tools have been developed in the Netherlands. They include COREX, a powerful and flexible software package for corpus exploitation developed in the context of the Spoken Dutch Corpus project, the comprehensive Language Archiving Technology tool set for making annotated multimedia recordings accessible for scientific research (both MPI), the Alpino parser[1], the L04 package for analysing pronunciation differences (both Groningen), a rich set of tools based on memory-based learning developed by Tilburg University (Daelemans & Van den Bosch 2005), the PRAAT tool for the analysis, generation and manipulation of speech signals (Amsterdam), and several tools for multimodal applications in the context of the IMIX programme. Of course, the success and the international visibility of the efforts dedicated to the development of corpora and tools is to a large extent due to the fact that the scientists driving and guiding the developments were already considered as leading in the field. This leadership position has only been strengthened by the research that can now be conducted thanks to the availability of these resources, witness for example the various studies on the basis of the

[1] The Alpino parser has also been used in a variety of applications, including a very successful question-answer system. See Bouma et al. 2005.

Spoken Dutch Corpus which already have resulted in a large number of master theses, PhD theses and many articles in conference proceedings and journals (e.g. Kuperman et al. 2006).

Development of Dutch language resources continued within the current joint Dutch-Flemish STEVIN programme (http://taalunieversum.org/taal/technologie/stevin/). One of the aims of the STEVIN programme is to ensure that by the end of the programme (2010) a complete basic language resources kit (BLARK) for Dutch will be in place. The highly positive 2008 mid-term evaluation of the programme underlines the programme's success and identifies it as leading in Europe. National research bodies in other countries look upon the programme as a showcase which they are eager to take as an example for how they should go about creating their own national language research infrastructure.

As more resources become available there is the ever more pressing need for integration: inter-operability is a necessary precondition when it comes to exploring the resources' potential to the full. In order to make existing and emerging resources, most of which have been (and are being) developed independently of one another and with specific applications in mind, available for scientific research and the subsequent development of civil applications, additional steps are needed. Independently constructed corpora for different languages or different variants of the same language, annotated with different terminology, need advanced tools such as taxonomies, ontologies and complex mapping functions to become inter-operable.

Several Dutch institutes and experts are leading partners in activities that are meant to overcome inter-operability problems, e.g., in the ISO TC37/SC4 work in Semantic Annotation, the Lexical Markup Framework standard, the ISO Data Category Registry model and the ISO standard for citation of and referencing to electronic resources.

The *Typological Database System (TDS)* project is an example which shows the advanced state in the Netherlands especially with regard to implementing inter-operability. Ten different typological databases were investigated, brought into a common infrastructure and semantically connected by an ontology, allowing users to do cross database searches, statistics etc. CLARIN-NL includes definite plans to enhance the scope of the successful integration methodologies developed in this project by interfacing the TDS with other data sources such as the archive at the MPI and data from Leiden and Amsterdam via state-of-the-art web services and semantic web technologies.

An integrated language resources infrastructure with inter-operability functionality is certain to spark off new, exciting and innovative research. CLARIN-NL seeks to develop such a national distributed language infrastructure that can be used in the humanities and social sciences, offering open access to a wide range of resources, services and applications along with supporting activities such as workshops and training courses. It thereby opens up the full potential for *eScience* research in this domain, which will further strengthen the already excellent Dutch scientific position in Europe. The infrastructure and language processing services that will be developed and integrated in CLARIN will make it possible to find and combine related pieces of information from seemingly unrelated text and multimedia sources, thus creating options for groundbreaking new discoveries and highly original views on data that will certainly lead to new scientific insights in the exact same way as we have witnessed in the physical and life sciences. Thus, CLARIN-NL will help advancing boundaries to a true eScience scenario in the Netherlands.

6. Talent Case

As described in the preceding section, the CLARIN infrastructure will help open up the full potential of doing eScience on language-related documents in the domain of the humanities and the social sciences. This will create abundant opportunities to develop new research methods and research lines and carry out groundbreaking research, which will be attractive to researchers and PhD students from other European countries as well as from non-European ones. It will also provide a highly stimulating educational environment for undergraduate and graduate students from the Netherlands and abroad. Since the Netherlands already has a leading position in CLARIN-EU, a position that will be strengthened by CLARIN-NL, other researchers and students will be eager to join the Dutch research groups. Though it is true that the location where the research is carried out in terms of access to data and services is made irrelevant by the very nature of the CLARIN infrastructure, the Dutch research groups remain interesting because they are in the lead in designing, constructing and exploiting the infrastructure, and they will be the first to have built up significant expertise in using the infrastructure. Users will be able to create virtual collections including resources from different institutes and performing operations on them, as well as to combine various tools to new

applications by workflow mechanisms. Expertise in using such advanced research methods will attract young researchers and foster not only language research in Netherlands, but it will also have an outreach to all disciplines working with similar paradigms and language resources.

In the budget, we have set apart a significant sum to stimulate such *brain gain* actively, by inviting leading experts from abroad as well as for creating openings for foreign PhD students and trainees. This will allow us to closely collaborate with foreign top researchers and infrastructure specialists, which will undoubtedly lead to a great deal of synergy, and an improved Dutch research environment, attractive to foreign students and researchers. We will seek additional funding through programmes such as the Marie Curie programme (International Training Networks), as well as regular projects in the SocioEconomic Science and Humanities and the Information & Communication Technologies sub-programmes of the FP7 Cooperation Programme.

CLARIN-NL is in a position to take a leading role in the CLARIN enterprise as it extends throughout Europe. A comprehensive set of language resources for Dutch are in place or will be so shortly (by 2010 when the current STEVIN programme comes to an end), the technology is there and together the leading participants in CLARIN-NL have all the know-how, experience and the drive that are needed to create a language resources infrastructure that puts the Netherlands at the forefront internationally.

7. Innovation case

The CLARIN-NL programme encompasses a dedicated effort to include in its infrastructure language-based cultural heritage resources, many of which are currently 'hidden' with individual researchers or institutes This will make these data more visible to the research community. In addition, it will improve the ways in which these data can be accessed, by making the access technically simpler, and by including tools and services that allow more intelligent ways of exploiting the data. It thus will create easier and more intelligent ways of accessing language-related cultural heritage data. The research infrastructure is aimed at humanities and social sciences researchers, so these will be the first to benefit from this infrastructure. It will make it possible for them to address completely new research questions and to approach old research questions in a completely new manner, i.e. it will revolutionize research processes in the humanities and social sciences. Accompanied by an intensive programme of dissemination and training and education of new students in using the infrastructure, it will create a whole new generation of researchers who become familiar with these innovative *eScience* methods at a very early stage, thus offering the best possible platform for achieving scientific breakthroughs and excellent career opportunities.

The global society has entered the Information Age, and we are already living in an Information Economy. While much of the information that is presently being used is sitting in structured databases, entered mainly by hand or by simple automatic data entry procedures, it is abundantly clear that there is much more (and perhaps more valuable) information that is hidden in unstructured documents such as newspaper articles, reports, blogs, audio and video news reports and other varieties of textual and multimedia information.

Social Climate Change is a hot issue in the current society: migration and mobility as well as cultural identity in an increasingly multi-cultural society pose new challenges. The CLARIN infrastructure with its sophisticated eScience tools will allow social scientists, once given access to the 'hidden' information in documents and audio and video recordings, to detect and analyze changes in society caused by such factors (e.g. signals of attitude change (e.g., using opinion mining techniques), language contacts, changes in language and language use, etc.) and thus provide a solid factual basis and analysis tools for scientific research and policy decisions in these matters.

The CLARIN-NL programme, will offer, by its innovative nature, new opportunities for researchers in other disciplines as well. Moreover, many people and groups in society can benefit from tools that can reveal information hidden in textual and multimedia data. Investment decisions on the stock market depend to a large extent on information about companies that is 'hidden' in reports, some of which are published by the companies themselves, many by persons and organizations that have some interest in a company. Internet Publishing has been known to affect the stock price of companies considerably as well as the popularity of artists and politicians. The eScience tools created in CLARIN will provide new opportunities to unlock the information hidden in textual and multimedia documents for other domains than humanities and social science proper.

In the past spin-off companies (e.g. Polderland, Textkernel, and Gridline) have successfully valorized the outcomes of language research. Though not primarily aimed at commercial applications, the CLARIN infrastructure with its easier and more intelligent access to texts, multimodal documents and cultural heritage data will offer new opportunities for commercial applications in the areas of education, entertainment and information services. The CLARIN infrastructure will provide the opportunity to develop and test such applications and services, after which companies (possibly spin-off companies) will exploit the ones that are most successful and appealing to a large audience. It is therefore reasonable to expect that CLARIN will spurn off similar start-up companies, as well as commercial and social innovations.

With innovative combinations of meta-data and content-based access to multimedia productions, newspaper articles and web pages it will be possible to develop new media products that are of interest to various users, from professional users to students in educational programmes. For example, it will be possible to create interactive museums, not only for artifacts that have been canonized as 'art', but also for artifacts of the very recent past. This would in effect put the entire world within the reach of anyone interested. Initial steps in this direction have already been taken in several projects in the NWO CATCH programme[2]. Examples include the CHIP project that aims to disclose a museum collection by providing personalized tours through the virtual cultural Netherlands, and MuSeUM and STITCH projects that attempt to disclose the content of a museum in a single, unified system even if all sub-collection are organized by systems that differ in purpose, design, tradition and access methods. Other CATCH projects (e.g. CHOICE) investigate automatically enriched metadata descriptions to increase accessibility.

Other examples are novel tourist packages to be developed in collaboration with the ANWB and VVVs, working with police organizations and security agencies to intercept terrorist activities and cybercrime, or the development of new media products in collaboration with the *Instituut voor Beeld en Geluid*, the *Koninklijke Bibliotheek*, and publishers such as PCM. In CLARIN-NL, commercial spin-offs will be stimulated actively. First by means of a wide range of dissemination activities aimed at humanities and social sciences researchers, but also beyond these groups, to create awareness of the CLARIN infrastructure and the potential it offers. We will welcome new users of the infrastructure, and help them use the facilities in the best possible way by providing courses and training sessions. We will also assist organizations or companies that consider using the CLARIN infrastructure and services in the context of other disciplines, for example by defining and carrying out projects aimed at developing CLARIN-related functionality for completely different domains together.

8. Partnership case

One reason why all branches of language research in the Netherlands are of top quality is the excellent organization of the field. Starting with the NWO Foundation for Linguistic Research and its Working Groups in the seventies, Linguistics has constantly improved the national coordination and collaboration. After the dissolution of the NWO Foundations, Linguistics has built on other organizations, such as LOT, CLIN, Anéla, etc. In addition, several linguistic disciplines have joined forces in government supported innovation programmes, such as SPIN, IOP, etc. Most recently, the Dutch-Flemish programme STEVIN has provided a very strong platform for collaborative research, resource creation and service development.

The CLARIN-NL partners can be divided in three groups:

- Institutes that focus primarily on research, including the development of new tools
- Institutes that focus primarily on data, more specifically on acquisition and storage of data as well as giving access to data
- Institutes that are involved both in research and in data acquisition, storage and access.

The CLARIN-NL partners with a primary focus on research and tools include the universities of Utrecht, Groningen, Leiden, Nijmegen, Amsterdam, Twente and Tilburg.

The CLARIN-NL partners with a primary focus on data include INL, DBNL, and DANS.

The Meertens, Huygens and Max Planck institutes are involved both in research and in data.

[2] http://www.nwo.nl/nwohome.nsf/pages/NWOA_7ANC86

Partners involved in research will contribute to CLARIN in two ways: (1) by providing language-based services developed around language technologies investigated in these institutes, and/or (2) by providing user requirements that can serve as input for specifying the CLARIN infrastructure thus ensuring that the eventual infrastructure will be optimally suited for their own research needs, and by evaluating preliminary versions of the CLARIN infrastructure in the context of their research practice.

The partners involved in data (i.e. acquisition, storage, access of data) will contribute to CLARIN by making available their data, converting and or enriching these in order to integrate them into the CLARIN infrastructure, and (for some) to host servers that allow researchers to deposit and access the data, and run services that form part of the distributed CLARIN infrastructure.

Two institutes have a somewhat special role: Utrecht University will be the general coordinator of CLARIN-NL (as it is for CLARIN-EU), while MPI has a special role as a major actor and coordinator in the design and implementation of the technological infrastructure. All CLARIN-NL partners in this proposal have already been active in the preparation of the CLARIN-EU project and have registered as Dutch participants. Of course, CLARIN-NL is open to new partners and maintains active contacts with other relevant players in the field, who may join CLARIN-NL in a later stage. For example, organizations recently approached such as KB, Fryske Akademy, KDC, *Veteranen Instituut*, IISG and IIAV have already shown their interest and willingness to contribute, and others (e.g. COGIS, CPG, and NIOD) are considering participating.

The research institutes will make contributions to CLARIN on the basis of their particular expertise. For Utrecht, this includes expertise around the integration of diverse sets of data gained (inter alia) in the TDS project, expertise on semantics, pragmatics and discourse, and in developing language services in the context of *e*Learning. For Groningen it involves expertise in parsing implemented in Alpino and applications built around it, and on visualization of dialect variation. For Nijmegen (CLST) it involves expertise in speech technology and in the creation and validation of spoken and written corpora, and the application of language and speech technology in computer assisted learning and augmentative communication. For Nijmegen and Twente, it involves expertise in finding structure in raw multimedia recordings and automatic annotation tools. For Tilburg, it concerns expertise in memory-based learning and a variety of language technologies based on this approach. For Amsterdam, it involves expertise on search, question-answering, and language modelling.

Research institutes will also provide input for the requirements and specification of the CLARIN infrastructure as well as for the evaluation of preliminary versions of the infrastructure against their research needs. This holds in particular for the Meertens and Huygens institutes, for Nijmegen University (CLS) and MPI, but also for all universities that carry out research in any sub-discipline of the humanities and social sciences.

As to data centres, it can be stated that the Netherlands is home to a number of centres that can develop into sustainable language resource centres. The most obvious ones are the *Instituut voor Nederlandse Lexicologie* (INL), the Meertens Institute, and the Max Planck Institute for Psycholinguistics. Perhaps somewhat less obvious, but still serious candidates for developing into language resource centres are the *Koninklijke Bibliotheek* (KB), the *Instituut voor Beeld en Geluid* (BG), archives, museums, etc., with DANS as an option for long-term preservation and access and as playing a gateway to other humanities disciplines. DBNL is a foundation that makes Dutch literature available electronically, and must therefore play a role in CLARIN-NL

These organisations are largely complementary in terms of areas of expertise and activities they focus on. By virtue of their charter INL and KB, as well as DBNL, are almost exclusively focused on written Dutch, DBNL focusing even more narrowly on Dutch literature. Thus, in the multi-dimensional space of language resources they occupy a clearly demarcated sub-space. Meertens shares with INL/KB a focus on the Dutch language area, but its focus is on regional variants, rather than on the general (official) language. In addition, the charter of Meertens includes ethnographic data in addition to what is conventionally considered 'language' data. The focus of MPI is on a wide range of fundamental aspects of linguistic research, including psycholinguistics, cognitive anthropology, language acquisition, etc. In ongoing and future research projects there is a focus on multi-modal interaction in which speech and gesture are naturally combined. Another aspect that sets MPI apart from INL and Meertens is the focus on minority and endangered languages. Therefore, MPI differs from the other emerging resource centres in that the focus is on multi-media recordings and languages other than Dutch, often languages for which no official lexicons and grammars exist. While BG also focuses on multi-media documents, their charter is limited to the preservation of Dutch media productions; it does not include research, nor the preservation of foreign media productions.

DANS is another partner with a somewhat special role. Since its foundation in 2003 DANS aims at storing research data for the Humanities and Social Sciences, and at securing permanent and persistent access to these data. For this aim DANS develops and maintains sustainable archiving services and helps other organizations to use existing DANS services or, if that is not feasible, to develop their own services that are compatible with DANS guidelines. In addition, DANS Works with data providers so as to maximize the accessibility of all kinds of data for research in HSS. One way to approach the goals is for DANS to participate in relevant projects, nationally and internationally. Thus, DANS is a member of CLARIN-EU, and a partner in CLARIN-NL.

The CLARIN infrastructure, thanks to its distributed nature, will allow these centres to remain independent, each with its own charter, and nevertheless to collaborate closely and create synergy by accessibility to each other's complementary data and various services to explore and manipulate these data.

On the European level, national governments collaborate in the CLARIN-EU project. This EU-funded project aims to support and coordinate various national initiatives as they are planning and setting up their national language resources infrastructures. To this end CLARIN-EU will concern itself inter alia with the implementation of a prototype and address questions such as 'how do we unite existing archives from all over Europe in a single federation?, 'how do we ensure that tools and other resources that originate from sometimes widely different research paradigms can be applied in different settings?', and 'what can we do to support languages that are less technologically advanced and for which there are as yet only few resources available?'. CLARIN-EU is a project in which 32 members from 22 EU member and associated states participate and the CLARIN network encompasses 131 organisations from 32 EU member and associated states. With the federation of archives, the sharing of resources and technologies among CLARIN partners, and the development and usage of common expertise and advisory centres, CLARIN has a large critical mass and can achieve things that no individual research institute or university, or, at the European level, any individual country could ever do, thus creating significant synergy and increasing the possibilities to carry out research on a very large set of (mono- and multilingual) data otherwise practically not accessible using state-of-the art intelligent methods of accessing and manipulating these data. It will stimulate cross-disciplinary work and allow the formation of virtual research groups, thus overcoming physical and geographic constraints and making it possible to carry out research in optimally balanced research teams.

Utrecht University and MPI were the major driving forces in the conception and preparation of CLARIN-EU (in cooperation with the Hungarian Academy for Sciences and the Oxford Text Archive) and are recognized as such by the 32 consortium partners and all network participants in the CLARIN initiative. This leading role will be significantly strengthened by the CLARIN-NL project, especially since the CLARIN-EU project must work with a modest budget. The CLARIN-NL partners will therefore be the prime instigators in setting up the design and prototype implementation to be developed in CLARIN-EU, and will, in the context of the CLARIN-NL project proper, extend it and develop the prototype into a full-fledged infrastructure, in particular for the Netherlands. This will also ensure a seamless integration of the Dutch part of the infrastructure into the larger European-scope infrastructure.

In addition, there is close collaboration and cross-participation with the *Fostering* La*nguage* Re*sources* Net*work* (FlareNet, http://www.ilc.cnr.it/flarenet/), which is starting up its activities early September 2008. FlareNet is a European forum to facilitate interaction among language resource stakeholders, more specifically to identify those priority areas of language resources that are of major interest for the public and that need public funding to develop or improve, in the form of a blueprint of actions that will constitute input to policy development both at the EU and national level. The leading role of the Netherlands in CLARIN and the participation of the European coordinator as well the national CLARIN-NL coordinator in FlareNet will therefore be highly influential for the course of this network project.

A related project is DARIAH. CLARIN and DARIAH will work closely together at the European level, and this collaboration should also be achieved at the Dutch level where CLARIN would see itself as one hub in an emerging landscape of infrastructures for the humanities and social sciences. It should be the primary task of DARIAH to reach out to various humanities and social sciences (sub-) disciplines and CLARIN should offer its services in this framework to all disciplines. As in other countries, an interaction with initiatives like the Council of European Social Science Data Archives (CESSDA) is intended and will be pursued in CLARIN-NL.

Although an inventory has been made of the types of activities that are necessary to develop and implement the CLARIN infrastructure and introduce it in the humanities and social sciences

research communities, it is in this stage not possible to specify in detail which partner(s) will be involved in which activity. In part this will also depend on the outcomes of the CLARIN-EU preparatory phase. The governance structure must therefore be set up in such a way that a range of calls for tender for selected well-defined tasks can be organized, as well as some open calls for projects that fit in CLARIN-NL but are not precisely predefined.

The governance structure of CLARIN-NL consists of the following bodies: a Programme Committee (PC), a Board, a Support Office (SO), and an International Advisory Panel (IAP).

- The CLARIN-NL PC initiates and executes a policy and procedures of assignment of funds to projects (organize calls, evaluate submitted proposals, prioritize them, etc.). It advises the Board. Composition: a representative sample of the important researchers in the field selected from CLARIN-NL partners and chaired by the national CLARIN-NL coordinator (approx. 8 persons).
- The CLARIN-NL Board takes decisions and monitors whether the other bodies adhere to the policy and procedures set-up by the PC. Composition: representatives of funding agencies + some experts in the field (approx. 6 persons).
- The CLARIN-NL SO supports the governance bodies, in particular the Board and the PC with the execution of the policy and daily follow-up of projects and other activities (such as dissemination activities).
- The IAP evaluates project proposals, and advises the PC and Board in this as well as in other matters. It carries out the mid-term and final evaluation of the CLARIN programme. Composition: internationally renowned experts in the field, e.g. top players in CLARIN-EU and/or other national CLARIN programmes.

Special measures and procedures will be set up to avoid conflicts of interests (since the members of the PC will very often also be involved in project submissions). The procedures used in the IMIX and STEVIN programmes form a solid basis for this.
The Board has the option to set up committees, chaired by a PC-member, if needed with participation of PC-external experts, e.g. to organize special events or to investigate specific matters.

This governance structure resembles the organization of several recent programmes. CLARIN-NL partners are familiar with this type of organization which has proven to function pretty well. Obviously, in the first phase of the project, the details of this governance structure will have to be worked out, as well as the relation with CLARIN-EU. CLARIN-NL will take care that its experts will participate in the various working groups that are discussing the actual work at the CLARIN-EU level to bring in the Dutch expertise, but also to guide the discussions at national level.

9. Business case

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. 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.

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 in Appendix A.

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
    - broad participation in the definition of standards
    - conducting validation projects to establish the requirements for the infrastructure
    - making an inventory and analyzing the needs of the national research community
- 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 in Appendix A.

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. Appendix B contains some concrete examples of this kind of application project. 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.

Finances

The budget for the proposed CLARIN-NL programme is specified in some detail in Appendix C. 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 2.85M€, for the NL-line 19.67M€, and 22.52M€ in total. The yearly exploitation costs in the exploitation phase are budgeted at 3.12M€, which also forms a good basis for determining the yearly exploitation costs beyond 2014.

Estimation of actual use of the infrastructure

All participants in this proposal will be users of the infrastructure. In addition, the expression of interest and willingness to contribute by KB, Fryske Akademy, KDC, *Veteranen Instituut*, IISG and IIAV extends the potential users to these organisations and a range of other HSS researchers making use of their data and services. The user-centred design approach will ensure that the infrastructure will provide data and services that are actually needed by the users. In the exploitation phase, user requests for extensions and modifications will be closely monitored, prioritized and responded to, ensuring that the infrastructure continues to serves its users. In short, we expect that the whole community of HSS researchers that need to work

with language-related data will be users of the infrastructure. In addition, as described elsewhere, the infrastructure will serve as an exemplary case and therefore be explored for use outside of the social sciences and humanities as well.

10. Technical case

CLARIN-NL intends to create the technical infrastructure that is needed to provide a sustainable and persistent *e*Science working environment for researchers in the Humanities and Social Sciences that want to make use of language resources and technology. Building an infrastructure is different from the conventional research projects in the Humanities and Social Sciences. However, we are convinced that constructing a Resource Provider Federation that will rely on a Service Oriented Architecture (SOA) and also Grid services (where necessary) will be challenging enough to attract top-ranking scientists and engineers to build the infrastructure. Especially the focus on SOA should be appealing, because this approach will help government agencies to avoid extremely costly failures with monolithic ICT projects. 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. We are confident that CLARIN-NL will result in the education of a much broader base of scientists and engineers who will be able to provide a sustainable and solid infrastructure.

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.

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 *e*Science 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, but many details remain to be investigated and defined. 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.

While it will be possible to hide the technical and operational complexity of SOA services for the Humanities and Social Sciences researchers, we are still facing the problem that future *e*Science 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 HSS research it is essential that scientists have access to the full and raw versions of 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,

The CLARIN-NL Governance Structure will make sure that the core partners in CLARIN-NL will receive the resources that are needed to fulfill their obligations at the national level. This holds especially for the data centres mentioned in section 8 on *Partnership.* Infrastructure services (such as archiving services, metadata registry portals, PID services, WAYF registry, ISOcat services, etc.) will initially mainly be hosted by DANS and MPI, due to their know-how, experience and their technical infrastructure. The future allocation of these services will be elaborated and defined in detail during the project. In this respect CLARIN-NL will follow the principles currently being established at the European level, where all prospective member institutes must submit a self-assessment in which they describe their capabilities and support. The CLARIN-NL Board will base their decision on an evaluation of these descriptions. It is the intention of CLARIN-NL to maintain at least one mirror site of the main CLARIN-EU services at a Dutch institute in order to build up knowledge and to remain central in all networking.

Similar strategies will be applied to prepare decisions about which groups will receive funding for tool development and LRT integration projects. The CLARIN-NL Board will issue calls for tenders, inviting Dutch research and data institutes to submit proposals for the development of specific tools and services. In defining the tenders the Board will take due account of the developments in CLARIN-EU. At the same time due measures will be taken to guarantee that developments at the European level will be integrated in the national infrastructure.

The CLARIN-NL Board will determine which of the existing resources (data and tools) will be adapted and integrated at which stage and what funds can be made available for the purpose. This approach will result in a rich and exemplary infrastructure of services in the Netherlands at the lowest possible cost. To prove the functionality of the infrastructure and to promote its use by HSS researchers a mix of demonstration projects and individual integration projects will be funded.

11. National Road Map

Unique position of the Netherlands

Utrecht University and MPI play a leading role in CLARIN-EU and are recognized as such by the consortium partners and all network participants in the CLARIN initiative in 32 EU and associated countries. In this situation, the Netherlands has an excellent opportunity to determine its own profile CLARIN-EU. CLARIN-NL will offer the opportunity to maintain and further extend this unique driving position of the Netherlands in the European context. Foreign organizations or institutes who might challenge this unique Dutch position are members of CLARIN and therefore it may be expected that there will be synergy rather than competition. CLARIN NL will add to the strong position Dutch institutes have in several integration activities. CLARIN-NL, however, is to start soon in order to avoid losing this forefront position to countries that already have national CLARIN projects running (e.g. Denmark, Germany).

Focus
CLARIN-NL is a concerted effort to create a language research infrastructure with a clear focus on the disclosure of information hidden in language-related data by applying human language technologies to unstructured documents and by combining seemingly unrelated documents and information categories.

Critical Mass
For the design, construction and exploitation of archives MPI and INL are internationally recognized experts, who have built up and demonstrated their expertise in these areas in a variety of national and international projects (e.g. DAM-LR, ISLE, INTERA, LIRICS, and DoBeS). In the areas of language and speech technology, (players in these fields will be both co-developers and users of the infrastructure) the Spoken Dutch Corpus project and the IMIX and STEVIN programmes are internationally recognized as being of the highest quality and even exemplary for the field. Linguistics (broadly construed) in the Netherlands is internationally renowned and backs CLARIN massively (15 organizations or institutes from 11 legal entities from the Netherlands have registered at CLARIN-EU, and the number is growing). Players in this field will typically be the first users of the infrastructure.
Important institutes that maintain large data sets such as INL, DBNL (Digital Library for Dutch Literature), MPI, the Meertens and Huygens institutes support and play an active role in CLARIN. DANS has a central role in the DARIAH infrastructure initiative bringing together institutes that want to offer their service to the humanities at large.

Embedding
As stated above, 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*), 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 Grant awarded to Pieter Muysken (RU), DoBeS, the Multilingualism Project (MPI & RU), and the Sign Language Project (RU et al).

Cooperation
There is an explicit desire of 15 institutes and organizations from 11 legal entities in the Netherlands to cooperate within CLARIN (they are registered participants in CLARIN-EU), and their number is growing. They have pronounced their willingness to contribute in one fashion or another (though this has not been formalized yet). Many of them participate in a project (FIDLR-Start) to prepare the FIDLR proposal for a project that will be directed at making linguistic databases suitable for carrying out linguistic research in the CLARIN infrastructure (to be submitted in NWO-Groot). In addition, 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. We have close contacts with the CLARIN coordinator in Flanders and will soon investigate what forms our collaboration could take.

Developments in Society
CLARIN can directly contribute to issues related to *Social Climate Change*, as described above. CLARIN offers possibilities to quickly discover changes in this domain. This allows the development of policies to avoid that groups of people feel alienated and measures can be taken to keep groups of people involved in the society, which is perhaps the most important factor in ensuring a pleasant and safe society.

New challenges are also posed by the increasing amount of information that is created and distributed in other domains. Language and speech technologies and language resource technologies will be key technologies to take on these challenges. Therefore CLARIN, in which these technologies play an essential role, will contribute to addressing these challenges and turning them into opportunities. Some of these have been described in the *Innovation case* section.

CLARIN will also contribute to innovation programmes such as *Zorg* (*Care*) and *Veiligheid (Safety)*. For example, it is well known that information about most of the preparations for the 9-11 attacks was available, but not used. In a similar vein information that could have prevented the murder of Theo van Gogh was available, but not used, because there was no adequate technology for analyzing and prioritizing documents. The CLARIN infrastructure, with its standardized metadata schemata and innovative way of uncovering information hidden in textual and multimedia documents using content-based access tools, offers excellent opportunities to address such problems, thus contributing to safety in the society.
Furthermore, many of the basic technologies behind advanced ways of accessing and manipulating data in the infrastructure are also applicable in other services and applications relevant in themes such as *Zorg* and *Veiligheid*. Many of these technologies are in fact already now put to use in these domains.
CLARIN will therefore, in various ways, also contribute to these innovation domains.

Finally, since the set-up of a language resources infrastructure necessarily requires the development of inter-operable data and services, and since CLARIN-NL has set aside a significant sum for dissemination and training, it directly contributes to the *ICT Agenda*, which has *inter-operability* and *eSkills* as key targets.

References

[Bouma et al. 2005] G. Bouma, I. Fahmi, J. Mur, G. van Noord, L. van der Plas, and J. Tiedemann (2005), Linguistic Knowledge and Question Answering. *Traitement Automatique des Langues (TAL)*, 2005/3.

[Bree et al. 2006] E.H. de Bree, F.N.K Wijnen. & W. Zonneveld (2006). Word stress production of three-year-old children at risk of dyslexia. *Journal of Research in Reading, 29*(3), 304-317.

[Daelemans & Van den Bosch 2005] W. Daelemans and A. van den Bosch (2005). Memory-Based Language Processing. Cambridge, UK: Cambridge University Press. ISBN-10: 0521808901 | ISBN-13: 9780521808903.

[Dunn et al. 2005] M. Dunn, A. Terrill, G. Reesink, R. Foley & S. Levinson (2005). Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 23 September 2005: Vol. 309. no. 5743, 2072 – 2075.

[Kuperman et al. 2006] V. Kuperman, M. Pluymakers, M. Ernestus, and R. H. Baayen (2006). Morphological predictability and acoustic salience of interfixes in Dutch compounds. *Journal of the Acoustic Society of America 51*: 2018-2024.

Appendix A: Action lines

This appendix lists the actions to be undertaken in CLARIN-NL for the European line of actions and for the NL line of actions.

EUROPEAN LINE

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)

NL LINE

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. integrate advanced infrastructure services into CLARIN, e.g.
    a. multimodal data (e.g. in collaboration with the Instituut voor Beeld en Geluid)
    b. special data (e.g. sign language)
    c. advanced language and speech technologies
    d. collaboration with libraries
6. creation and operation of two centres of expertise
7. creation and operation of dissemination, education and awareness facilities
8. setting up and operating national coordination point

Appendix B describes a number of projects that implement some of these actions, and mentions potential participants in such projects (based on the expertise and/or data they have)

Appendix B: Infrastructure and Demo-Applications

CLARIN is a research infrastructure initiative that wants to establish a domain of integrated and interoperable services. It can be compared with building an infrastructure for high speed trains that need new tracks, sophisticated signalling and a new generation of operators. Of course, the functioning of such an infrastructure can only be shown when there are these new types of high speed trains. This is the reason why CLARIN from the beginning wants to start developing a few web applications serving two purposes: a) they can be used to derive requirements for the infrastructure and b) they can be used to indicate how typical applications need to be built.

CLARIN-NL wants to follow a two-fold approach: a few applications have been defined in this proposal, together with the respective roles for different partners. In a later phase, further applications will be developed. To this end specific calls for tender will be published. Thus we will be able to benefit from more advanced insights into the design of the infrastructure and the needs in other communities than the linguistic community.

Application 1: Audio/Video Case
A number of audio/video services will be offered that will allow users (inter alia)

- to align a speech signal and a given text even for small languages
- to do automatic indexing given a set of event types and a speech stream resulting in an additional annotation
- to automatically annotate speech data on the basis of a specification of acoustic features
- to transcribe speech automatically
- to translate a video stream into a storyboard of images representing content changes
- to detect predefined visual patterns in a video stream

In this sub-project RU and UT will play a major technology role. BG will be the main content provider, but also Meertens and MPI will carry out tests with their material.

Application 2: HLT Case
Services will be offered that will allow users

- to arrive at an extended annotation which includes tokenization and POS tagging
- to annotate data with morphological information
- to make use of advanced machine learning techniques
- to parse a given utterance of Dutch
- to carry out statistical analyses on textual data

In this sub-project Tilburg and Groningen will play a major role as technology providers. At first instance all centres working on Dutch language variants will be the main content providers and testers. They will integrate all relevant Dutch resources into the infrastructure.

Application 3: Dialect Case
Services will be set up that allow that will allow users

- to upload dialectal material
- to execute HLT operations such as calculating statistics on the material
- to visualize dialectal differences with GIS support

In this sub-project Meertens and Groningen will play a major role as resource and technology providers. Tilburg and UvA may join as technology providers. Part of the task is to integrate all available dialectal resources and tools into the infrastructure.

Application 4: Semantic Case
Services will be offered that will allow users

- to integrate the major concepts found in MPI collections into the DCR framework
- to integrate the TDS definitions into the DCR framework
- to integrate all ontological relations into a registry framework
- to cross-walk between typology databases and the archived resources

Utrecht, UvA and MPI will play a major role as resource providers and semantic integrators. This sub-project will deliver a functional infrastructure for semantic interoperability based on the ISO DCR framework and a simple application showing the types of cross-walk that are possible.

Application 5: Curation and Preservation Case
This sub-project will

- curate and upload all Leiden data into the MPI repository and make them available in the infrastructure
- establish a functioning data exchange based on standards such as OAI PMH and METS between MPI and DANS so that all Dutch data can be archived by DANS as well
- offer the methodology and supporting services to other CLARIN participants as well

The major players are Leiden, DANS and MPI where DANS and MPI offer their archiving capabilities and different access technologies serving different purposes.

Application 6: Literature/History Case
This sub-project, to be defined in more detail, will allow DBNL, Huygens, KB, and possibly others to integrate their data in the infrastructure and to establish show cases in which it is shown how research questions originating from literary studies can be approached with HLT technologies applied to these data.

Application 7: Enrichment Case
This sub-project is meant to establish a resource enrichment scenario that allows authorized users to

- add and store comments on resource fragments
- add and store relations between resource fragments
- to graphically visualize such joint relational spaces and to navigate in them

The major players are INL, Meertens and MPI as resource providers. Existing technology can be used to build the cross-repository relation stores.

Application 8: Quality Case
This sub-project is dedicated to the following functions:

- establish guidelines for long-term preservation and service providers
- carry out self-assessments of service providers along the data seal of approval guidelines
- carry out quality checks on the resources and services

The major partners in this sub-project are DANS and SPEX.

Application/Infrastructure Distribution
In the following description we want to indicate which applications can be seen as test cases for the various typical infrastructure services which CLARIN will provide.

We can distinguish a few generic infrastructure services that will be applied in almost all cases such as

- selecting services based on a joint metadata domain
- identifying resources on the base of persistent IDs
- sign-on based on a distributed authentication service
- certification of all participating servers
- solution for IPR issues

The following infrastructure services are only an issue in a few sub-projects:

- archiving services for resources are addressed in particular in 3 and 5
- format curation and transformation services will be addressed in almost all sub-projects, in particular in 5
- semantic inter-operability services will be addressed in particular in 4
- enrichment services will be addressed in particular in 7
- quality checks for all that is done in CLARIN will be addressed in 8

Appendix C: Budget

CLARIN-NL 2009-2014 **all amounts in M€**	**preparation** **2009_2010**			**construction** **2011_2012**			**exploitation** **2013_2014**			**all periods** **2009_2014**			**exploit per year**
EU-level	*labour*	*other*	*total*	*labour*	*other*	*total*	*labour*	*other*	*total*	*labour*	*other*	*total*	*M€*
Management and coordination	0.06	0.14	0.20	0.34	0.17	0.51	0.30	0.10	0.40	0.69	0.42	1.11	0.20
Technical coordination	0.02	0.07	0.08	0.15	0.07	0.22	0.15	0.07	0.22	0.32	0.21	0.53	0.11
Linguistic coordination	0.02	0.04	0.05	0.15	0.07	0.22	0.15	0.07	0.22	0.32	0.18	0.50	0.11
Outreach coordination	0.02	0.04	0.06	0.15	0.09	0.24	0.15	0.09	0.24	0.32	0.23	0.54	0.12
Internationalisation	0.01	0.03	0.04	0.03	0.04	0.07	0.03	0.04	0.07	0.07	0.11	0.17	0.03
Total EU level in M€	0.12	0.32	0.44	0.82	0.45	1.26	0.78	0.37	1.15	1.72	1.14	2.85	0.58
NL-level	*labour*	*other*	*total*	*labour*	*other*	*total*	*labour*	*other*	*total*	*labour*	*other*	*total*	*M€*
Technical construction	0.90	0.11	1.01	0.90	0.13	1.03	0.15	0.05	0.20	1.95	0.30	2.25	0.10
Data conversion & creation	1.05	0.14	1.19	1.05	0.11	1.16	0.45	0.13	0.58	2.55	0.38	2.93	0.29
Tools and services	1.05	0.14	1.19	1.05	0.11	1.16	0.45	0.11	0.56	2.55	0.37	2.92	0.28
User needs and usage cases	1.05	0.11	1.16	1.05	0.11	1.16	0.33	0.05	0.38	2.43	0.27	2.70	0.19
Advanced LT services	0.30	0.14	0.44	1.20	0.34	1.54	1.20	0.34	1.54	2.70	0.83	3.53	0.77
Expertise centres	0.15	0.07	0.22	0.60	0.18	0.78	0.60	0.18	0.78	1.35	0.44	1.79	0.39
Dissemination and training	0.30	0.20	0.50	0.60	0.25	0.85	0.30	0.25	0.55	1.20	0.69	1.89	0.27
Coordination & management	0.45	0.15	0.60	0.41	0.15	0.56	0.38	0.12	0.49	1.24	0.42	1.65	0.25
Total NL level in M€	5.25	1.07	6.32	6.86	1.39	8.26	3.86	1.23	5.09	15.97	3.70	19.67	2.54
Grand total in M€	5.37	1.39	6.76	7.68	1.84	9.52	4.64	1.61	6.24	17.69	4.84	22.52	3.12

Appendix D: List of Acronyms

Acronym	Expansion (Translated)	URL
Anéla	Dutch Association for Applied Linguistics	http://www.anela.nl/
ANWB	Dutch automobile association	http://www.anwb.nl/
BG	Netherlands Institute for Sound and Vision	http://portal.beeldengeluid.nl/
BIGGrid	Dutch eScience Grid	http://www.biggrid.nl/
BLARK	Basic Language Resource Kit	
CATCH	NWO Programme Continuous Access To Cultural Heritage	http://www.nwo.nl/catch
CELEX	Dutch Centre for Lexical Information	http://www.ru.nl/celex/
CESSDA	Council of European Social Science Data Archives	http://www.nsd.uib.no/cessda/home.html
CHIP	Project in the CATCH programme	http://www.nwo.nl/catch/chip
CHOICE	Project in the CATCH programme	http://www.nwo.nl/catch/choice
CLARIN	The CLARIN infrastructure	http://www.clarin.org/
CLARIN-EU	The European CLARIN enterprise	http://www.clarin.org/
CLARIN-NL	The Netherlands national project for CLARIN	http://www-sk.let.uu.nl/pc/clarin_nl/
CLARIN-prep	The currently running European CLARIN preparatory phase project	http://www.clarin.org/
CLCG	Center for Language and Cognition Groningen	http://www.rug.nl/let/onderzoek/onderzoekinstituten/clcg/
CLIN	Computational Linguistics in the Netherlands (local conference)	http://www.let.rug.nl/~vannoord/Clin/
CLS	Centre for Language Studies	http://www.ru.nl/cls/
CLST	Centre for Language and Speech Technology	http://www.ru.nl/clst/
COGIS	Dutch expert centre on the (psycho)social effects of war, persecution, aggression and violence	http://www.cogis.nl/Index.aspx
COREX	Corpus Exploitation Tool	http://lands.let.kun.nl/cgn/doc_English/topics/corex/info.htm
CPG	Centre for Parliamentary History	http://www.ru.nl/cpg/
DAM-LR	Distributed Access Management for Language Resources	http://www.mpi.nl/DAM-LR/
DANS	Data Archiving and Networking Services	http://www.dans.knaw.nl/
DARIAH	Digital Research Infrastructure for the Arts and Humanities	http://www.dariah.eu/
DBNL	Digital Library for Dutch Literature	http://www.dbnl.org/
DCR	Data Category Registry	
DoBeS	Documentation of Endangered Languages	http://www.mpi.nl/DOBES/
EAGLES	Expert Advisory Group on Language Engineering Standards	http://www.ilc.cnr.it/EAGLES96/edintro/node1.html
ELAN	EUDICO Linguistic Annotator (Annotation Tool)	http://www.let.ru.nl/sign-lang/ECHO/ELAN/ELAN_intro.html
ERC	European Research Council	http://erc.europa.eu/
ESF EURYI	European Science Foundation European Young Investigator Awards	http://www.esf.org/activities/euryi.html
ESFRI	European Strategy Forum on Research Infrastructures	http://cordis.europa.eu/esfri/

FIDLR	Framework for the Interoperability of Dutch Language Resources	www.meertens.knaw.nl/fidlr
FIDLR-Start	Project to prepare the FIDLR project proposal	www.meertens.knaw.nl/fidlr
FlareNet	Fostering Language Resources Network, European Network project.	http://www.ilc.cnr.it/flarenet/
FP7	Framework Programme 7 (European Union)	http://cordis.europa.eu/fp7/
GIS	Geographic Information System	
HLT	Human Language Technologies	
HSS	Humanities and Social Sciences	
IAP	International Advisory Panel	
ICT	Information and Communication Technology	
ID	Identifier	
IIAV	International Information Centre and Archive for the Women's Movement	http://www.iiav.nl/
IISG	International Institute for Social History	http://www.iisg.nl/
IMDI	ISLE MetaData Initiative	http://www.mpi.nl/IMDI/
IMIX	NWO programme for Interactive Multimodal Information eXtraction	http://www.nwo.nl/imix
INL	Institute for Dutch Lexicology	http://www.inl.nl/
INTERA	Integrated European language data Repository Area	http://www.mpi.nl/INTERA/
Interop	NSF Community-Based Data Interoperability Networks Programme	http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=502112
IOP	Innovation-targeted Research Programmes	http://www.senternovem.nl/iop/
IPR	Intellectual Property Rights	
ISLA	Intelligent Systems Lab Amsterdam	http://www.science.uva.nl/research/isla/
ISLE	International Standards for Language Engineering	http://www.ilc.cnr.it/EAGLES/isle/ISLE_Home_Page.htm
ISO	International Standards Organisation	http://www.iso.org/
ISO TC37/SC4	ISO Subcommittee focusing on Language Resource Management	http://www.tc37sc4.org/
ISOcat	Project to implement the ISO DCR standard	http://www.isocat.org/
KB	National Library of the Netherlands	http://www.kb.nl/index-en.html
KDC	Catholic Documentation Centre	http://www.ru.nl/kdc/
KNAW	Royal Netherlands Academy of Arts and Sciences	http://www.knaw.nl/
LIRICS	Linguistic Infrastructure for Interoperable Resources and Systems	http://lirics.loria.fr/
LMF	Lexical Markup Framework	http://www.lexicalmarkupframework.org/
LOT	Netherlands Graduate School of Linguistics	http://wwwlot.let.uu.nl/
LRT	Language Resources and Technologies	
LT	Language Technology	
METS	Metadata Encoding and Transmission Standard	http://www.loc.gov/standards/mets/
MPG	Max-Planck-Gesellschaft	http://www.mpg.de/
MPI	Max Planck Institute for Psycholinguistics	http://www.mpi.nl/

MuSeUM	Project in the CATCH programme	http://www.nwo.nl/catch/museum
NIOD	Netherlands Institute for War Documentation	http://www.niod.nl/
NWO	Netherlands Organisation for Scientific Research	http://www.nwo.nl/nwohome.nsf/pages/SPPD_5R2QE7_Eng
NWO-Groot	NWO Infrastructure Investment Fund	http://www.nwo.nl/nwohome.nsf/pages/NWOA_4YJD5U
OAI	Open Archives Initiative	http://www.openarchives.org/
OLAC	Open Languages Archive Community	http://www.language-archives.org/
PC	Programme Committee	
PCM	large publisher in the Netherlands (Perscombinatie Meulenhoff & Co)	http://www.pcmuitgevers.nl
PID	Persistent Identifier	
PMH	Protocol for Metadata Harvesting	http://www.openarchives.org/OAI/openarchivesprotocol.html
POS	Part of Speech	
R&D	Research and Development	
RU	Radboud University Nijmegen	http://www.ru.nl/
SAMPA	Speech Assessment Methods Phonetic Alphabet	http://www.phon.ucl.ac.uk/home/sampa/
SARA	Foundation Academic Computing Centre Amsterdam	http://www.sara.nl/
SO	Support Office	
SOA	Service Oriented Architecture	
SPEX	Speech Processing Expertise Centre (RU, Nijmegen)	http://www.spex.nl/
SPIN	Research programme	
STEVIN	Dutch-Flemish programme for realizing the BLARK and HLT research for Dutch	http://taalunieversum.org/taal/technologie/stevin/
STITCH	Project in the CATCH programme	http://www.nwo.nl/catch/stitch
SURFnet	University Network Services Organisation	http://www.surfnet.nl/
TDS	Typological Database System	http://languagelink.let.uu.nl/tds/index.html
TEI	Text Encoding Initiative	http://www.tei-c.org/
TERENA	Trans-European Research and Education Networking Association	http://www.terena.org/
UIL-OTS	Utrecht Institute of Linguistics OTS	http://www-uilots.let.uu.nl/
UT	University of Twente	http://www.utwente.nl/
UvA	University of Amsterdam	http://www.english.uva.nl/
VSNU	Association of Collaborating Dutch Universities	http://www.vsnu.nl/
VU	VU University of Amsterdam	http://www.vu.nl/
VVV	Touristic Office	http://www.vvv.nl/
WAYF	Where Are You From	http://www.switch.ch/aai/support/tools/wayf.html