CLARINO

Coordinator: University of Bergen, Prof. Koenraad De Smedt

Research Council of Norway – Research Infrastructure

Participation in international research infrastructure (ESFRI construction phase)

Project description, March 20, 2012

1 Relevance

1.1 Relevance in ESFRI

CLARIN[1] is a European infrastructure initiative on common language resources and technologies. It was listed as a mature proposal on the ESFRI 2006 roadmap and received an EC grant for the preparatory phase from January 2008 to June 2011, during which the University of Bergen has been a consortium member. CLARIN has now entered its construction and exploitation phase. For this purpose a European Research Infrastructure Consortium (ERIC) is being established where a consortium of countries will make long-term commitments towards construction and operation of the CLARIN infrastructure.

The following three layers are distinguished in CLARIN:

1. The overall **governance and coordination** layer is the responsibility of the ERIC, with a General Assembly as the main decision making body, a Board of Directors responsible for the execution of the strategies, policies and work programme decided by the General Assembly.

   CLARINO intends to fully participate in CLARIN. It implies a request to the Norwegian government to consider joining the CLARIN ERIC.

2. The **infrastructure** layer is fully financed by the participating countries, with no funding coming through the ERIC. Each member country will commit itself to creating a national consortium and will specify which centres it will establish and which infrastructure services these will offer in the CLARIN context.

   CLARINO presently proposes such an infrastructure layer with a national consortium, a centre structure and a specification of initial services. It will link to other centres in the European CLARIN network.

3. The **content** layer is the responsibility at the national and institutional levels, mainly through projects which generate suitable primary and secondary data by data collection (or digitization/restoration) and annotation, based on well founded methods validated in research projects.

   CLARINO does not apply for funding the creation of new content, but includes the upgrading and cataloguing of existing content as well as integration of content currently being collected and generated by other projects (e.g. INESS, NorDiaSyn, Menotec and the The Language Technology Resource Collection for Norwegian – Språkbanken).

1.2 Relevance to Science and Society

The Humanities must use new eScience approaches to language in order to address theoretical questions as well as societal challenges in a new way. Some fundamental questions in the sciences of language relate to the universality of human language, the question of how our conception of the world is structured in linguistic terms, a more detailed understanding of language change and more refined models of language processing and learning by humans and computers. Many of these questions are not only relevant to a deeper understanding of the human mind, but also to solving practical problems of communication in today's knowledge society.

[1] http://www.clarin.eu; deliverables at http://www.clarin.eu/deliverables

Furthermore, questions in literary, historical, cultural and philosophical scholarship are also dependent on an understanding of language and on the analysis of the content and structure of texts, such as studies related to the diffusion of ideas and cultural trends, from letters between scholars in the Middle Ages to blogging today. Scholarship dealing with editions of large literary works or archives are in need of more powerful tools for filtering, comparison and visualization of different interpretations and editions.

New data-intensive computational methods have the potential to further our knowledge in language-related areas. Imagine a research project seeking to contrastively investigate sociolinguistic properties and attitudes in a specific genre or domain, say climate related publications, or imagine a study on public opinion in the political sciences, based on mining large quantities of newspaper text, or an investigation of the so-called blogosphere in the media sciences. To establish a foundation for such work, a project might identify language samples from digital materials available from existing corpora, or harvested from the web, or from other sources. A dependable infrastructure is needed for collecting, accessing and analysing such materials, which may be large amounts of language data spanning several languages and archived in different countries.

Language data can be of a very diverse nature, such as digital and digitized writings, sound and video recordings with transcriptions, dictionaries and concordances, historical archives, experimental data from EEG, fMRI or reaction time measurements, and many others. These materials may be encoded in particular ways and carry annotations at different levels. This heterogeneous landscape and its increase in magnitude has drastically transformed the requirements for their archiving, publication, exploration and long-term maintenance.

These requirements are at present not met by satisfactory solutions. On the contrary, many potentially useful collections of language materials in Norway and elsewhere are either acutely endangered or lost, or their reuse is severely impaired by technical and organizational obstacles. While other projects (at the content layer of the infrastructure) take responsibility for recovery and digitization efforts, it is presently important to prevent similar problems in the future by better curation of research results and promoting their reuse in further research or in practical applications.

An eScience approach in the Humanities disciplines (eHumanities) is necessary to close the gap between, on the one hand, the increasing demands on theory development and practical applications, and on the other hand, the deluge of information in need of analysis. An infrastructure with national and international coordination and supported by sound scientific and technical approaches, must be put in place to store, preserve, describe and make publicly available this huge volume of data in an open, user-friendly and trusted way.

Language infrastructures represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architecture.

1.3 Political Relevance

The Norwegian Parliamentary White Paper no. 35 (2007–2008; adopted by the Parliament on April 28, 2010) clearly places language technology on the national research agenda as an important domain for language policy, by encouraging the participation of Norwegian language users in cultural, socio-economic, and political communication on new technology platforms. Furthermore, cultural, economic and political relations between Norwegian-speaking and other linguistic communities require extensive translingual and multilingual approaches to information, communication and humanities scholarship, both within Norway and in an international context.

2 Vision and Scientific Goals

The purpose of CLARINO is to realize Norwegian participation in CLARIN by establishing a Norwegian network of centres offering infrastructure services. CLARIN aims at overcoming the current fragmentation by establishing a more integrated Europe-wide infrastructure of language resources and technologies enabling eHumanities. Its characteristics are the following:

- integrated: the resource and service centres are connected via Grid technology and form a virtually integrated domain;
- interoperable: the resources and services will use web services to overcome format, structure and terminological differences;
- stable: the resources and services are offered with a high availability;
- persistent: the resources and services are planned to be accessible permanently so that researchers can rely on them;
- accessible: the resources and services are easily searchable and accessible in forms tailored to the needs of user communities;
- extendable: the infrastructure is open so that new resources and services can be added easily.

The expected impact of the CLARIN infrastructure is a much improved accessibility of data and tools for eHumanities through consolidation and harmonization. The potential of CLARIN may be comparable to that of the introduction of libraries: CLARIN intends to do for the digital word what libraries have done for the printed word. Norwegian participation in this European effort will enable Norwegian scientists to gain access to a pan-European infrastructure as well as to preserve and publish Norwegian research data, thereby improving the visibility as well as the capability of Norwegian research, especially in the Humanities.

Furthermore, CLARIN has written a cooperation agreement with META-NET, whereby the latter will promote the sharing of language data by industrial actors for the purpose of developing applications for communication in the information society. The University of Bergen is presently participating in META-NET through the META-NORD project. The sharing of language data will positively affect the language technology marketplace for developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially SMEs.

3 Scientific and Technological Status

Although Norway has a limited research environment in terms of budget and number of researchers, the state of the art in Norway is on a high international level and the conditions for participating in CLARIN are presently met. The University of Bergen (UiB), the University of Oslo (UiO) and the Norwegian University of Science and Technology (NTNU) have more than four decades of experience in working with digital annotated language data for Norwegian and other languages. They have earned a reputation for their collections of corpora, digital text archives and other digital Humanities resources. Also the University of Tromsø (UiT), the Norwegian School of Economics (NHH) and other institutions in Norway have a number of collections and tools. During the preparatory phase of CLARIN (supported at a national level by RCN), the University of Bergen has compiled a preliminary overview of more than 150 existing language resources in Norway.

The National Library of Norway (NB) is in the process of digitising its entire collection, but does not presently have any significant amount of annotated data; however, initiated by parliamentary decision, NB was in the budget bill for 2010 commissioned to establish the Language Technology Resource Collection for Norwegian – Språkbanken. This collection intends to be an important asset for Norwegian R&D and its materials will be accessible also in CLARINO.

Examples of annotated primary language data include a wide range of monolingual and multilingual language corpora, annotated literary archives, historical archives, audio and video recordings of language learners and users, etc. Secondary language data include, for instance, electronic dictionaries, termbanks, named entity lists, etc. Language tools include taggers, parsers, concordancers, as

well as systems for machine translation, automatic summarization, author attribution, speech recognition and production, corpus management and search, text encoding, etc. It is important to note that despite a natural focus on the Norwegian language, there has also been important work in Norway on several other languages (including not only the minority languages Sami and Kven but also many languages around the world), on multilingual resources and on language-independent tools.

Paradoxically, the early start of Norwegian digital language research has left the country lagging behind, since many of the digital language data produced over the past half century are poorly accessible, endangered or already unrecoverable. Secure data storage, archiving, and providing for varied reuse scenarios are typically beyond the scope of individual, short term research projects producing new language data. An additional problem is that the size and complexity of datasets is sharply increasing so that massive storage and sophisticated analysis tools are needed, requiring access to High Performance Computing. The CLARINO infrastructure will provide services to overcome these problems by means of archiving, long-term secure storage, high performance computing (HPC) and support for better access, exploration and reuse.

Some issues can be illustrated in the case of minority languages. There are three different Sami languages spoken in Norway: North, Lule and South Sami. At present, the amount of text electronically available for the three languages adds up to less than ten million words, collected at UiT. The text corpus, mostly unpublished material, is fragmented into small files, and much of it is encoded in some pre-Unicode format; manual improvement of the metadata is a prerequisite for an efficient exploitation of the corpus. The small amount of text, especially for Lule and South Sami, makes it necessary to collect in principle all existing text. For Sami, the digitization of NB means an increase of the available material, but access to an infrastructure with better LRT services to automatically analyse and annotate the material is becoming an acute need. Additionally, there is a general problem in accessing library material due to privacy protection and IPR.

With respect to terminology, there is at present no national infrastructure, despite the fact that an increasing number of scholarly domains face the threat of domain loss. This effort will take an important step towards overcoming the fragmentation problem in the terminology field and the need for coordination of terminological efforts at international and national levels. In CLARINO, NHH wants to establish a national infrastructure to harmonize terminological language resources and technologies, in response to the national responsibility now given to higher education to develop and disseminate Norwegian terminology.[2] This will be a natural extension of already existing resources originating from the Norwegian Term Bank which in recent years have been further developed in projects such as KB-N, Mikroøkonomen, Termportalen and CLARA. It is not the aim to develop new terminologies per se, but to provide the technical architecture needed for their integration and accessibility, and to vouch for their interoperability with the CLARIN and ISO/TBX standards.

As for scholarly editions of digitized archives with literary, historical or philosophical importance, a large number of XML-TEI-conformant texts already exist. National examples include the Wittgenstein Archives Bergen (WAB), the Medieval Nordic Text Archive (MeNoTa), the Ibsen text archive, the Nordic Holberg project; international examples include the resources published in the international Philosource federation, the Austrian Brentano project, the French-Polish ELV project, the German Leibniz edition, and many others. However, without appropriate access, filtering and visualization tools, these archives will not be fully and adequately exploited; if the needs of different users (e.g. scholars of literature vs. philosophers) are not catered to, the investment in digitization and encoding may have been largely wasted.

As part of gaining access to the content of digital language materials, tools for linguistic analysis are playing an increasingly important role. Computational language analysis is the automated process of exposing the underlying linguistic structure in observed samples of human language, be it written or spoken, with the ultimate aim to 'make sense' of human language. Many tools developed in the past have, however, been constructed for specific R&D purposes, or are limited to certain

[2] Cf. Act relating to Norwegian higher education §1-7.

platforms, data encodings or size of data sets, have been insufficiently documented, cannot be linked to other tools, and have a high treshold for deployment by targeted users. There is a need for better support and improved interoperability of tools and their compatibility with datasets through a grid-based web services architecture. This is becoming feasible today, since current language technology has matured to a point that enables large-scale eHumanities applications for Norwegian, English, and maybe other languages. Key motivations for this part of the infrastructure are (a) to lower the barrier to entry for the utilization of existing language technologies, (b) to establish a Web-accessible demonstration centre for a range of language analysis tools, and (c) to connect language technology providers and users alike to the emerging national high-performance computing and storage infrastructure (HPC)—particularly focusing on use patterns in the Humanities and Social Sciences.

Current efforts, also in Norway, are partly addressing some of the above issues. The Language Technology Resource Collection for Norwegian – Språkbanken, which is an integrated part of and established at the National Library of Norway, aims to collect, further develop and disseminate language resources for use by research and industry. It primarily targets the Norwegian language (and to some extent, Sami and multilingual resources) and will be integrated in the CLARINO registry, but it does not intend to cover all eHumanities needs. Furthermore, the INESS, NorDiaSyn and Menotec projects, financed by RCN, are building specialized databases with eHumanities facilities, exemplifying the CLARINO approach. The NorDiaSyn results have already proved their national and international importance, and the Nordic Dialect Corpus can already be accessed via a pipeline from a CLARIN-project in Amsterdam (Edisyn). INESS and Menotec are running infrastructure projects at the content layer and will provide data which will be accessible in CLARINO. Finally, university libraries are increasingly involved in digital archiving and dissemination of research data. The Bergen University Library is already participating in activities such as the Holberg project, Grind.no, Faghistorisk dokumentasjonsprosjekt etc. and is in a dialog with the Faculty of Humanities to take over some research archives.

4 Description of the Research Infrastructure

The CLARINO infrastructure is new in its presently proposed scope and organization, particularly in relation to the European CLARIN project, but will build mostly on existing collections and well proven tools. The infrastructure will adhere to guidelines and standards by CLARIN related to technologies (persistent identifiers, identity federations, repositories, grid technologies), metadata encoding, and legal aspects (IPR, privacy).

4.1 Centres

The construction of CLARIN is planned as a distributed infrastructure based on a network of centres offering data and technology services. Centres are of type A, B, C or R depending on the level of services and integration in the European infrastructure.[3] CLARINO is a distributed effort with a clear division of national responsibilities. We propose the following Norwegian main centres in CLARINO which are committing themselves to long-term operation of the infrastructure:

A. The different roles of type A centres will not be fulfilled by a single organization, but by a cooperation between the following centres:

   1. **NB.** NBs primary task will be to run a national metadata registry, harvesting national providers, providing an OAI PMH gateway for metadata exchange with other national CLARIN nodes in Europe, and providing a repository for very long-term archiving and curation. It will therefore fulfill an important role of a type A centre.

[3] The centre types are described in CLARIN Deliverable D2R-1b, centre formation in D1R-1a.

2. **UNINETT.** UNINETT will through Feide offer authentication and authorization services to the CLARIN AAI federation. It will through NOTUR, NorStore and future eInfrastructure offer access to national HPC and storage facilities. These are important type A centre roles.[4]

B. Type B service centres are part of the CLARIN AAI federation but still act as individual centres not taking over responsibilities for the federation. Their resources will be maintained and made accessible by appropriate interfaces in a well-structured repository system with a long-term commitment and support for metadata harvesting and tool integration.

1. **The Text Laboratory (UiO).** In cooperation with USIT (UiO), this centre will provide services for a wide variety of corpus data, tools and derived language resources.
2. **EDD, the Unit for Electronic Documentation (UiO).** This centre, operating in cooperation with the Dept. of Philosophy at UiB and with the projects MENOTEC/MeNoTa and Norsk Ordbok 2014, will primarily provide services for digital editions of literary and historical texts and dictionaries, and tools for their construction.
3. **The Bergen Centre.** UiB, in cooperation with NHH and Uni Research, will provide language data resources services with special national responsibility for treebanks and terminology resources. The Research Group on Language Models and Resources (LaMoRe) will provide scientific coordination at the Bergen centre. This centre will be technically anchored at the Bergen University Library (UBB) which will define and test its role as an institution-wide repository for research data. It has newly created a Section for Digital Systems and Services which will participate in CLARINO.
4. **LAP, the Language Analysis Portal, located at IFI (UiO).** In cooperation with USIT (UiO) and the national facilities for HPC (represented by UNINETT), this centre will primarily provide language technology services.

C. Type C centres will be NTNU and UiT. Type C and R centres will be an open set providing data, tools and preferably a base address for OAI PMH or XML based metadata harvesting to the other centres.

The CLARINO centres will together provide the services described in detail in the following sections.

4.2 Infrastructure Services

a. **A National Language Data Registry.** There is currently no national (let alone international) registry of language resources, so that it is impossible to ask, for instance, where to find corpora that include speech by teenagers between 13 and 16 years of age. CLARINO will have a national data registry facility providing an OAI PMH gateway for central harvesting from all nodes in the Norwegian network and a system for metadata exchange (synchronization) with other national CLARIN registries, so as to offer an up to date Europe-wide catalog of language resources. A dedicated interface for access to the national CLARIN metadata registry will be provided at NB. Metadata for existing and new materials, including also licensing conditions, will be compatible with the inventory requirements by CLARIN and META-NORD.

b. **Language Resources Preservation Services.** The need for data archiving and curation depends on long-term storage with continuous technology migration. Such an infrastructure is present at NB, which has facilities for secure triply redundant storage and regular copying schedules for its

[4]The involvement of UNINETT will also contribute to avoiding parallel setups for eInfrastructures. However, the communication between European infrastructures needs to be organized mainly on the European level. The DASISH project (Data Service Infrastructure for the Social Sciences and Humanities) represents an important step in this direction. CLARINO will cooperation with DASISH.

entire collection. NB will use similar storage solutions for The Language Technology Resource Collection for Norwegian – Språkbanken, thereby avoiding duplication of effort. A repository with metadata and PIDs will be set up.

c. **Online Language Data Storage and HPC access.** Whereas NB will store static data, there is also a need for for dynamically accessed data that is indexed or structured with database management tools for online access, e.g. interacting with the LAP (see below) and connected to High Performance Computing facilities for efficient search and processing. Some data may be stored locally, but larger databases will be stored at national storage facilities (NorStore).[5]

d. **Trusted Authentication and Authorization.** LRT providers in Norway have so far implemented local login mechanisms which have resulted in a multitude of incompatible usernames. CLARIN presupposes that repository systems will interact with a Shibboleth resource provider instance to participate in the distributed CLARIN AAI federation. Since Norway has a good national identity provider federation (Feide) and agreements with other federations (e.g. Kalmar) these national solutions will be implemented at the CLARINO centres with support from UNINETT. Furthermore, authorization based on licensing conditions metadata will be matched to trusted user identification to support authorized access to resources. A national PID service will be run.

4.3 Language Data Resource Services

Data access and management will be a core component of the infrastructure services. However, the heterogeneity of language data implies that their exploration cannot be supported by a single platform. Multiple systems depending on the data, the discipline, and system functionality will be put in place, mostly based on existing, successfully tested solutions. Some existing infrastructure initiatives, e.g. the INESS, NorDiaSyn and MENOTEC infrastructure projects come with their own exploration services, developed and financed within these respective projects. They will be integrated into the CLARINO network. Other on-line services for exploring data in CLARINO are described in the following list. These do not, however, form a closed set of services with fixed specifications, since new needs will certainly continue to arise for new data types and scholarly purposes.

a. **An Electronic Editions Platform.** This platform will enable non-technical philologists to work with digital editions of literary and historic documents. Such documents often need to be inspected at various levels of interpretation, such as facsimile, multiple editions and scholarly annotations, which creates a need for comparisons between editions, visualization of specific textual elements, etc. The Interactive Dynamic Presentation (IDP) system, which has been tested at the Wittgenstein Archive Bergen (WAB), will be generalized and will provide important functionality for TEI-conformant texts. Its user steered, customizable presentation of textual materials marks a new approach to the understanding of scholarly editing by engaging the user in filtering and visualization of source text characteristics and their metadata. Through IDP it is possible, e.g., to transform textual data from an item of Wittgenstein's Nachlass into a customized e-book tailored edition in a specific format (for example diplomatic), with a specific sequence of its text parts (e.g. chronological or in the order of the original manuscript), with particular items (e.g. names) highlighted, with links to other sources Wittgenstein makes reference to, and with notes by other scholars. The system will offer search without the necessity to know the exact codes of the stored elements and attributes, so that non-specialists will be able to search metadata, full text, and content structure, and present the results in user tailored ways, including IDP. Other services, e.g. dictionary services, will also be included. The Menotec project will from its own funding provide content and tools to integrate the Menotec infrastructure into this platform. This platform will be offered at EDD and the IDP will be developed in cooperation with UiB and Uni Research.

[5]The UNINETT Sigma director responsible for NorStore welcomes an application from CLARINO for this purpose. The actual size of the resources (cycles, bytes) needed by CLARINO is probably small compared to other communities (e.g., life sciences and climate), though the language community will require the necessary user support.

b. **Glossa.** This advanced search and results management system for the exploration of linguistic corpora interacts with the Open Corpus Workbench system and is in use worldwide for a wide range of written and spoken, multimedia (with links to audio and video), monolingual, multilingual and parallel corpora. It focuses on user friendliness and flexible interaction with other tools, including statistics and innovative integration with Google Maps and automatic translation. Its inclusion as middleware in the CLARINO infrastructure lies in its comprehensive range of services offered to corpus researchers. Glossa will be expanded from the current model where corpora and other language resources are managed by the Text Laboratory (UiO) to a service where users freely build and expand corpora using loosely integrated CLARIN resources to enrich the data. Users will be able to provide their own language material and potentially their own tools for text analysis or management (tokenizers, aligners etc) and share the resource they have built with other researchers, transfer it to other CLARIN repositories or dynamically expand it with further data or analysis. The system will provide fully automated import of text material, corpus building and running of integrated analysis tools, in addition to mechanisms for expanding the corpus material after it has been built. Furthermore, the user interface, which is currently configured manually for each corpus, will be built automatically from corpus meta data. Collaborative functionality such as access control, various channels of user communication and search functionality among the shared resources will be provided. The backend storage and indexing will be independent of the rest of the system so that Corpus Workbench can be replaced with other backends such as ANNIS2 or Corpuscle (see below). Glossa will be hosted at the University of Oslo.

c. **Corpuscle.** This corpus management tool solves some of the fundamental limitations in the current generation of systems (Open Corpus Workbench, Manatee etc.). It is operational today but is being further developed in relation to user needs. Corpuscle's query engine with innovative algorithms based on suffix arrays and pre-filtering results in query execution speed comparable to or better than Corpus Workbench for most types of queries, in particular those with high-frequent initial position or involving certain types of regular expressions. It also has a powerful query syntax, extending that of Corpus Workbench (CQI), supplemented with menu-driven query building. Furthermore, it features seamless integration of manual corpus annotation and editing, with live querying, at least for smaller corpora, backed by a relational database. Several other features set Corpuscle apart from other corpus systems, among them its support for hierarchic data and the integration of manual editing and annotation. This makes Corpuscle very well suited for special-purpose research corpora that need extra functionality (as proven by the adaptations made for ASK, the Dialektendring corpus and the Norwegian Newspaper Corpus). The corpus management system and web interface is developed as a single seamless program, but the Corpuscle backend will also be compatible with the Glossa frontend. Corpuscle will be hosted at UiB and will be developed in cooperation with Uni Research.

d. **Terminology Services ('CLARINTerm').** CLARINO will include a technical architecture that integrates a variety of structured mono- and multilingual terminology bases via a common interface. The infrastructure will provide public and controlled unified access to distributed resources in a national grid-based architecture, with links to European terminology resources via CLARIN. Both distributed and centralised solutions will be offered. CLARINTerm will provide access to autonomous bases with a common interface and search system, as well as integrate a variety of existing termbases into a central knowledge base that contains entries for concepts and conceptual descriptions such as main terms and synonyms, definitions and concept relations. It will be the decision of each data provider to either have the data fully integrated into the central knowledge base or made accessible as a distributed base. Moreover, hybrid solutions may be offered, in which an instantiation of CLARINTerm draws from both a local source and the central knowledge base, but is presented uniformly as one resource.

e. **Treebanking Services.** Treebanks are language data with highly complex hierarchical anno-

tation. A national (and international) platform for treebanking services will be provided by the INESS project with its own funding under an existing RCN grant (cf. INESS project description at RCN). The INESS platform for treebanking services will be integrated in CLARINO.

4.4 Language Technology Services

The CLARINO infrastructure will offer custom services for the analysis of language data in an eScience spirit. At present, tools for processing complex language data are often narrow in scope, language-, genre-, domain-, and platform-dependent, or designed primarily for expert users. If available at all, most existing technologies require users to download and install locally. These obstacles will be overcome by a portal offering easy and flexible access to a wide range of language technologies, enabling scholars without in-depth technological expertise to perform language analysis and thus to conduct effective research on a very large scale.

The CLARINO Language Analysis Portal (LAP) will provide web services for both interactive and batch processing, e.g. as plugins for language data resource services (such as Glossa or Corpuscle), providing on-the-fly annotations for short examples, or through large-scale asynchronous processing for larger data collections. The LAP will stream language data through a pipeline of user-selected analysis tools. Such analysis might comprise a number of steps, including extraction and segmentation of textual content, tokenization, utterance-level analysis, discourse analysis and aggregation of results.

A user interface combines access to common, pre-existing language resources, web harvesting facilities, and the upload of user-provided data. It includes flexible access to a set of tools for content extraction and layout analysis (from common file formats), as well as a comprehensive repository of language analysis tools—ranging from token- to discourse-level processing, depending on available technology for individual languages, to be collected in CLARIN and META-NET.

Explicit knowledge about interdependencies among component tools, standardized interchange formats and converters, predefined common analysis pipelines, and a flexible web-based configuration manager will make it possible for users to experiment with composite analysis workflows fitted to individual requirements. In practice, users will be able to connect tools and data which previously were incompatible in format or structure.

Through a web browser GUI, the resources and tools involved will be configured and submitted to the national grid infrastructure, where HPC resources for computation and storage are readily available on a scale traditionally inaccessible to users in the Humanities.

4.5 Critical factors

The CLARINO centres will depend on good cooperation with other centres in Norway as well as those in the participating European countries. At international level, it is expected that the ERIC will play a coordinating role, as long as the participating governments are willing to make long-term national commitments. At national level, the centres will have interdependencies, in the sense that the services provided by one centre may depend on access to resources from another. A strong national management and provisions in the consortium agreement to secure interoperability of centres are therefore critical.

CLARINO will apply for access to national grid computing and national storage (NorStore) using regular procedures. Continued operation of the CLARINO services will depend on the continuation (and improvement) of the national eInfrastructure.[6]

Acceptance of the CLARINO infrastructure by end users is dependent on the service level offered and training provided. For that reason, the CLARINO plan pays attention to user interface design, functionality of the infrastructure in relation to common user needs, documentation and user support. Training in the construction and use language resources is a necessary component to increase

[6]Cf. the report *The scientific case for eInfrastructure in Norway* by the eInfrastructure Scientific Opportunities Panel established by the eVITA Programme Committee.

the value of the infrastructure for future research. Norway participates in such researcher training through CLARA (a Marie Curie ITN), but wider training will be needed through curriculum changes at the bachelors and masters level as well. CLARINO will organize researcher training tutorials at its annual meetings, but relies on cooperations with local, national and international PhD schools in the long run.[7]

5 Impact on Science, Technology and Innovation

Participation in CLARIN will enable eHumanities in Norway, offering language data access and analysis to the Humanities disciplines, thereby causing a broad paradigm shift towards large scale empirical methods which have previously been accessible only to very few researchers in Norway. Furthermore, participation in the CLARIN AAI will provide Norwegian researchers with direct access to thousands of research collections in other European participating countries, and vice versa, thus promoting internationalization in a big way. Advanced access to language resources will also attract new researchers to the field, both from Norway and abroad, as demonstrated earlier.[8]

CLARINO services will boost the output of research projects in terms of quality and quantity of analysis. Linguistic analyses can build on the language data resource services or the language technology services offered by CLARINO, depending on the concrete needs.

IDP will increase among users the awareness that the texts they work with and base their research upon, are the results of scholarly processes and hence open to methodological criticism; it will thus generally increase critical awareness of one's choices of method and approach. It will also lead to new research questions and new ways of following up long standing research questions, since it will provide hereto unknown tools to filter and present different textual aspects. On a more general basis, it will significantly help to migrate the humanities more fully into the digital medium, since IDP based user customizable research will only be possible in the digital realm. Demonstrating the value of text encoding, it will encourage humanities projects to strongly consider text encoding as a relevant option for recording and making available their data. On a general level, it will boost the use of digital humanities resources since it will allow device tailored use.

For terminology, The ambition of CLARINO is to provide the technical solutions needed for the overdue coordination, integration and further development of terminology nation-wide. This is seen as relevant for national bodies like the Language Council of Norway and the Norwegian Association of Higher Education Institutions (UHR), which will be given the necessary technical tools required for their stated goals of a coordinated national effort in terminology, as recently expressed in governmental white papers. CLARINO will be a direct response to the recent proposal by UHR to initiate such a national terminological infrastructure.[9] Finally, the infrastructure is needed for what is also a strategic objective of academic institutions such as NHH and UiB, which have recently established their institutional language policy documents advocating the use of parallel terminology in Norwegian and English and taking responsibility for specific domains. By offering a comprehensive and consensus-based theoretical and methodological framework for conceptual structuring, definition writing and delineation of domains and subdomains, the effort will streamline and unify terminology efforts across fragmented teams of developers.

[7] *"The CLARIN infrastructure will make it easier for starting researchers to get access to national and international language data and analysis tools as a basis for new research. When this is becoming a reality, we foresee that training in using language resources and tools must become more integrated in researcher training programs in the years to come."* — Dr. Kari Haugland and Prof. Øivin Andersen, directors of the PhD Research School in Linguistics and Philosophy.

[8] The *Transnational Access to European Research Infrastructure Wittgenstein Archives at the University of Bergen* attracted 35 projects by visting researchers from 12 countries over only two and a half years.

[9] Cf. the UHR board meeting of Aug. 25, 2010.

6 User Groups

The CLARINO infrastructure will serve all project participants and the larger European CLARIN academic community, as well as other national and international language data users, content providers and technology developers. It is expected that in the foreseeable future, the infrastructure will be virtually 'on the desk' of every scholar dealing with annotated or structured language data. Thus, CLARIN will in the long run have the same enabling function as national and university libraries, with which the project intends to cooperate. Eventually some parts of the infrastructure will even have the potential of reaching high school students and transform high school language curricula by using digital technology for more than cosmetic changes to teaching methods.

Language data resource services will cater to a variety of user groups. For end users of text archives, corpora, termbases etc., from field experts via translators and interpreters to the general public, CLARINO will provide unified access to a wide range of updated data, hence eliminating the need to check a variety of web locations in the search for specialist field knowledge. Furthermore, the proposed system will allow owners of digital documents, corpora, dictionaries, term lists and term bases (such as translation bureaus) to upload their data in simple formats and have it converted via dynamic web-based tools to exchangeable formats and in accordance with existing industrial standards. This, in turn, will allow for distributed editing, controlled accessibility and integration with other sources, again at the request of the data owner.

IDP access to archives at EDD will be directly relevant to any scholar who is interested in utilizing the full spectrum of data and metadata contained in an XML-TEI encoded text source. A great number of such electronic texts are already available, and an increasing number are being prepared.

CLARINO will address large-scale developers of structured termbases, such as the Norwegian Ministry of Foreign Affairs, to disseminate their resources and have the terminology data made accessible nationally via CLARINO with a transparent international throughput to CLARIN. Moreover, developers of termbase technology and standards, such as Standards Norway, will have the opportunity to align their technology in accordance with national and international recommendations and best practices. CLARINO is also relevant for enterprises and research groups developing language technology systems for knowledge management, for checking of terminological consistency, for multilingual technologies such as machine translation, etc.

National CLARIN projects have been ongoing in various European countries. The proposed LAP shares its goals with the German national CLARIN initiative (D-SPIN). Establishing LAP as an open national infrastructure, building on existing initiatives such as NOTUR, NorStore, and NorGrid, will offer a common point of reference or 'clearing house' for other CLARINO sub-projects, as well as help establish an interface to ongoing national and international eInfrastructure activities. Our LAP vision takes as its point of departure the experience of the proposers in HPC-scale language technology, as well as the positive outcomes of providing an abstractly similar portal for Computational Biology (BioPortal) which has a good user base.[10]

Within Norway, a national *user community* for the use of HPC in language processing has been established after a model used in other scientific fields in Norway, to promote best practice and sharing of solutions. This group is led by Stephan Oepen, who is also representing the field on the eSOP panel for the future Norwegian eInfrastructure roadmap.

7 Partners

7.1 Consortium Participants

The CLARINO consortium consists of the following institutions, representing the major research groups on language resources in Norway:

[10]Other related projects include the Japanese Language Grid and various multi-national initiatives towards the use of UIMA as a common language technology middleware.

1. **The University of Bergen (UiB) (coordinator).** UiB has been the only Norwegian consortium member and national contact point in CLARIN. It has extensive experience in national and international projects in LRT and has coordinated the NO-CLARIN preparatory action in Norway. UiB has also become the Norwegian consortium partner in META-NORD, liaising with META-NET and META-SHARE. Units involved at UiB are the Dept. of Linguistic, Literary and Esthetic Studies (LLE), the University Library (UBB), the Dept. of Foreign Languages (IF) and Dept. of Philosophy (FoF)).

2. **The University of Oslo (UiO).** UiO has probably the largest collection of annotated language data in Norway at the Dept. of Linguistics and Scandinavian languages (ILN, which includes the Text Laboratory and EDD, the Unit for Digital Documentation). The Text Laboratory has expertise on taggers and development of corpora and corpus tools, and has the president of the Nordic Association of Language Technology (NEALT). EDD has expertise on electronic lexicography, scholarly text encoding and ontologically based annotation. USIT (the IT department at UiO) contributes with technical expertise and computing resources. The Dept. of Computer Science (IFI) has expertise in large-scale NLP applications and has a representative on the national eInfrastructure panel eSOP.

3. **The Norwegian School of Economics (NHH).** Its language department has received a national responsibility for terminology and has extensive expertise with digital termbases and corpora for specialized domains.

4. **The University of Tromsø (UiT).** UiT has a national centre of Sami language technology (Sámi giellatekno) with special expertise, resources and tools for Sami, in addition to an internationally known linguistics environment.

5. **Uni Research AS.** Its Uni Computing department has over four decades of experience in the field. Its role will be mainly in software development and data and metadata conversion.

6. **The Norwegian University of Science and Technology (NTNU).** The Department of Electronics and Telecommunications has long experience, resources and tools for spoken language processing, while the Department of Computer and Information Science and the Department of Language and Communication Studies have expertise in Computational Linguistics research.

Committed national partner institutions cooperating in the project with their own funding are the following:

1. **The National Library of Norway (NB).** NB is an organization with a very long-term stability and an established national responsibility for storing, curating and disseminating language materials to researchers and the general public. Their experience with cataloguing and secure, long-term storage capacity makes them the best qualified for setting up the national registry and long-term archiving for the infrastructure. NB also has the responsibility for The Language Technology Resource Collection for Norwegian – Språkbanken, which will provide its metadata. NB will contribute in CLARINO entirely from its own funding, as part of its national responsibility to curate and disseminate texts and audiovisual materials.

2. **UNINETT Sigma AS.** UNINETT has national responsibility for academic computing services such as networking, identity federation (Feide), HPC (NOTUR) and large storage (NorStore). It will contribute in CLARINO with its own funding as part of its role as technology and networking provider. NorStore access will be applied for separately by the consortium (see also above).

Internationally, all CLARIN nodes will be implicitly committed partner institutions through the ERIC. They will participate in the CLARIN AAI federation and the exchange of metadata. Working relationships will be established through the CLARIN working groups. International CLARIN partners, including the Max Planck Institute for Psycholinguistics (Nijmegen) and D-SPIN (Tübingen) will provide expertise to CLARINO on technical integration issues with the CLARIN infrastructure. Furthermore, international cooperative efforts in the EU-funded DASISH project, in which UiB participates, will promote common solutions for a range of large-scale infrastructures in the Humanities and Social Sciences.

8 Project Management

Overall project management and administration will be at UiB, which is currently the national contact for CLARIN. The scientific coordinator for CLARINO will be Koenraad De Smedt (see enclosed CV), who will also link internationally to CLARIN. The coordinator will also liaise to related projects such as META-NORD and DASISH. The consortium and partners have been allocated tasks which fit their competencies, and centres are localized at established centres of expertise. Janne Bondi Johannessen (UiO, ILN) has led the Text Laboratory for many years and has through many projects built up a competent centre in language resources and tools construction and dissemination. She will lead WP5. Christian-Emil Ore (UiO, ILN) has a long record in leading the EDD unit and its precursors and has through the Documentation Project and subsequent activities built up a significant portfolio. He will lead WP4. Stephan Oepen (UiO, IFI) has extensive R&D project management expertise (including the LOGON and WeSearch projects). He will lead WP8 and WP9. Gisle Andersen (NHH) has been director of Unifob Aksis and has led projects such as the Norwegian Newspaper corpus. He will lead WP7. Paul Meurer (Uni Research) has developed numerous language software tools and will lead WP6. Kristin Bakken is director of the research department at NB and will take responsibility for WP2. Jacko Koster is director of UNINETT Sigma and will supervise WP3. Trond Trosterud (UiT) has long experience in building Sami resources and tools which he will contribute. He will lead WP10. Björn Gambäck will be the contact at NTNU and provide expertise in Language Technology. His colleague Torbjørn Svendsen is Norway's authority on speech data and processing. Other individuals at the consortium and partners will contribute with their expertise.

A steering group will be established with one representative for each centre/partner; the coordinator will chair this group. This body will take strategic decisions on project planning, adjustment of goals, resolution of conflicts, and remedial actions in case of problems. It will be especially important for the project to continually check that the work remains in line with the various dependencies between infrastructure components. An advisory council will be established consisting of CLARIN representatives, a representative of the Language Council of Norway, a representative from industry, and a representative of the government responsible for the Norwegian contribution to the ERIC. Day to day management (see WP11) will be at the coordinator (UiB) who will appoint a project manager under the supervision of the scientific coordinator. The management bodies will have yearly meetings. In case of unforeseen major problems, the coordinator or working group leaders may call for extraordinary meetings. Further working communication will normally proceed by telephone and email to keep travel costs limited.

A consortium agreement will describe responsibilities, rights, and procedures.

Coordination and data harmonization activities will be undertaken in cooperation with the CLARIN ERIC and the international nodes. The coordinator and/or working group leaders will attend relevant CLARIN meetings.

9 Plan for Access and Use, Data and Knowledge Management

CLARIN is a distributed infrastructure where CLARINO is a transparent subset of the CLARIN network. All potential CLARIN users are therefore potential CLARINO users. Usage will be through

portals at the centres, using a CLARIN-wide AAI. All Norwegian universities and research institutions will be directly connected to the AAI through Feide. Most or all of this usage will be free of charge, but access rights will depend on existing licences regulating the use of some materials. It is not expected that usage patterns will lead to capacity problems for CLARINO nodes, since on the one hand, good provisions are made for HPC deployment in the current proposal, and on the other hand, many Norwegian resources will not have an overly large target group abroad.

Knowledge transfer and dissemination within the CLARIN network is important for an efficient construction and effective deployment. A practical cooperation is already in place through the various CLARIN working groups, in which the project participants will actively participate through distance communication as well as workshop attendance. The CLARINO project will, in agreement with CLARIN, organize some of these workshops.

In order to reach primary target user groups in the Humanities and Social Sciences, and secondary user groups among developers of language based applications and services, easy access and dissemination activities will play an important role. A project web site will be established with up to date information on progress, events and results, and the service centres will establish their own portals.

Since the project itself is not primarily research, the potential for scientific publications will be limited, although it is expected that some research publications demonstrating the use of the infrastructure will be published already within the construction phase. The project will be presented in national and Nordic arenas (such as MONS, NordTerm, NoDaLiDa) and publications (such as Språknytt) reaching a wide Norwegian audience. It will also be presented at relevant international events (such as LREC, EACL, NEERI and CLARIN conferences). CLARINO will organize annual dissemination events in conjunction with its project meetings, and will attract media attention.

10 Time Schedule and Deliverables

The project will run over five years of construction and testing. The work will be divided into the following main Work Packages. Each work package will involve the following which will not be repeated in each WP description:

- gathering user requirements, drafting specifications (taking into account CLARIN standards) and checking these with other workpackages;
- prefinal testing of pilot with the involvement of end users (where appropriate);
- prefinal checking of compatibility with other infrastructure components;[11]
- writing of documentation (normally online) and reporting to coordinator.

Detailed timing of deliverables is listed in the main electronic application form.

WP1 **Centres Setup.** All centres offering resources will make their schemas explicit, refer to the vocabulary specified in accepted data category registries and set up to offer appropriate metadata descriptions (see also 10). Some will support the OAI PMH protocol, others will support XML harvesting. Their resources will be maintained in a well-structured and documented repository system with a long-term commitment and associated with an accepted PID service (either in the institute or by making use of registrations at another accepted PID service site). Versioning will guarantee that references will remain valid. All type A and B centres will take part in the CLARIN AAI to allow authentication and authorization of users registered in Feide and the CLARIN AAI federation. Advisory services, ingest services, preservation services, resource discovery services, access services and language technology services will be provided by all type A and B centres except for UNINETT. Centres will provide support and training,

[11] By the end of Y2Q1, the participants in WP8, WP9 and WP10 will produce a joint specification of interfaces across these components.

and online documentation for all services they offer. By the end of Y1, the implementation of the centres will be tested and stable enough to serve as a basis for use by other WPs.

WP2 **National Registry and Long-Term Archiving.** NB will set up a national data registry facility providing DC mapping and an OAI PMH gateway for metadata exchange with other national CLARIN registries and will set up OAI PMH or XML based harvesting from all Norwegian nodes providing resources. NB will set up an interface allowing users to search the registry for language resources based on specified metadata. NB will set up a national long-term archive coupled to the registry, with secure redundant storage and media migration. The registry and archive will be operational by the end of Y1, but interaction with other components will be implemented as needed in subsequent years.

WP3 **Trusted Authentication and Authorization Infrastructure (AAI).** UNINETT will implement the centres' interaction with a Shibboleth resource provider instance to participate in the CLARIN AAI. It will interact with other federations at the international level to work towards future European integration in these matters. It will provide expertise and software support the A and B centres' implementation of authentication solutions based on Feide. It will also implement a user authorization system based on user identification and the licensing data stored in the national registry. This component will be ready after Y1.

WP4 **Electronic Editions Platform with IDP.** A platform for electronic editions will be implemented by extending the work on IDP and making it generally applicable by a web service which on the basis of an XML-TEI-conformant text and additional text-specific DTD/XML schemata characteristics generates user-steered transformations with XSL stylesheets. This work at EDD (in cooperation with FoF at Bergen) involves a generalization of current scripts and user interface optimization. An XML database with search functionality (e.g. Exist) will be created to allow structured searching for codes in the XML documents with a user-friendly interface; this work will be contributed by Uni Research. The Menotec project will from its own project efforts contribute the data, schemata and metadata for its own texts. The platform will gradually be put into operation as its functionality is being expanded and will be fully operational after five years.

WP5 **Glossa Integration.** The current Glossa corpus interface and analysis tool will be integrated in the CLARINO infrastructure through the following extensions, carried out by UiO (at the Text Laboratory, ILN, in cooperation with USIT): (a) automation of corpus building by creating the necessary server software, (b) a framework for the use of internal and external text analysis and management tools, and integration of this framework into the corpus construction pipeline, (c) a framework for metadata-based user interface and a port for metadata harvesting by the national registry, (d) integration of the user management and access control features into a federated user authentication framework, (e) a fully self-service workflow for creating and managing the user's corpora, and (f) user communication features such as self-service access control and online documentation, comments, blogs, etc. Glossa will be provided as an online web-based service hosted in cooperation with USIT at the University of Oslo and will function in the infrastructure after Y3.

WP6 **Corpuscle Integration.** The Corpuscle corpus management tool, which is already operational and will be fully functional by April 2011, will be integrated and maintained in the CLARINO infrastructure through the following tasks, carried out by Uni Research: (a) integration into a federated user authentication framework and authorization system to manage levels of access according to licenses and user groups, (b) interface for self-registration and self-serviced addition of new material, (c) download and export possibilities in several relevant formats, (d) a meta-framework with user interface and a port for metadata harvesting by

the national registry, (e) troubleshooting and regular updates of the software as needed according to evolving user needs and depending on updated libraries, compilers and hardware, (f) parallellization of algorithms for query execution to exploit the computing power of HPC clusters, and (g) improvement of various functionality (e.g multimedia support, online documentation) based on user feedback. Corpuscle will be integrated in the Bergen centre and will be gradually put into operation, having full functionality after Y3.

WP7 **Terminology Integration.** The ClarinTerm component will implement a dedicated repository for terminology-related LRT, giving users access to terminology services that enhance the value of their resources. Services will be provided for distributed management of termbase entries, for ontology building, for efficient data conversion, for linking termbase entries to relevant domain-specific text corpora, for neology identification, for segmentation of multiword entries and for mono- and multilingual term candidate extraction. CLARINO will function both as a repository for existing termbases (including 'historical' resources that are no longer being updated) as well as a portal giving direct access to and management of current and continuously updated terminology. Standard XML-based knowledge representation schemes will be deployed, including RDF, DAML and OWL. The project will also disseminate guidelines and standards (such as ISO 16642, the TBX exchange format and CLARIN's terminology standards), as well as offer training in and administration of T-LRTs. The best practice will include methodological aspects such as methods for annotation of terminological data, recommendations for enhancing poorly architected systems (such as spreadsheets), sense mapping techniques, definition writing, handling of synonymy and preferred/deprecated terms, harmonising terminological resources at different levels, including terms, concepts, subdomain/domain/topic classifications at varying granularity, as well as workflow. This will be a five year project at the Bergen centre with primary input from NHH.

WP8 **LAP.** This work package comprises the following sub-tasks: (a) user requirements and technology survey (including the wider CLARIN and META-NET contexts); (b) architecture specification; (c) middleware design and implementation; (d) user interface design and implementation; (e) integration of language analysis components; (f) support and training to component providers and users (including documentation); (g) portal maintenance and operations; and (h) liaising with other CLARIN(O) participants.

The analysis portal presupposes authentication and authorization services in WP3, access to the data in the repositories at the content providers, and access to compatible tools (WP9). Furthermore, for mid- and long-term operation, bulk allocation of storage and computation resources from the national eInfrastructure (NOTUR, NorStore, and NorGrid) is anticipated.

In terms of linguistic coverage, LAP will focus on languages actively used in Norway, e.g. Norwegian Bokmål and Nynorsk, Sami, other Scandinavian languages, and English, initially at least with a focus on written language. The portal will reach out to national and international developers of processing tools, seeking to install the broadest possible range of technologies, and encompassing both rule-based and statistical approaches. Besides tools in active use across Norway today (e.g. finite-state and constraint grammar analysis, or the ParGram and DELPH-IN toolchains), results from the VerdIKT project *WeSearch: Language Technology for the Web* will be integrated in the LAP to enhance functionality and avoid duplication of effort.

LAP will assist in license clarification, interface definitions for conversion from and to interchange formats, 'wrapping' of tools for HPC use, general 'hardening' to improve scalability and interoperability, and on-line documentation. Technology developed (or currently used) in Norway will be complemented with standard language analysis tools available from international players, with a bias towards general-purpose, open-source systems. Establishing the portal will also comprise a limited amount of in-house tool development or adaptation of existing technology, for example to provide a broader range of interfaces to common representation

formats, to support users in gauging expectations for a specific technology and evaluating results, or to facilitate better customization of component tools, e.g. adaptation or retraining for a specific domain, genre, or task.

A key element in this last strand of R&D will be the provision of methods and technology for intelligent and flexible combination of components, aiming for custom language analysis performance and cross-domain portability well transcending off-the-shelf technology. Such combination can build on ensembles of components at parallel analysis levels, for example multiple taggers working hand-in-hand, or 'stacking' of parsers to improve analysis accuracy and robustness to variation in inputs. Establishing LAP will take 5 person years at IFI and USIT, with appropriate checkpoints on the way.

WP9 **Tool Adaptation.** Existing grammars and taggers will be adapted so as to produce a CLARIN compatible format and become more easily integrated in pipelines within the LAP, as well as in other conformant platforms. This work will be carried out by tool providers in cooperation with IFI (UiO). Rule-based taggers for Norwegian Bokmål and Nynorsk (comprising a tokenizer, morphological analyser, guesser, and rule-based disambiguator), as well as statistical taggers for orthographically transcribed spoken Norwegian, Swedish, Danish, Icelandic, and Faroese will be adapted. The tagger adaptation will be carried out by the Text Laboratory (UiO), while some work on the Oslo-Bergen Tagger will be done in cooperation with Uni Research.

Furthermore, the constraint grammars for Sami will be adapted and optimized by UiT. For the three Sami languages, there are efficient morphological transducers and, for one of them, grammatical and dependency parsers available. Especially important are older texts, since these represent forms of the language less influenced by Norwegian. In order to automatically analyse these texts, new versions of the grammatical analysers will be needed, adjusted to the numerous orthographical and grammatical standards used in earlier text. In order to make the Sami analysers available in CLARINO, the build process must be standardized, the different components must be documented in a uniform way and the API must be made compatible with the LAP in order to get access to HPC.

Existing tools at NTNU for manual and (semi-)automatic transcription and annotation of audio, multilingual and multimedia content will be adapted to conform to CLARIN formats and to accommodate different audio quality as well as different annotation and audio formats. Finally, the NorGram grammar and XLE parser, as well as other auxiliary tools at Bergen, will be adapted by Uni Research and UiB to function in the LAP. The tool adaptation WP will be performed throughout the consortium and will deliver its results at the end of Y2.

WP10 **Data and Metadata Adaptation.** Every resource owner participating in CLARINO will provide metadata to be harvested by the national repository. Existing metadata will be mapped to CLARIN format and standards. This can partially be performed automatically, but the decisions on designing the mapping between existing categories and CLARIN standards must be done manually in each case, while missing metadata will be provided manually. At the type B, C and R centres, there are automatic and semiautomatic procedures set up for harvesting text, and for receiving text from different donor instances. For most texts, especially the important texts for which there are parallel versions, manual metadata addition and control is needed.

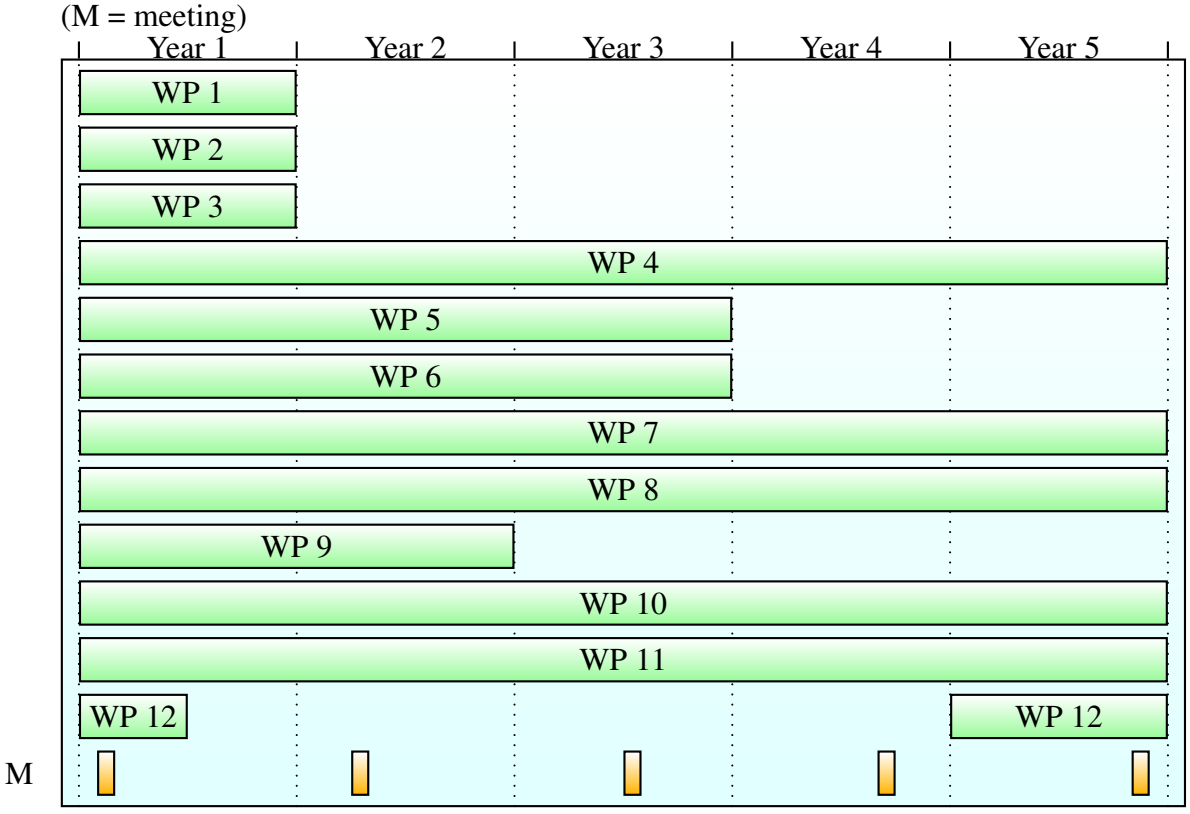
Furthermore, not only metadata, but also data itself will have to be adapted and streamlined. Texts often contain symbols, markup and encoding, even alphabets, that will be an obstacle when transferring data between tools. Older corpora and other language resources (including audio and multimedia) will need to be converted, re-encoded and re-tagged to follow CLARIN standards before being archived and imported in various tools and services. For messy data, e.g. newspaper corpora, filtering must be improved and semi-automatic metadata generation

must be improved at article level. Metadata is incremental work throughout the consortium and will gradually go from construction to a maintenance phase.

WP11 **Management and Dissemination.** The coordinator will delegate day-to-day management to a project manager, who will monitor progress, ensure communication between partners, keep a record of results, activities, and problems for further management and reporting, prepare meetings and workshops, keep an internal and external website updated and promote dissemination. The consortium will have annual meetings, each combined with a dissemination event (workshop/conference), if possible also a media event (see also above). The project participants will attend international CLARIN activities and the project will attempt to organize at least one significant international CLARIN event in Norway.

WP12 **Use Cases, Evaluation and Delivery.** Besides continuous testing activities in each WP, there will be an overall testing of the infrastructure in Y5. All partners will cooperate in this activity under leadership of the coordinator. As one of the guiding elements for evaluation, a list of use cases will be collected during the first six months of the project, including ideas for future research projects based on the infrastructure, but also simple empirical research questions which concretely exemplify the need for intelligent handling of one or several data sources. Early in Y5, the infrastructure will be evaluated and adjustments will be made. The infrastructure will be officially delivered during the final public event.

The following chart is a timetable of *main* work to be done in each Work Package. Continued testing and tuning will extend beyond these timelines.



11 Budget and Funding Plan

The *annual* budget that CLARIN expects for the infrastructure layer is over 45 M NOK for a country like Norway.[12] The current proposal has a lower budget representing a structural investment to be supplemented by appropriate funding for additional projects at the content layer.

[12] Cf. CLARIN Deliverable 8S-2.1a, adjusted for personnel cost rate.

The budget for CLARINO is about 41.3 M NOK for Y1–5, with a requested RCN funding of 25 M NOK (about 60% of the total financing) including the ERIC membership fee.

11.1 Construction Costs

Total construction costs in Y1–5 are nearly 40 M NOK, most of which is spent on salary costs. The consortium and partners will contribute with considerable, increasing funding of their own during the construction phase and beyond. UiB is contributing with one coordinator month, two senior researcher months and 2.5 senior library staff months per year during ten years of construction and operation. UiO IFI is contributing with 1.2 senior academic staff months during construction, half of that during operation, as well as three months for graduate level researchers. UiO ILN is contributing a total of 84.5 person months to the project. Two senior engineer months during the entire ten years will be contributed to WP8 by USIT, and an additional month per year during the construction phase to WP5. NHH and UiT will both contribute 20% of two senior academic positions during the construction phase and half of that during the operative phrase. NTNU will contribute one senior academic staff month per year during the construction phase and half of that during operation. NB and UNINETT are not asking for any funding from RCN, but will finance their own activities in CLARINO as part of their national responsibilities. Their estimated contributions are 48 and three person months, respectively.

Non-salary costs include the following:

- Investments for interactive application servers and laptops for development and dissemination, totalling 512 K NOK. The use of additional computer facilities at NB and access to national eInfrastructure (NOTUR/NorStore) are not included in the budget. NB will provide storage and computer use free of charge. NOTUR/NorStore access will be applied for, using the standard application procedures for access to national eInfrastructure facilities.

- A pool of minimally 8 months for visiting researcher stipends, mostly for active developers at other CLARIN nodes in Europe, is meant to secure technology transfer and coordination especially during the first year; additional exchanges in subsequent years will be applied for through other channels.

- Travel for coordination within the consortium (232 K NOK yearly), including the organization of workshops and events, and for cooperation with other CLARIN nodes in Europe is motivated by the need to achieve a high degree of national and international harmonization and integration. Also the participation in ERIC technical committees will require travel.

11.2 Operational Costs

Each project participant has committed itself to continued operation of its part of the infrastructure after the construction phase, sustaining a long-term integrated infrastructure. Together, the project participants contribute about 14.5 M NOK in Y6–10 to deploy and operate the infrastructure, most of which goes to salary. With this sum of resources and responsibilities, a usage model based on free access, compatible with present free access to institutional and national libraries, is realistic under present conditions. In this model, the consortium's resources must be matched with continued adequate access to national *eInfrastructure* (HPC, storage, middleware and support), as well as a stable synergy with the library sector which is expected take an increasing share of long-term archiving and cataloguing in the form of *digital libraries*. It is difficult to accurately foresee the needs for language research data and the conditions under which the infrastructure will operate five years from now. Therefore, the responsibilities for Y6–10 and beyond are not to be cemented, but may need readjustment after experience and ERIC decisions on continued national funding of investment and operation in the face of evolutions in the research data landscape during the construction phase.

The CLARIN consortium has proposed an annual contribution to the ERIC of EUR 34500, indexed by 2% yearly, from countries such as Norway.[13] This amount has been included as in the budget as an operational cost. In Y1–5, the membership fee is provided through the RCN project grant, while responsibilities after Y5 await further agreements.

11.3 Financial Contributions from Others

CLARINO, being a small part of the CLARIN infrastructure, will in principle have equal terms for data access from and to its European partners. Norwegian researchers who will gain access to the data and services from the pan-European CLARIN infrastructure, collectively funded by all participating countries, will thereby experience the other countries in CLARIN as substantial *in kind* contributors. An accurate budget for CLARIN as a whole is not available, but the planned investment over the coming five year period is in the order of 100 to 200 M EUR.

12 Environmental and Ethical Perspectives

The project will have no consequences on the natural environment. An ethical and legal issue concerns the privacy of speakers and authors in language collections. The consortium will take appropriate measures to assure personal privacy for all materials to be distributed.

[13]The estimated annual ERIC cost will be around 1 M EUR. A breakdown of the proposed cost is described in CLARIN Deliverable D8S-2.1a.