Universiteit Utrecht

Heidelberglaan 8, Utrecht

College van Bestuur

O&O, Postbus 80125, 3508 TC Utrecht

NWO
Commissie Update Nationale Roadmap
Onderzoeksfaciliteiten

Ons kenmerk
O&O/11.20986
Faxnummer
030 253 77 52
Telefoon
030 253 42 71
E-mail
e.i.stiekema@uu.nl

Datum
29 augustus 2011
Onderwerp
Aanvraag voor de update Nationale Roadmap Onderzoeksfaciliteiten

Geachte voorzitter van de commissie,

Naar aanleiding van de oproep voor het indienen van aanvragen voor projecten voor de Update Nationale Roadmap voor Grote Onderzoeksfaciliteiten bieden wij hierbij de aanvraag aan voor het project *Common Lab Research Infrastructure for the Arts and Humanities* (CLARIAH).

Met het CLARIAH project zal Nederland zijn leidende positie in Europa op het gebied van onderzoeksinfrastructuren voor de geesteswetenschappen verder versterken en een drijvende kracht zijn achter een fundamentele verandering in geesteswetenschappelijk onderzoek: het grootschalig toepassen van computationele technieken op enorme hoeveelheden data (Data Intensive Science).

Het CLARIAH project is een aanvraag gedragen door alle geesteswetenschappers in Nederland, die verenigd zijn in de projecten CLARIN en DARIAH die reeds op de roadmap staan. Een kort voortgangsverslag van het project CLARIN-NL is opgenomen in deze aanvraag.

Het college van bestuur ondersteunt de aanvraag voor het project CLARIAH van harte en is ervan overtuigd dat hiermee de geesteswetenschappen in Nederland hun sterke positie verder kunnen uitbouwen. De universiteit zal bij toekenning van de aanvraag aan alle verplichtingen voldoen die voortvloeien uit de aanvraag.

Mocht u na beoordeling van de ingediende projecten besluiten niet over te gaan tot plaatsing op de roadmap en/of financiering van het CLARIAH project, dan verzoeken wij u de projecten CLARIN en DARIAH wel op de roadmap te laten staan.

Mocht u nog vragen of opmerkingen hebben, dan kunt u ons bereiken via bovenstaande coördinaten.

Hoogachtend,
het college van bestuur,

mr. Yvonne C.M.T. van Rooy,
voorzitter

c.c.: prof. dr. W.J. van de Akker
prof. dr. M. Everaert
prof. dr. J. Odijk
prof. dr. P. Doorn, directeur DANS

Application form

National Roadmap for Large-Scale Research Facilities

2011

General information

CLARIAH - Common Lab Research Infrastructure for the Arts and Humanities

Kind of proposal *(please choose only one type of proposal)*

New facility (not in NL Roadmap 2008)

- ☐ inclusion in the roadmap only (no request for funding) **(type 1)**
- ☐ inclusion in the roadmap and request for funding **(type 2)**

Facility from NL Roadmap 2008

- ■ progress report and request for funding **(type 3)**
- ☐ progress report only (no request for funding) **(type 4)**

Main applicant

Title(s)	Prof. dr.
First name	Jan
Initials	J.E.J.M.
Surname	Odijk ☒male ☐female
Address for correspondence	Universiteit Utrecht Trans 10, 3512 JK Utrecht
Telephone number	+31 30 253 6006
Fax	+31 30 253 6000
Email	j.odijk@uu.nl
Website (optional)	http://www.uu.nl/hum/staff/JEJMOdijk/0

Co-applicants

Organisation		Signatory Letter of Intent	Function	Contact (if empty=signatory)	e-mail
1	UU	prof. dr. Wiljan van den Akker	Dean, Faculty of Humanities	prof. dr. Jan Luiten van Zanden	j.l.vanzanden@uu.nl
2	UvT	prof. dr. Arie de Ruijter	Dean, Tilburg School of Humanities	prof. dr. Jaap (H.J) van den Herik	h.j.vdnherik@uvt.nl
3	UvA	drs. P.W. Doop	Acting president CvB	prof. dr. Rens (L.W.M.) Bod	rens.bod@uva.nl
4	UT	prof. dr. P.M.G. Apers	Scientific Director, CTIT	prof. dr. Franciska (F.M.G.) de Jong	f.m.g.dejong@utwente.nl
5	UL	prof. dr. H.W. van den Doel	Dean, Faculty of Humanities	Kurt (K.F.K.) de Belder, MA MLIS	k.f.k.de.belder@library.leidenuniv.nl
6	VU	drs. R.M. Smit	President CvB	prof. dr. Piek (P.T.J.M.) Vossen	p.vossen@let.vu.nl
7	RUG	prof. dr. G.C. Wakker	Dean, Faculty of Arts	prof. dr. ir. John Nerbonne	j.nerbonne@rug.nl
8	RU	prof. dr. Paul (P.L.M.) Sars	Dean, Faculty of Arts		facbureau@let.ru.nl
9	EUR	prof. dr. Dick Douwes	Dean, ESHCC		douwes@eshcc.eur.nl
10	UM	prof. dr. Rein de Wilde	Dean, FASoS		r.dewilde@maastrichtuniversity.nl
11	Huygens ING	dr. Henk Wals	Director	drs. Joris van Zundert	joris.van.zundert@huygens.knaw.nl
12	DANS	dr. Peter (P.K.) Doorn	Director	dr. Ingrid (I.G.) Dillo	ingrid.dillo@dans.knaw.nl
13	MI	prof. dr. Hans (H.J.) Bennis	Director	Marc (M.) Kemps Snijders	hans.bennis@meertens.knaw.nl
14	NIOD	prof. dr. Marjan (m.) Schwegman	Director	drs. Edwin Klijn	m.schwegman@niod.knaw.nl
15	NIAS	prof. dr. Aafke (A.C.J.) Hulk	Director		a.hulk@nias.knaw.nl
16	IISG	prof. dr. Erik-Jan (E.J.) Zürcher	Director	prof. dr. Kees Mandemakers	ezu@iisg.nl
17	FA	prof. dr. Reinier Salverda	Director & governor	prof. dr. Arjen (A.P.) Versloot	rsalverda@fryske-akademy.nl
18	eHg	prof. dr. Sally Wyatt	Programme Leader		sally.wyatt@ehumanities.knaw.nl
19	MPI	prof. dr. Wolfgang Klein	Director	ir. Daniel (G.) Broeder	daan.broeder@mpi.nl
20	INL	dr. Jeannine (J.C.T.) Beeken	Director	drs. Remco van Veenendaal	remco.vanveenendaal@inl.nl
21	KB	dr. Bas (J.S.M.) Savenije	Director General	drs. Paul Doorenbosch	paul.doorenbosch@kb.nl
22	B&G	Hans (H.) Westerhof	Deputy Director	drs. Johan (J.F.) Oomen	joomen@beeldengeluid.nl
23	NLeSC	Prof. dr. Jacob (J.) de Vlieg	CEO		j.devlieg@esciencecenter.nl
24	NA*	Mr. Martin (M.J.) Berendse	National Archivist		martin.berendse@nationaalarchief.nl

* The participation of NA is conditional until clarity on its financial obligations is obtained (ult. on October 1, 2011).

The CLARIAH proposal is supported by a substantial number of organisations both from the public and the private domain. Their Letters of Support are incorporated in this document. The organisations in the public domain are: the research schools LOT, WTMC and Posthumus, UKB, the library of the UvA, the Libratory Consortium, DEN, DBNL, DIV of the House of Representatives. The organisations in the private domain are: STM, Brill, Telecats, Tessella, Microsoft, IBM, ArborMedia, Teezir, NOTaS, Knowledge Concepts and Furore. The supporting partners will be involved in the users panels and reflection groups of CLARIAH. During the course of the project their respective roles and activities will be determined in more detail.

1 General information
2 Research proposal
3 Timetable
4 Declaration/signature

Abstract

Summary

In the Common Lab for Research in the Arts and Humanities (CLARIAH) we aim to design, construct, and exploit a facility for *eHumanities* research. This virtual 'Common Lab' will provide a sustainable *eHumanities* research environment, which will provide researchers and research groups with integrated access to unprecedented collections of seamlessly interoperating digital research resources and innovative tools to process them in virtual workspaces, thus enabling *Data Intensive Science* in the humanities.

The Common Lab will provide researchers with a wide variety of resources and services, e.g. intelligent access methods for exploring resources and innovative ways of combining different resources into virtual collections. These services enable researchers to disclose and analyse information hidden in unstructured textual, image and multimedia resources, in combination with structured databases with qualitative and quantitative information. Interoperability of resources and services is a key element in the infrastructure. The infrastructure will be easy to access and use for scholars with a limited technical training. Through dissemination activities, educational programs and training sessions, a new generation of researchers and students will be able to acquaint themselves with new research methodologies. This will create the potential for groundbreaking research. The CLARIAH infrastructure will enable new research questions to be asked and old questions to be posed in new ways, boosting the quality, the effectiveness and the efficiency of research within the arts and humanities. The Common Lab will therefore attract top-researchers and students from abroad.

The CLARIAH consortium is of unprecedented breadth: it includes all humanities researchers in the Netherlands, the most relevant research libraries, heritage organizations, data centres and specialists in infrastructural matters, including the Netherlands eScience Center.

A wide range of public organisations and companies supports CLARIAH. It directly contributes to important policy issues in society and to the Top Sectors 'High Tech' and 'Creative Industry'. Through an outreach programme CLARIAH will involve these and other companies in the facility and create new opportunities for innovative commercial products and services.

CLARIAH is the national counterpart of two European research facilities on the ESFRI Roadmap and the Dutch Roadmap 2008: CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities).

With the CLARIN and DARIAH initiatives, the character of humanities research is going to change dramatically and forever. The Netherlands is already playing an important leading role in these developments. CLARIAH will strengthen the position of the Netherlands even more, and will enable our country to take on a unique driving role in the transformation process that the arts and humanities research is currently going through.

Summary of the research proposal in layman's terms

Rapid digitization of massive quantities of formerly analogue sources (text, images and audio-visuals) for research is revolutionizing the humanities. Top-quality humanities scholarship of today and tomorrow is therefore only possible with the use of sophisticated ICT tools. CLARIAH aims to offer humanities scholars, from literary researchers to historians and from archeologists to linguists, a 'Common Lab' that provides them access to large collections of digital resources and innovative user-friendly processing tools, thus enabling them to carry out ground-breaking research.

Key words

eHumanities, Virtual labs, Digital resources, Tools

Research proposal

Detailed description

1 Science case

1.1 Description of the facility

In the Common Lab for Research in the Arts and Humanities (CLARIAH) we aim to design, construct, and exploit a facility for *eHumanities* research. This 'Common Lab' will provide a sustainable *eHumanities* research environment, which will provide researchers and research groups with integrated access to unprecedented collections of seamlessly interoperating digital research resources and innovative tools to process them in virtual workspaces, thus enabling *Data Intensive Science* in the humanities. The Common Lab is virtual, i.e. the data, tools and facilities, as well as its developers and users are distributed over various locations and institutes. The distributed nature of the Common Lab, however, will not be visible to its users, who will have access to it via Internet. The Common Lab is preferably to be integrated in the evolving National eScience Infrastructure.

The Common Lab will provide researchers with a wide variety of resources and services, e.g., intelligent access methods for exploring resources and innovative ways of combining different resources into virtual collections, so that information hidden in unstructured textual and multimedia documents, in combination with structured databases with qualitative and quantitative information, can be disclosed and analysed. Interoperability of resources and services is a key element in the infrastructure. The infrastructure will be easy to access and use for scholars with a limited technical training. Through dissemination activities, educational programs and training sessions, a new generation of researchers and students will be able to acquaint themselves with new research methodologies, thus creating the potential for groundbreaking research. The CLARIAH infrastructure will enable new research questions to be asked and old questions to be posed in new ways, boosting the quality, the effectiveness and the efficiency of research within the arts and humanities. The Common Lab will attract top-researchers and students from abroad.

CLARIAH builds upon the only two large research facilities in the domain of the humanities on the Dutch Roadmap 2008: CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities). CLARIN received funding for a national project (CLARIN-NL) at the Roadmap update in 2008. CLARIAH differs from CLARIN-NL, which is focused on language, in two respects (1) it joins forces with the Dutch DARIAH community and its infrastructural needs, especially in the area of data and tools for structured historical data, and (2) it deals with aspects not or hardly covered in CLARIN-NL, in particular audio-visual data and tools, and facilities for virtual workspaces. The proposal combines the best and strongest aspects of CLARIN and DARIAH, as will be described in the sections that follow. CLARIAH will e.g. determine a set of CLARIAH standards and best practices based on CLARIN and DARIAH standards to ensure full interoperability.

A schematic overview of the CLARIAH infrastructure is provided in Fig. 1. The data and tools that form the instrumentation of the Common Lab can be grouped in three kits reflecting the most common data types used by scholars in the arts and humanities: text, structured data and audio-visual resources:

- Text Kit, focused on textual documents and images thereof (because texts are often linked to surrogates in digital image form of the originals). This kit will contain data and tools for the study of textual documents, which are probably the single most important resource type in humanities research.
- Structured Data Kit, focused on databases and other tabular resources, such as spreadsheets, and their associated tools. Such structured data play an important role in quantitative humanities research, for instance in social and economic history, but also in linguistics, e.g. for lexical databases, typological databases, databases of dialectical variation, etc.

- Audio-visual Kit, focused on audio, video, visual images, and multimedia. The kit will contain resources and tools for the treatment of, for instance, speech data, oral history interviews, and digital reproductions of paintings.

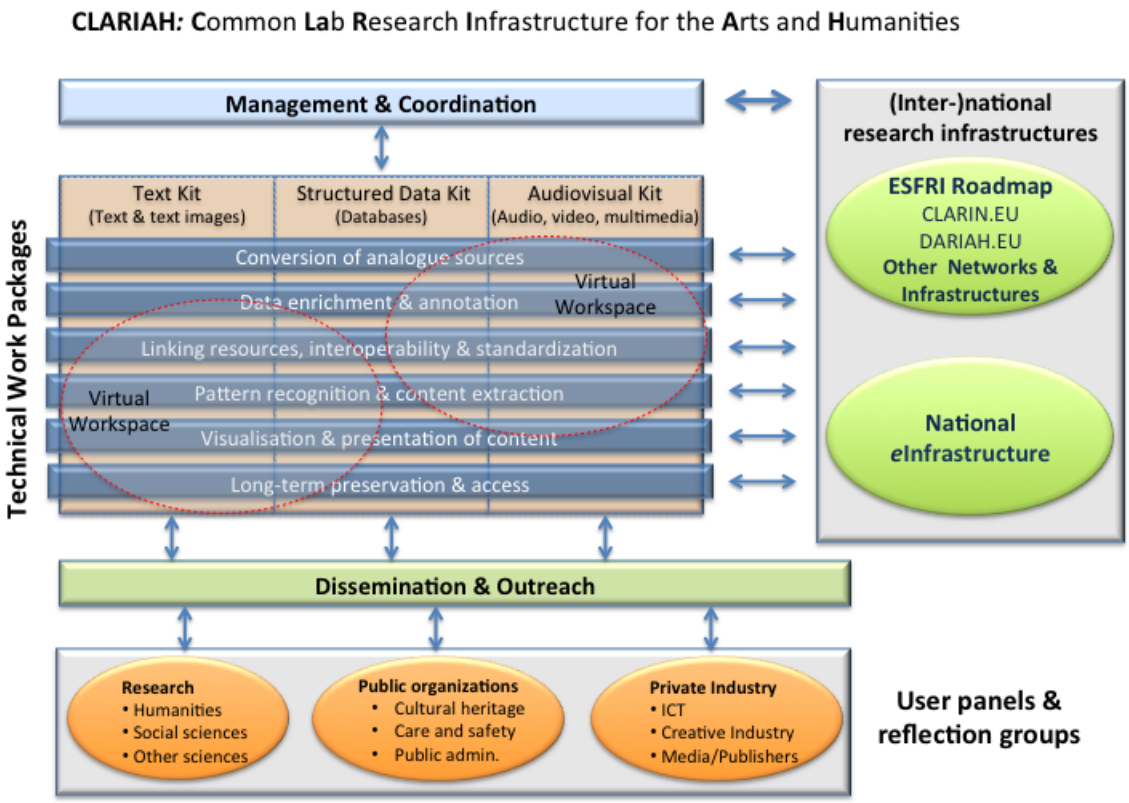


Fig. 1: A schematic overview of the CLARIAH infrastructure

Actual humanities research will often involve not just one data type, but a mixture of textual, image, audio-visual, semi-structured and/or structured data. The fact that all the data and tools are part of one and the same Common Lab and must meet requirements that the Common Lab levies or recommends will ensure that they are visible and accessible via a single portal and able to interoperate seamlessly.

In order to fill the kits, both existing and new data and tools will be documented, made visible, accessible, and referable in the Common Lab, in compliance with CLARIAH standards and procedures, and they will be made available whenever possible under open access licences[1], conforming to best practices for digital durability. The data shall be made formally and semantically interoperable in a stepwise manner. Newly developed tools shall be open source, and existing tools shall be adapted to become web services and/or web applications, in accordance with CLARIAH guidelines[2]. Principles of sustainability of services will be applied as much as possible. This should make it possible to re-purpose data to arbitrary virtual collections, and to combine tools and/or services in new virtual workflows, all being offered via the Common Lab to be part of the research and innovation process. The tools should moreover be user-friendly to users with limited technical background while also be supportive for experts.

Inventories of required and available data collections and tools will be made and prioritised for inclusion in the Common Lab (e.g. more generic tools and tools needed in independently financed research projects will get higher priority; other criteria will be developed in the project). If crucial tools or data are lacking, they will be created

[1] Certain data and tools, however, may require other licensing models

[2] Dependent on technology innovation and application areas locally running applications will be relevant to researchers.

(provided that this does not require a separate research activity). Existing tools must, if needed, be adapted to make them scalable to deal with data of widely different volumes and complexity.

The Common Lab will be hosted on (virtual) servers of selected partner institutes, in particular CLARIAH Centres (i.e. CLARIN Centres and the DARIAH Virtual Competency Centre on Scholarly Content Management (VCC 3)). Its data and tools are preferably stored on the emerging national eInfrastructure proposed by SURF as a successor to BiG Grid. The CLARIAH-centres will play a vital role in the curation of the resources by guaranteeing permanent access to the tools and data. The Common Lab will provide facilities for individual researchers and groups of cooperating researchers to create a personal (or group-specific) virtual research environment in which they can select the required data, subsets of data and new combinations of data in virtual collections, store such virtual collections, and get support for the maximally automatic metadata description of such virtual collections. The researchers will be able to select tools (e.g. as web services in a workflow system, as web applications or where necessary as local applications) and apply them to the data. Tools to facilitate the interpretation of the results by non-experts will be incorporated. It will be possible to store the results of the application of these tools in the personal research environment and they will be enriched with metadata and provenance data in a maximally automated manner, so that results of computationally intensive processes or processes that require a lot of manual effort (e.g. semi-manual annotation) can be re-used as first class CLARIAH objects by the research group itself as well as by other researchers. The Common Lab will thus support the whole data life cycle, adhere to basic design principles and nicely fit with the emerging Digital Object Architecture (DOA).

The CLARIAH centres with their stable repositories and archives will form the backbone of the Common Lab, preferably based on the national eInfrastructure for all disciplines as proposed by SURF. They will host, harvest and make available CLARIAH-compatible metadata, provide registries for syntactic and semantic interoperability, and use PIDs and bona-fide long-term preservation technologies and protocols.[3] They can also deal with a single sign-on authentication and authorisation process, provenance information, different versions and representations, so that the Common Lab provides collaborative spaces where users can follow the genesis of resources and data sets and obtain suitable resource representations. Here we will collaborate with and build upon results of the projects EUropean DATa (EUDAT) and Data Service Infrastructure for the Social Sciences and Humanities (DASISH) in which these issues are tackled and generalised repository APIs are being developed.[4]

When a new set of tools becomes available, existing independently financed research projects will be requested (and supported) to work with them to test whether they offer the required functionality and the desired user-friendliness. Their experiences will be taken into account to make improved versions of the tools. Some large-scale resource will have to be disclosed using these tools in order to test the suitability of the infrastructure for *Data Intensive Science* in the humanities.

1.1.1 Text Kit

Text is the primary source of research data and the predominant means of communication in the humanities. Text in our society is presently all but equated with *digital* text. In the not too distant future all unique books in the world (estimated at roughly 130 million items) will be digitally available. Digital text as resource of information and researchable data is already abundant. The size of the web is estimated at roughly 17 billion web pages (all containing text in different degrees), at least 740 million of these in Dutch. That is just the web visible to search engines; a manifold of that information is stored in back ends of databases (journals, digital libraries, archives etc.). Without appropriate tools to find, select, mine, analyse and visualize textual resources nobody in the near future will be able to claim to have taken into account sufficient relevant textual information available about any subject of research.

The mass of textual data becoming available offers humanities researchers exciting new possibilities to pose new research questions and tackle previously unanswerable ones. With the Text Kit we want to provide the humanities

[3] E.g. the Data Seal of Approval http://www.datasealofapproval.org/

[4] http://www.lat-mpi.eu/latnews/tag/eudat/

with cutting edge digital technology to make sense of the multitude of textual information becoming available, and to address new research questions with evidence based approaches, in short Data Intensive Science. The technologies take three forms.

1) Improved tools for the curation of textual resources. Part of our cultural heritage is at risk of 'falling off the radar' because industry level digitization is not able to mass digitize historic forms and formats of texts, nor to provide the adequate metadata needed for research and preservation. Pushing the bar for semi-automated high-end digitisation of textual resources should be a primary objective to guarantee the availability of a balanced and well-represented digital footprint of the human textual record.

2) All kinds of (semantically shallow) tools exist for data mining and extraction of information from textual resources. But also smarter ways of mining and analysing relevant textual information from large datasets are becoming available. Higher level pattern recognition tools based on knowledge gained in the field of linguistics and natural language processing (NLP) are emerging. These tools should be incorporated in the Common Lab, and the Common Lab should provide facilities and guidelines to include new versions with even better functionality that will be developed in the near future, thus raising the bar to automated semantic analysis to be able to create information agents that can discover and associate meaning instead of just identifying superficial word similarities in mass textual resources.

3) With the increasing capabilities of the tools, it must also be closely monitored whether existing standards of data models for representations of text and meso-text (i.e. annotations and interpretations of text) are still appropriate, and whether extensions and modifications are required, in order to express lexical, syntactical and semantic levels of text, and allowing for the discrete expression of multiple levels of annotation and interpretation.

The tools to be incorporated will be reusable components allowing researchers to create purpose specific virtual research environments. Reusable and generic components also provide for further support of multidisciplinary approaches.

1.1.2 Structured Data Kit

The Common Lab will offer facilities enabling researchers to work with integrated databases and other tabular resources, such as spreadsheets. It will offer easy access to tools for searching, processing and/or enriching these data. The Structured Data Kit will provide facilities to carry out *Data-Intensive Science* for structured data sources in the humanities, e.g. quantitative data in historical studies, qualitative and quantitative data in language-typological databases, lexical databases, databases of dialectical variation, etc.

Existing data and tools will be documented, made visible, accessible, and referable. The data will also be made formally and semantically interoperable. Particularly important in this context are visualisation tools as a means for exploration and dissemination. It will be investigated which functionality the visualisation tools will have to offer, and how these tools can best be obtained. The focus will be on two data collections, which were part of the 2008 DARIAH-NL proposal for the national roadmap: Clio-Infra and HSN (the Historical Sample of the Netherlands). These data sets will be incorporated in the Common Lab in accordance with the CLARIAH standards and procedures, and, since large parts of the required data are still lacking, they will be significantly extended.

The procedures, protocols, methodologies and tools developed in Clio-Infra will be made part of the Common Lab, so that they become available for collaborative work on large distributed databases for other domains as well. Furthermore Clio-Infra will be significantly extended beyond the possibilities within the current project funded by a NWO large investment subsidy acquired in 2010, by creating three essential data hubs and a facility for linking the data to journal articles. The three hubs are: (1) A **migration** hub, consisting of large datasets on migration flows, individual migrants and migrant organizations; (2) A **sustainability** hub covering the links between economic development, climate change and other forms of environmental stress; (3) A **technology** hub covering indices of technology. Moreover, within CLARIAH the Clio-Infra data will be connected to the literature by facilitating journals (in economics and economic, social and demographic history) to enforce a Data Availability Policy (DAP), implying that papers can only be published if authors make the underlying data available, first to the referees, and after publication to the scholarly community. CLARIAH will develop and offer a service to such journals, which makes it

possible to upload datasets that are archived and made available to other scholars, in the process creating 'enhanced publications'. Partners in the CLARIAH consortium have developed expertise in the area of e- publications in various national and European projects (such as SHARE, DRIVER, OpenAire).

The current database of the HSN will be extended in three directions: time, kin, and health/living standards. (1) Time. The family reconstitutions contained in LINKS ('**Link**ing **s**ystem for historical family reconstruction) will be used to reconstruct life courses of the early HSN cohorts, thus allowing a truly long-term perspective on social and demographic change. (2) Kin. Currently, HSN provides data only on co-residing kin and there is no way of knowing the size, composition, let alone the role, of kin beyond the household. (3) Health and living standards, made possible by the anthropometric data contained in militia registers.

1.1.3 Audio-visual Kit

The Common Lab will enable researchers to work with a range of audio, video and visual image data and will offer easy access to tools for searching, processing and/or enriching these data, thus enabling the researcher to carry out *Data-Intensive Science* for audio-visual resources in the humanities.

An inventory of required and available data collections and tools will be made. If crucial tools or data are lacking, they will be created (if they do not require a separate research activity). Existing tools must, if needed, be adapted to make them scalable to deal with data of widely different volumes and complexity. And all data and tools will be adapted to make them seamlessly interoperable. Tools and their user interfaces will be made user friendly so that they do not require special technical skills. CLARIAH has major content providers for audio-visual material such as B&G and MPI among its participants and will closely co-operate with *Verteld Verleden*[5] in which the CLARIAH partners B&G, DANS and MI participate, and with *Beelden voor de Toekomst*[6] in which B&G participates.

The visibility and accessibility of these resources via CLARIAH-centres also offers a basis for making provisions for media resources or fragments of these to be referenced from publications either as individual files (as examples) or as complete virtual collections for verifiability or sharing with others. In the case of enhanced publications there is the need for tighter integration of media data in the publication. To that end, we have to incorporate tools and apply standards that allow creating such media (fragment) references and collections in the process of (e-)paper authoring but also of other publication platforms such as for instance layered maps.

The media object references from publications are of course only part of the total network of references to media that will be used by VREs. There is also a network of relations between media objects or media and textual objects captured in annotations or in the form of open linked data that needs to be accessible and exploitable with advanced data mining technologies.

1.1.4 Interconnections: shared methodologies, management, outreach

The three resource kits are interconnected by a set of shared methodologies, structured according to the 'digital research cycle'. These methodological layers take the form of working groups managing technical work packages that cut across the resource types. They are:

- Conversion of analogue sources: the overwhelming majority of humanities resources, most of which are located in cultural heritage organizations such as archives, libraries and museums, is analogue. A shared facility which offers expertise in the most effective and efficient conversion strategies, either by applying innovative automated recognition technologies (speech recognition, optical character recognition, image recognition, structure & format recognition for tables), or by applying manual data entry (still the best option for most handwritten data). Mass digitization in itself is not the aim of CLARIAH; rather, in dedicated areas, selected materials need to be made digital in order to complement gaps in already existing digital collections, such as those of the Historical Sample of the Netherlands or the vital economic history data in Clio-Infra. Coordinator: UVA

[5] http://www.verteldverleden.org/

[6] http://beeldenvoordetoekomst.nl/ According to this website the amount of its digitised audio-visual material is over 159 million hours (Aug 30, 2011) and counting.

- Data enrichment & annotation: in most humanities research, data can only be sensibly analysed when it is enriched with interpretative information, and when the appropriate metadata is added. Automation of this process makes it less time-consuming, and CLARIAH aims to offer intelligent tools to support this labour-intensive phase in the preparation of digital resources. Annotation tools and tools for automated metadata extraction belong to this layer. Coordinator: Huygens ING
- Linking resources, standardisation and interoperability. Coordinator: MPI. It consists of two different but closely related topics:
    - Standardisation and interoperability: dealing with standards for formats, standards for semantic interoperability (ISOCAT, RELCAT, etc.): many digital humanities resources are currently living (and dying) in isolation in separate silos or containers. Making them interoperable both on the syntactic and on the semantic level is probably *the* challenge of the coming decade, but it is essential to be able to make scientific breakthroughs. CLARIAH will offer the facilities and protocols for making resources syntactically and semantically interoperable, based on widely accepted standards and best practices, and will test the standards against real data. No doubt, new types of data and tools not envisioned so far will emerge, which requires continuous extensions and possibly adaptations of the CLARIAH standards.
    - Linking resources: Interoperability provides a prerequisite for linking resources: relating resources to other resources, data to publications, etc. This includes virtual collections, citation options, relations with linked open data, data fragment identification, etc.
- Pattern recognition & content extraction: in the interpretation and analysis of (digital) humanities data, the essential question is: what does it all mean? How to make sense of the data? How can we distinguish information from noise and detect high-level patterns? This is also a big challenge for the next generation of self-learning systems under development in information and computing science, and a field where collaboration between CLARIAH, humanities researchers and High Tech Industry will be fruitful. CLARIAH will offer tools to extract valuable information from heterogeneous, complex and fuzzy sources. Coordinator: UvT
- Visualisation & presentation of information content: visualisation tools are essential to present large quantities of complex information. Enhanced publications, in which text (such as an article, dissertation or monograph) has been enriched with additional material (databases, images, audio-visuals, etc.) is innovating the transfer of knowledge in exciting new ways. CLARIAH will offer tools for visualisation and support enhanced publications, for example GIS-tools for visualising historical and other humanities data over space and time. It will also employ gamification techniques for increasing the appeal of the sources to a wider audience, and in order to use crowdsourcing for the elaboration of resources. Coordinator: MI
- Long-term preservation and access: The data and software created and used in the Common Lab need to be managed, curated and archived, making sure that the substantial investments in preparing and presenting the data and tools will be durable. For permanent storage preferably use will be made of the existing BiG Grid infrastructure and of its successor e-science infrastructure for storage capacity and processing power. The possibility to use this layer will moreover be an integral part of the CLARIAH Centres. CLARIAH will make sure that digital humanities resources will be uniquely referable, that interoperable metadata profiles, joint standards and policies will be used, that quality standards (such as the Data Seal of Approval) are met, and that access conforms as much as possible to open source and open access licenses. Legal expertise is an element of this layer. Coordinator: DANS.

A management and coordination layer governs the Common Lab, as described in more detail in section 4 (Partnership Case). The existing CLARIN.NL office will be expanded to serve CLARIAH as well, thus avoiding duplication and creating synergy. The layers will be integrated by overarching facilities for virtual workspaces. Virtual workspaces must enable groups of cooperating researchers to create their own virtual research environment in which they can select the required data and tools, process the data, store the results, etc. This integration will be the special responsibility of the CLARIAH technical director. For browsing and searching in data and metadata, as well as a metadata registry and associated tools, the facilities being developed in CLARIN will be used.

Finally, the facility will have a joint dissemination and outreach layer described in more detail in section 2 (Talent case).

1.2 Scientific background and research questions, expected results and breakthroughs

In e-Humanities[7] research, computational methods and techniques are applied to core issues in the humanities[8]. It includes pattern recognition, sequence analysis in text and historical data, modelling and simulation, and the presentation of the results in images (visualisation; animation) and sound, all encoded in algorithms. It also includes innovative ways of data acquisition, validation, storage, documentation (annotation), processing and dissemination.

Computational approaches enable the researcher to address important overarching research questions in the humanities[9] in completely new ways. They challenge the humanities by reformulating old problems and formulating new problems that address core areas and by proposing new ways to solve these problems by perspectives and new methods including reduction, embedding, transformation, and simulation.

Large parts of the humanities are already making effective use of information technology for data processing, storage and access, and to extract information from data. In ten years from now, this will be the normal way of working in the humanities. CLARIAH provides the Netherlands with the opportunity to play a leading role in this process. Observational studies and experiments, imaging and digitalisation are generating vast volumes of data[10], which are only overwhelmed by the even larger quantities of naturally occurring and digitally born data relevant to the humanities in the Information Age. Vast amounts of data are of limited value if data mining technologies are not available or access is restricted, or if the knowledge infrastructure does not yet exist to create new knowledge from data. It takes teams of skilled personnel to compare data, to detect patterns, to describe and comprehend the processes that are revealed in the data, to capture in a model the essential features of a pattern or process and to separate it from secondary features and noise. The 'humanities perspective on information' may lead to a breakthrough in ways to maximise our insight from complex, fuzzy and incomplete data through the CLARIAH project. The proposed facility will also have an integrating effect on the study of language, music, art, history, religion, and archaeology. But in order to give more specific examples of expected scientific breakthroughs, we have to look at the resources.

Text

The Text Kit will be significant in a wide range of domains, varying from social and cultural history to the study of politics, social sciences, literary studies, the history of science, and art history. Spin off tools for higher grade data mining and knowledge extraction are expected to serve interest in both media and IT industry as well as non-profit areas (cultural heritage institutions and research institutes/communities).

Research questions that may be answered using the Text Kit vary widely. We can only list a few here: Mining the historical written record enables us to trace the development of ideas through time (cf. CKCC project). Research on comparing and associating syntactic and semantic units from different texts in large corpora may clarify properties of successful literature, relevant for the understanding of cultural dynamics. Identifying patterns and motives in large text corpora results in new insights in the origination and spread of political ideals and ideas in current and historic times. (cf. the Political Mashup project (UvA) and the War in Parliament project (CLARIN)). Semantics-enabled tools identifying patterns and associations provide information on (changing) attitudes and interpretation. (for example to enrich Dr L. de Jong's impressive series *Het Koninkrijk der Nederlanden in de Tweede Wereldoorlog* facilitating research on Dr De Jong's methodology and provide further information on the much debated question of De Jong's changing research perspective over time). Tools for the analysis of discrepancies between 'stories', e.g. between the accounts of victims vs. criminals of war as part of historical research address the challenge of semi-automatically identifying narrative structure, for which IBM has also shown interest. A historical search engine that offers clues about which information could have been available to contemporaries at a certain moment in time and at a certain place is of particular interest to various fields of historical study.

---

[7] We will not go into the subtle differences between the terms e-Humanities, computational humanities and digital humanities here, nor will we dwell on the different connotations of e-Science and computational science.

[8] We liberally quote here from the KNAW report on 'Computational Humanities', Amsterdam, 2009/10.

[9] E.g. as described in the KNAW Computational Humanities report.

[10] Even though this will be only a fraction of the available analogue data.

The ability of research based primarily on textual sources (e.g. history, literary studies, politics research, etc.) to provide generalized conclusions and hypotheses is much forestalled by the heterogenic and complex nature of textual data. State-of-the art semantic analysis tools are breaking this barrier and will enable researchers to derive generic trends and patterns from large scale unstructured data. CLARIAH thus facilitates the progress from data mining to knowledge mining. This will allow researchers of history, literature, politics, etc. to analyse and simulate processes of ideas interacting and shifting. This will allow a leap in the ability to understand the dynamics of societal processes like ideology uptake, cultural trends, and causes of historic events.

Structured Data

Examples of research questions from historical research that make use of structured data and tools that are the focus of this kit are the following:

CLARIAH will contribute significantly to enhancing our understanding of the origins, causes and character of the process of global inequality. A research question (from the research by Prof Jan Luiten van Zanden, 2003 Spinoza laureate, 2011 Academy Professor Prize KNAW) such as: 'How is the *creation* of economic wealth connected to its unequal distribution?' can only be answered by using large and collectively created global datasets. Global inequality is one of the key problems of the contemporary world. New theoretical insights in economics, such as new institutional economics, new economic geography and new growth theory, as well as the rise of global economic and labour history, mean that these processes can and should be studied on a world scale. These issues can only be tackled on the basis of global datasets for analysing patterns of economic performance and their causes. These global datasets are now being created in collaborative projects such as Clio-Infra, and CLARIAH will incorporate its results (data, tools, procedures, protocols, collaborative modes of working) and contribute significantly to extending the datasets.

A second example where structured data play a crucial role is related to the Historical Microdata Population Community. The crucial research question here is: how do societies change over time? Such a question can be addressed using information on changes in life courses, which provide a unique insight in this matter. They make it possible to analyse the nature, timing and relative importance of the constituent elements of social change. The micro data revolution that has reshaped scientific research in the social sciences during the last forty years has in the last two decades also reached historical research. Databases with individual level data have transformed historical research of societal change. The Historical Sample of the Netherlands (HSN[11]), coordinated by the IISG, collects this kind of data for the Netherlands. Although far from complete, the HSN-database has already been very successful. It has provided the basis for an ever–growing number of studies, often published in high–impact journals. The HSN has put the Netherlands in a unique position in the international research world. The database has grown into one of the most important ones in its field and plays a major role in the development of new methodologies for increasing the comparability of longitudinal data. In 2010 the HSN was the first Dutch database to be awarded with the DANS data prize.
The proposed extensions of the databases will serve two main functions. First, it allows new, *in depth research* into economics, health, mobility and kinship for a really long-term, allowing us to compare the pre-industrial period with industrialization. Second, by studying the connections between the developments in these fields, and using the material from the HSN, which is already available, we are able to provide a new *synthesis* of Dutch society in the past two hundred years.

Audio-Visual

Audio-visual data and tools to work with these data are important for conducting research in a range of humanities disciplines. The following are examples of humanities disciplines using audio-visual data: Linguistics (for language documentation, acquisition, impairment and attrition), anthropology (as their primary data), phonetics (for the analysis of speech), Sign Language (as an essential data source), history (as an important source of empirical evidence (oral history)), archaeology (for the documentation of their sites and for analyses of the recorded objects),

[11] www.iisg.nl/~hsn

art history (e.g. for analysis of paintings via processing high resolution images), and media studies (semantic video analysis, e.g. rhetorical analysis of mainstream media using automatic concept detection).

These disciplines are well-represented in the consortium, e.g. Descriptive linguistics, anthropologists and ethnologists at UL, RUN, and MPI; Sign Language Studies at RUN and UvA; phonetics and linguistics broadly construed) at UU, UvA, UG, UL, UvT, MI, FA; historians at Huygens ING, IISG, EUR, NIOD, UU; language acquisition, impairment and/or attrition at UG, UU, RUN and MPI. Many of the consortium members worked with cultural heritage institutes in the CATCH programme (e.g. KB, MI, RUN, UvA, UT, UvT) and MI co-ordinates the successor CATCHPlus project.

Many research questions in these disciplines can only be addressed effectively if there are sufficient audio-visual data that are richly annotated and if there are tools to explore, process, analyse and enrich such data. We will illustrate this using two disciplines: language acquisition and oral history. Such questions, however, can also be produced for the other disciplines.

Language Acquisition: The first seven years of a child are a crucial period to acquire language. The only way to obtain data of this period is by audio and video recordings. If there are sufficient data, if they are richly annotated and easily explorable, if they can be combined to broaden the empirical base and establish unexpected connections, then research questions such as: 'Which aspects of language acquisition are genetically determined and which ones are learned?', 'Are there default grammatical categories for semantic concepts and what role does this play in language acquisition (Semantic Bootstrapping Hypothesis)?', 'Does indirect negative evidence (absence of specific phenomena in a critical period) play a role in language acquisition?', 'How does frequency of specific phenomena play a role in parameter setting, if at all?', 'Is the way grammatical properties of words are established dependent on the acquisition stage?' can be addressed in a much more efficient manner than currently possible.

Oral History: In the case of oral history, using the example of recorded interviews with war veterans, research questions such as 'How are metaphors used by respondents to describe military issues?', 'How do respondents react to moral dilemmas?', 'What is the relation of descriptions of respondents ('oral history') to 'official' history?', 'How are traumatic events related to war formulated?' can be addressed in a much more efficient and systematic manner than is currently possible.

The Audio-visual Kit will facilitate the production of new data needed for research (making it easier and more efficient) so that future research projects can increase the focus on the actual research questions instead of collecting. Moreover, the increased interoperability of the various audio-visual data makes it possible to use much more data. Even data originally recorded for a different discipline, may be used. An example is a set of Oral History interviews with nuns in their cloisters during WWII in the Netherlands. The recordings, although meant for historic research, may be used for dialectological research as well.

The Audio-visual Kit will make it possible to carry out *Data Intensive Science* and will therefore boost the research in the disciplines mentioned, which may lead to breakthroughs in the disciplines mentioned above.

It will do so especially because it will make it possible to tap from a larger data source than ever before: the information contained in unstructured audio-visual data will become available to the researchers. In addition, tools will make it possible to automatically analyse, pre-process and aggregate the information contained in such data, turning these disciplines into true examples of humanities eScience. It thereby facilitates the interpretation process and boosts the opportunities for discovering unexpected relations.

Audio-visual data are large, and processing them is compute-intensive. This requires special facilities and organization that will be discussed in more detail in section 1.4. But the seamless inclusion of large-scale compute facilities will enable researchers to carry out compute-intensive operations without them even being aware of it.

1.3 Impact on specific science fields at international level

The CLARIAH project is part of the European CLARIN and DARIAH initiatives, which are widely supported in Europe and even beyond Europe. The methodologies and techniques developed in these projects will become *de facto*

standards for conducting humanities eScience. With the CLARIAH project the Netherlands will not only fully participate in these developments, but – given its leading position in the CLARIN and DARIAH projects - will be a driving force behind these developments.

The infrastructure envisaged will enable *Data Intensive Science* in the humanities across national borders. The facility will enable inclusion of data and tools from other countries since many will participate in Digital Object Architectures and use accepted standards promoted by CLARIN, DARIAH and EUDAT. The trend towards improved mappings at the semantic level based on agreed standards (semantic interoperability) will enable cross-corpus (hence even more data intensive) science and thus creates opportunities for relating different phenomena in new and unexpected ways.

The structured data collections and the tools to analyse and visualize them, belong internationally to the top. At the macro level, the databases brought together on world economic development have an impact on global economic history as a discipline, and Dutch researchers take a leading position in this field. At the micro level, the data on individual life courses have a transforming impact on how social history and historical demography are being carried out. Also the methods for harmonizing the data are exemplary. The databases are also having an impact on neighbouring fields, for instance on (historical) epidemiology and genetic research, linking historical population data to information on the spread of genetic diseases.

With the envisaged infrastructure, the science conducted will also be raised to a higher quality level because claimed research results can for the first time be replicated systematically in independent verification experiments.

In short, with the CLARIN and DARIAH initiatives, the character of humanities research is going to change dramatically, and forever. The Netherlands can play an important and leading role in these developments with the CLARIAH project.

1.4 Required and complementary expertise, embedding with expertise of partner institutes

We first list required expertise for CLARIAH as a whole, and then discuss required expertise specific for each resource type.

First of all, CLARIAH requires expertise in general infrastructural functionality, such as resource curation which includes making data and tools visible, accessible in a persistent way and syntactically and semantically interoperable, and ensure sustainability of data and tools. For these aspects the expertise is abundantly present and it has been further developed largely in the CLARIN-prep and CLARIN-NL projects with heavy involvement of MPI, MI, UU, DANS, INL and others. Some CLARIAH partners participate in European projects such as EUDAT and DASISH[12] and the *DAITF* initiative that is in preparation.[13] CLARIAH will of course closely collaborate with these projects and reuse their results. In short, Dutch institutes are present in the core international infrastructure initiatives creating momentum to strengthen the situation nationally as well via the CLARIAH project.

Second, expertise on data dissemination, publication and visualization is needed and also abundantly present in the consortium (DANS, MI, MPI, INL, KB, B&G and others). Some work on visualisation maybe outsourced (see below).

Third, large scale data and computing techniques (like cloud computing and grids) require expertise from specialized technology organisations such as the Netherlands eScience Centre (partner in CLARIAH), SURF (eInfrastructure), Nikhef, and SARA, continuing and extending the excellent collaboration with these organizations already initiated in CLARIN-NL and DARIAH. The nature of certain data (esp. audio-visual data) and the tools to process them probably warrant the use of somewhat further evolved services architecture for data processing than was thus far promoted in CLARIN. Special arrangements with the organizations mentioned are therefore particularly important for this type of resources. The CLARIN-NL pilot project with BiG Grid testing dynamic deployment of web-services can serve as a

[12] DASISH: http://www.lat-mpi.eu/latnews/tag/eudat/ with participation by DANS and MPI

[13] A worldwide Data Access and Interoperability Task Force initiative with participation of MPI and SARA from the Netherlands.

starting point for solving this problem. The EUDAT project is also highly relevant in this context and cooperation with this project, using their intended development of 'standardised' access APIs and procedures looks promising for the CLARIAH case. SARA and MPI take part in this project.

Text Kit
The implementation of the Text Kit requires extensive knowledge and expertise of historic and textual scholarship methodology, abundantly available within the humanities faculties of universities and in the KNAW institutions in the consortium. The libraries (KB, university libraries) possess specific expertise in the organisation and techniques of text conversion and structuring.[14] It also requires expertise on the social shaping and embedding of (digital) technology provided through the KNAW eHumanities program. The essential knowledge in linguistic computation (NLP), semantics modelling, digital textual representation, open annotation, linked data, and digital text analysis is provided through partners such as UvA, UvT, RUG, UU, VU, Huygens ING, MPI, the Computational Humanities program of the KNAW, and others. We will seek active collaboration with an external partner such as NBIC (Nanopub, BioSemantics) to exchange expertise on semantic modelling for textual research resources.

Structured Data Kit
First, expertise is required on databases, data structuring and data modelling. Second, expertise is required on data processing techniques, including efficient data conversion methods (and organisation of the conversion process), harmonization, data cleaning, treatment of fuzzy data, error correction and interpolation, data integration and semantic or 'linked data' techniques. Since the historical domain is the primary application area, knowledge of historical sources is required. The core partners focusing on this Kit, such as the IISG, UU and DANS jointly possess most of the needed expertise. Software development work for production systems will probably be outsourced to preferred external developers. Among these are Gapminder (Sweden) for visualisation of multidimensional data in graphs and Statplanet (Australia) for geospatial visualisations.

Audio-Visual Kit
The Audio-Visual Kit requires, first of all, expertise in the areas of recording and processing of audio, video and images. This expertise is abundantly present at the partners MPI, UT, B&G, RUN, UvT and DANS, who participated in a wide range of relevant projects.

The activities for the Audio-visual Kit will extend the activities in this domain already initiated at a very small scale in CLARIN-NL. It will nicely complement research activities in these areas, e.g. the AVATech project[15], a consortium of MPI Nijmegen and Fraunhofer institutes in Germany in which technology is being developed for semi-automatic annotation of audio and video recordings; the project *Socially enriched access to cultural media* of the COMMIT programme[16] in which B&G is a major partner and in which solutions are provided to enrich collections using Internet content curation and access through search engines and advanced recommendations. It will also build forth on the data and tools developed in the CATCH projects[17], and obviously cooperate closely with the CATCHPlus project[18] (Meertens Institute), as is currently already being done in CLARIN-NL.

2 Talent Case

The instruments and data offered by the CLARIAH infrastructure will act as a Pole of Attraction for international experts in the field of interpretation and analysis technologies. The Netherlands is at the forefront of international

[14] E.g. KB co-ordinates the EU IMPACT Project on this topic. http://www.impact-project.eu/
[15] http://www.mpi.nl/avatech
[16] http://www.commit-nl.nl/Messiaen53.htm
[17] http://www.nwo.nl/catch
[18] http://www.catchplus.nl/

eHumanities and is already attracting talented scholars from abroad. CLARIAH will further establish the Netherlands as a major player in eHumanities.

An intensive programme of dissemination, training and education of new students in using data and tools of the Common Lab will accompany the infrastructure. CLARIAH will continue and expand the efforts initiated by CLARIN-NL to make courses for working in this infrastructure part of the regular humanities curricula. As virtually all humanities faculties are participating, CLARIAH offers a unique chance to collaborate and specialize in offering innovative teaching and training courses. The support for CLARIAH by graduate research schools such as the Netherlands Graduate School of Linguistics (LOT), N.W. Posthumus Institute and the Netherlands Graduate School of Science, Technology and Modern Culture (WTMC) underlines their eagerness to incorporate new Common Lab technologies in the graduate programmes.

The infrastructure will therefore create a whole new generation of researchers who become familiar with these innovative *eScience* methods at a very early stage, thus creating the best potential for achieving scientific breakthroughs and excellent career opportunities. Having such an attractive infrastructure and its experts in the Netherlands will contribute to keeping talented new researchers in the Netherlands and to attracting top researchers from abroad to our country.

In the budget, we have set apart a significant sum to stimulate such *brain gain* actively, by inviting leading experts from abroad and by creating openings for foreign PhD students and trainees. This will allow us to closely collaborate with foreign top researchers and infrastructure specialists, which will lead to a great deal of synergy, and to an improved Dutch research environment, attractive for other students and researchers. In this way CLARIAH will build human capacity, and be of particular benefit for early stage researchers. As the Netherlands has a leading position in CLARIN-EU and DARIAH-EU, researchers and students are eager to join Dutch research groups.

Though the location where research is carried out in terms of access to data and services is made irrelevant by the distributed nature of the Common Lab infrastructure, Dutch research groups are intellectual growth poles because we are in the lead in designing, constructing and exploiting the infrastructure, and will be the first to have built up significant expertise in using it. The communities in the humanities will be directly involved in developing the courses needed. Users are already creating virtual collections combining resources from different institutes and performing operations on them, as well as integrating various tools to new applications by workflow mechanisms. Such improved mechanisms of accessibility will attract young researchers and foster innovative humanities research in the Netherlands,

We will continue to seek additional funding through programmes such as the Marie Curie programme (International Training Networks), as well as in regular calls in the Humanities, ICT and Research Infrastructure Programmes of the 7th and 8th Framework Programme of the EC and of the ESF. How this works is illustrated by the following examples, focusing on several humanities user communities. CLARIAH will both benefit from and build on these examples:

a) The Network for Digital Methods in the Arts and Humanities (NeDiMAH), which just started in May 2011, is working closely with DARIAH-EU and CLARIN-EU to examine the practice of advanced ICT methods in the arts and humanities across Europe, articulating these findings in a series of training outputs and publications. To accomplish this, NeDiMAH is providing a locus of networking and interdisciplinary exchange of expertise among the trans-European community of digital arts and humanities researchers, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. The programme is bringing together practitioners in a series of thematic Working Groups, which will examine the use of formal computationally-based methods for the capture, investigation, analysis, study, modelling, presentation, dissemination, publication and evaluation of arts and humanities materials for research.

b) The active role of the community in educational activities to train a new generation of researchers in working with the language resource infrastructure of CLARIN is clear from a range of events:

- MPI already participates in the Marie Curie CLARA programme, which organizes several educational activities to train a new generation of researchers who will be able to cooperate across national boundaries on the

establishment of a common language resources infrastructure and its exploitation. MPI hosted the 2010 and 2011 CLARA summer schools in the Netherlands.[19]

- CLARIN-NL has organized many tutorials and workshops that attracted international participants even though they were mainly intended for researchers working in the Netherlands.
- Several seasonal school lectures were held in which the CLARIN-NL infrastructure played an important role.

c) An educational outreach and training programme for young scholars, based on the Clio-Infra facilities is already part of the proposal funded by Large NWO Investments. Training will target graduate and post-graduate level researchers who wish to advance, refine and apply data in social science history. The new facilities offered by the Clio-Infra infrastructure (such as new, interlinked, global datasets on world economic development) will moreover attract and challenge high-quality students to test new theories and develop new perspectives on European and global developments. Training facilities under the title of 'Advanced Seminars in Data Archiving and Publication Strategies' will be offered in conjunction with existing networks for PhD students in economic and social history (Research Training Network – RTN; ESF Research Networking Programme GLOBALEURONET; European graduate school ESTER). This will attract Dutch and foreign students interested in working with large datasets.

d) At the national level, the Historical Sample of the Netherlands has demonstrated its power to attract and organize multidisciplinary collaborations between social scientists, historians, demographers, geographers, economists and others on a scale unprecedented in this area. The unique character of the database with *complete* life histories representative at the national level, and covering two centuries from 1812 till the present, has been attracting researchers from within and outside the Netherlands. The quality of the data and of the research community working with them will create a stimulating environment for more high-quality scholars, students and post–docs interested in working with longitudinal data. Training facilities will be built for them in cooperation with existing training centres like the Interuniversity Consortium for Political and Social Research in Ann Arbor, Michigan (ICPSR) and the graduate courses of the NW Posthumus Institute.

3 Innovation Case

3.1 Contribution to government priorities

The core instrumentation of the CLARIAH Common Lab is directly relevant to the Top Sectors 'High Tech' and 'Creative Industry'. The humanities are traditionally not seen as instrumental for industrial progress. Yet the interest of high tech and new media companies to collaborate with and to support the CLARIAH proposal is remarkable. We will set up user groups and reflection panels in which private companies, public sector organisations and eHumanities researchers will participate.

The development of intelligent, self-learning systems is one of the innovative directions in ICT within the Top Sector *High Tech*. Several CLARIAH partners are participating in the COMMIT programme, which brings together leading researchers in search engines, parallel computing, databases, interaction in context, embedded systems and knowledge technology. Microsoft Research sees that '4th paradigm' data-driven science extends to the humanities as well. Open innovation networks are globally emerging as a highly effective approach towards innovation and exploitation of eScience/eHumanities and technology.

The involvement of the Netherlands eScience Center in CLARIAH guarantees the cross-fertilisation of 'humanists' and 'scientists' in innovative approaches to the tackling of complex data issues. Humanities scholars have a long tradition in working with complex or even 'messy' data, meaning that the sources available are often incomplete, inaccurate, unstructured, vague, contradictory or sometimes false. The rapid extraction of maximal information content from such complex data sources and assessing their explanatory power in intricate relationships, while reducing noise to a minimum, is a challenge not only for humanities scholars, but is also one of the frontiers of IT

[19] http://clara.uib.no/

development in big companies such as IBM. IBM has expressed its explicit interest in collaborating in this area with eHumanities researchers.

3.2 New products and services

Humanities knowledge is more often turned into commercial products than most people think: many books, documentary films, TV-programmes, games and a lot of tourist information are based on humanities research. The CLARIAH infrastructure will use elements of gamification to present its results in attractive and playful ways, thereby seeking collaboration with key players in the *Creative Industry* (including new media companies and publishers), a second Top Sector of the Ministry of EL&I. Games and crowdsourcing foster new ways of interaction between cultural heritage and the public, and will improve the quality of research data, for instance by asking the public to validate or improve OCR output. New paradigms for improved content access will arise by integrating machine labelling, human input, social interaction (gamification) and semantic links between media items.

CLARIAH actively contributes to the open data and open source policy of the Dutch Government and international organizations, such as the OECD, whose *Declaration on access to research data from public funding* (Paris, 30 January 2004) offers additional arguments for the CLARIAH innovation case. Digitale Agenda.NL of the Ministry of EL&I sees open data as fuel for growth and innovation. An optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation. Open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers. Open access will maximise the value derived from public investments in data collection and digitization efforts. New business models for publishing companies are emerging, but CLARIAH will also provide tools for linking data to articles (supporting persistent identifiers and digital author identifiers) that follow more traditional pathways in collaboration with publishing companies.

In CLARIAH, commercial spin-offs will be stimulated actively. First by means of a wide range of dissemination activities aimed at humanities and social science researchers, but also beyond these groups, to create awareness of the CLARIAH infrastructure and the potential it offers. We will welcome private sector users of the infrastructure, and help them use the facilities in the best possible way by providing courses and training sessions.

New research opportunities made possible by CLARIAH will enable the development of new products and services. To this end CLARIAH aims to continue and intensify the already existing contacts and fruitful collaboration with the IT-industry (e.g. in the STEVIN and CATCH programmes). This collaboration will consist of user-groups, meetings and small demonstration projects where academia and industry work together. The CLARIAH resources and infrastructure are important for Dutch companies, and several of them therefore support the CLARIAH proposal and/or have indicated their willingness to contribute actively in user groups. See the full list in the General Information section and the letters of support attached to this proposal.

Many tools that are being incorporated in the CLARIN infrastructure and that will become part of the CLARIAH infrastructure can and will be used in commercial applications as well. For example, in the CLARIN-NL project 'War in Parliament', opinion mining and content analysis tools are being improved and a demonstrator is being created to investigate the attitude of parliament members towards war using the parliamentary minutes. The commercial potential of this application lies in the power to evaluate the public perception of for instance a company, a product or a political party. A variety of companies have expressed their interest in and support for this reason. They include big companies such as Philips, but also smaller software companies active in the area of opinion and sentiment mining such as BuzzCapture, Hippo, Trendlight, and TotalRequest, as well as companies that are active in language and speech technology such as GridLine, RightNow! (formerly Q-Go), TextKernel, and TeleCats.

In the past, spin-off companies (e.g. Textkernel, Telecats and Gridline) have successfully valorised the outcomes of language research, such as automatic speech recognition. The CLARIAH 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, including services for augmented reality. The CLARIAH infrastructure will provide the opportunity to develop and test such applications and services in the humanities domain, after which companies can exploit the ones that are most successful and

that appeal to a customer base. It can be expected that CLARIAH will spurn off additional start-up companies, as well as commercial and social innovations. Of course, even though Open Source Software will be the norm, IPR to new inventions need to be regulated and appropriate licenses will be formulated.

With clever combinations of meta-data and content-based access to multimedia productions, articles and web pages, it will be possible to develop new media products and enhanced publications that are of interest to various users, e.g. in publishing and (new) media. Initial steps in this direction have already been taken in several projects, for instance in the NWO CATCH programme[20].

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 by CLARIAH partners and supporters, such as B&G, KB, and publishers such as PCM.

Many people and groups in society will benefit from tools that reveal information hidden in textual and multimedia data. For instance, investment decisions in the stock market depend to a considerable extent on information about companies that is 'hidden' in reports. The eScience tools created in CLARIAH will provide new opportunities to unlock information hidden in such textual and multimedia content.

4 Partnership case

4.1 Positioning of the facility in the (inter)national research landscape

The partners in CLARIAH cover practically the whole of humanities Netherlands: all universities with a humanities faculty are represented, and so are all humanities institutes of KNAW, as well as the other important humanities institutes (MPI for Psycholinguistics, INL), libraries and data centres (KB, NA, B&G). The participation of NLeSC will make sure that eHumanities and eScience will closely cooperate. Many related organizations and selected private companies from Top Sectors support the proposal.

The CLARIAH initiative, as a national counterpart to CLARIN and DARIAH-EU, is directly and deeply embedded in the Europe-wide ESFRI enterprise. CLARIN and DARIAH are the only two humanities infrastructures under development on the ESFRI Roadmap and are both on the 2008 national roadmap.

On the CLARIN side, this enterprise was initiated by the just finished CLARIN preparatory project (CLARIN-prep[21]) and is to be continued by a consortium of national infrastructure projects united at the European level in the so-called CLARIN ERIC[22], A formal request to establish the CLARIN ERIC is underway and is expected to start early 2012. The Netherlands played an important role in CLARIN-prep, and the CLARIN ERIC will be hosted by the Netherlands.

There is wide support for CLARIN in Europe and beyond: in CLARIN-prep 36 members from 26 EU member and associated states participated and the CLARIN network encompasses 208 organisations from 33 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.

On the DARIAH side, The Netherlands was also the coordinator of the 'Preparing DARIAH' project, in which 14 partner organizations from 10 countries originally participated[23]. Also DARIAH is currently in the transition from preparation to construction. Together with Germany and France, The Netherlands will play a leading role in the

[20] http://www.nwo.nl/nwohome.nsf/pages/NWOA_7ANC86
[21] www.clarin.eu
[22] ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric
[23] www.dariah.eu

construction. France and Holland will jointly be responsible for one of the four 'Virtual Competency Centres' of DARIAH-EU, viz. the one on Scholarly Content Management (long-term preservation of and access to humanities resources). At present, 13 countries have indicated they want to join the DARIAH ERIC by signing a Memorandaum of Understanding. It is expected that the ERIC proposal will be submitted this autumn, and that construction will start in spring 2012.

There has been close collaboration, both at the national and at the international level between the CLARIN and the DARIAH projects, witnessed by jointly organised conferences such as 'Networking European Research Infrastructures' and 'Supporting Digital Humanities'.[24] For many features highly similar approaches are needed. It is therefore natural to join forces, and to strengthen each other in a common project such as CLARIAH. The DASISH project (Data Service Infrastructure for the Social Sciences and Humanities) is preparing for further collaboration between the humanities and social science ESFRI projects, such as CESSDA (Council of European Social Science Data Archives).

There is no space here to describe the networks, nationally and internationally, in which the (digital) humanities communities that CLARIAH is building on and will serve, participate. We limit ourselves to just a few:

- There has been close collaboration with the *Fostering Language Resources Network* (FlareNet, http://www.ilc.cnr.it/flarenet/; 2008-2011).
- The Clio-Infra and HSN communities play influential roles in the most important economic and social history networks.
- The exchange facility META-SHARE is currently being set-up in the EU META-NET project. The META-SHARE exchange facility is targeted at academic and industrial researchers and developers of language and speech technology, and focuses on facilities for *research into and development of* these technologies.
- The Dutch partners in CLARIN-EU and DARIAH-EU participate in a number of related EU-funded projects, such as EHRI (European Holocaust Research Infrastructure)
- CLARIAH members play a leading role or participate in Europeana and related digital heritage projects, which offer a multi-lingual online collection of millions of digitized items from European museums, libraries, archives and multi-media collections.
- Of course immediate connections exist with our American counterparts in 'cyber humanities', such as Project Bamboo, which is a multi-institutional, interdisciplinary effort that brings together humanities scholars, librarians, and information technologists to tackle the question: *'How can we advance arts and humanities research through the development of shared technology services?'*

4.2 Organizational structure and governance of the facility

The **General Assembly (GA)** will be the highest governing body of CLARIAH and will consist of one representative of each consortium member. The GA will decide on the yearly budget and work plan, thus having the power to approve the general CLARIAH policy. It will meet twice a year. As in CLARIN-NL, a range of subprojects will be started up in CLARIAH in the form of calls, tenders, or direct assignments to specific (groups of) participants prepared by the Board of Directors and decided upon by the GA. Special measures and procedures have already been set up and successfully applied in CLARIN-NL to avoid conflicts of interests, which will be adopted in CLARIAH. The GA will have the option to set up committees if needed with participation of external experts, e.g. to organize special events or to investigate specific matters.

CLARIAH will have a **General Director** for overall management of the project, including WP Management & Coordination.

Each Technical Work Package will have a **WP leader**. They will be directed by the CLARIAH **Technical Director** (or CTO), whose prime responsibility it is that there will be technical cohesion and coordination among these WPs.

[24] http://www.csc.fi/english/pages/neeri09; http://ztwweb.trans.univie.ac.at/sdh2010. For 2011 a new SDH is planned: http://cst.ku.dk/sdh2011/

The **Outreach Director** will lead the Dissemination & Outreach WP and will be responsible for the interaction of DARIAH with the user communities (of humanities researchers as well as interested companies) through **Users Panels & Reflection Groups**. These panels will act as an important source for advice on the priorities for data and tools needed, and on the valorisation of CLARIAH outputs. They will consist of researchers (humanities, social and other science) and of representatives from public organizations (cultural heritage, care & safety, public administration) and from private enterprises (ICT, creative industry, media and publishers).

The three directors together form the **Board of Directors**, which is responsible for day-to-day running of the CLARIAH facility. They will have regular meetings with the Dutch DARIAH-EU director and with the CLARIN-EU director. The CLARIAH office will be co-located with the CLARIN office, providing secretarial and administrative support.

CLARIAH will ensure that its experts will participate in various working groups that are working at the CLARIN and DARIAH-EU level to bring in the Dutch expertise and to ensure suitability of design decisions to the needs of Dutch researchers, but also to guide the discussions at the national level.

Finally, an **International Advisory Panel** will advise both the CLARIAH General Assembly and the Board of Directors**.**

After the CLARIAH proposal is awarded funding, CLARIAH and CLARIN-NL will discuss with NWO how to formalize the mutual cooperation (e.g. by partial or complete merger) in order to maximize the synergy.

The CLARIAH governance structure is graphically depicted in Figure 2.

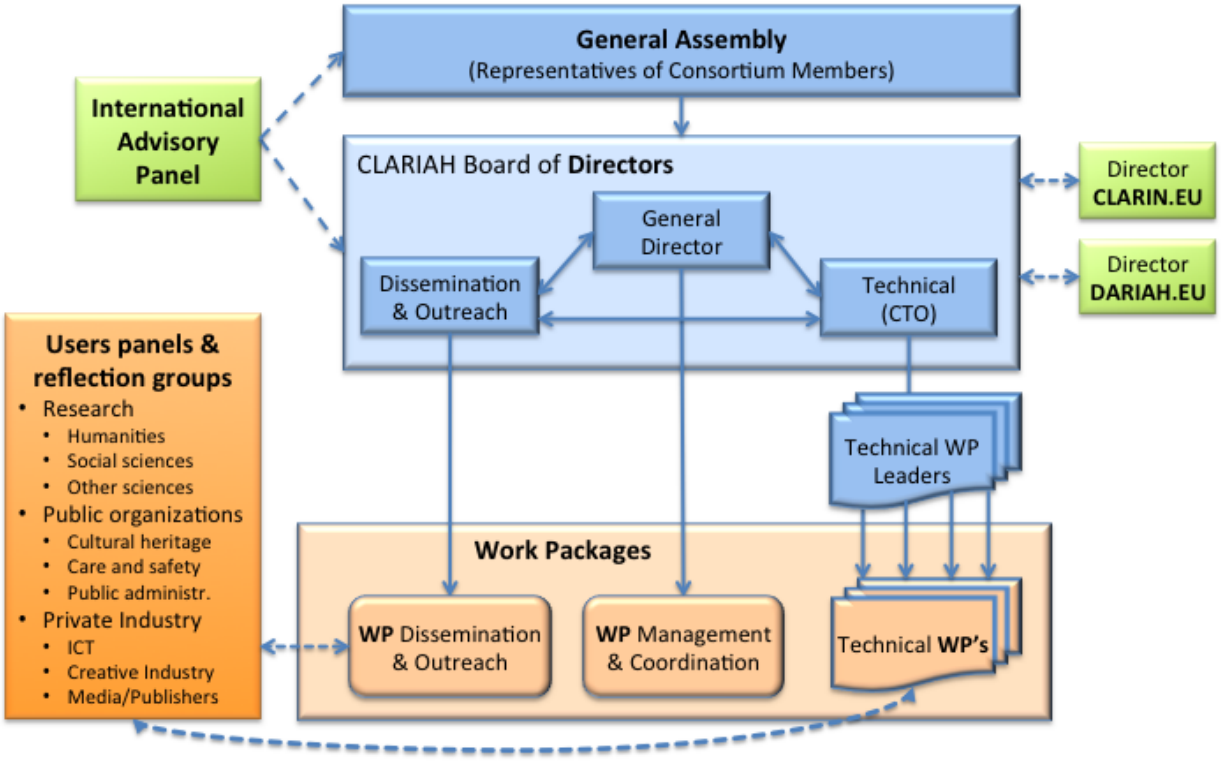


Figure 2. CLARIAH organisation

5 Business case

5.1 Total Costs

In the budget below we made a calculation according to the categories distinguished in the NWO standard form for year 1-5 (2012-2017) and for year 6-10 (2017-2022). Calculations per Work Package are given for the first five years only.

Budget according to NWO categories, year 1-5

Year 1-5 (2012-2017) costs (Euro*1000)		***YEAR 1 2012/13***	***YEAR 2 2013/14***	***YEAR 3 2014/15***	***YEAR 4 2015/16***	***YEAR 5 2016/17***	***TOTAL yr 1-5***
Construction and investment costs		€ 1875	€ 2625	€ 2500	€ 2600	€ 1950	€ 11550
Running costs	Maintenance	€ 125	€ 210	€ 350	€ 520	€ 650	€ 1855
	Upgrades	€ 0	€ 175	€ 500	€ 975	€ 1300	€ 2950
	Personnel costs	€ 375	€ 315	€ 1400	€ 2080	€ 2275	€ 6445
Decommissioning		N/A	N/A	N/A	N/A	N/A	N/A
Other costs		€ 125	€ 175	€ 250	€ 325	€ 325	€ 1200
Total / year		**€ 2500**	**€ 3500**	**€ 5000**	**€ 6500**	**€ 6500**	**€ 24000**

The budget is expected to grow from M€ 2,5 in the first year to reach a maximum of M€ 6,5 in year 4 and 5, after which year operation costs will dominate construction and it will go down to M€ 4 per year from year 8 onwards. The construction and investment costs will amount to 75% of the total budget in the first two years, and after that gradually go down to about 30% in 2016/17. The running and maintenance costs will already start at about 20% in the first year, as CLARIAH builds on existing components that need maintenance. The maintenance costs will increase to 65% by the end of the first five years. Personnel costs are a substantial component of this. Other material costs (travel and subsistence, publication costs, conference & workshop organisation, invitations of international top specialists) are estimated to amount to 5% of the total budget.

Budget according to NWO categories, year 6-10

Year 6-10 (2017-2022) costs (Euro*1000)		***YEAR 6 2017/18***	***YEAR 7 2018/19***	***YEAR 8 2019/20***	***YEAR 9 2020/21***	***YEAR 10 2021/22***	***TOTAL yr 6-10***
Construction and investment costs		€ 1100	€ 675	€ 400	€ 400	€ 200	€ 2775
Running costs	Maintenance	€ 550	€ 450	€ 400	€ 400	€ 400	€ 2200
	Upgrades	€ 1100	€ 900	€ 800	€ 800	€ 800	€ 4400
	Personnel costs	€ 2475	€ 2250	€ 2200	€ 2200	€ 2400	€ 11525
Decommissioning		N/A	N/A	N/A	N/A	N/A	N/A
Other costs		€ 275	€ 225	€ 200	€ 200	€ 200	€ 1100
Total / year		**€ 5500**	**€ 4500**	**€ 4000**	**€ 4000**	**€ 4000**	**€ 22000**

The management and coordination costs are budgeted at 8% of the total, the dissemination and outreach at 5% of the total budget. This leaves 87% of the total budget available for the technical work packages, which form the six methodological layers of CLARIAH. A division is also made according to the three Kits (Textual, Structured Data and Audio-visual), which are approximately of equal size (table not included here).

Budget according to Work Packages

Year 1-5 (2012-2016/17) costs (Euro*1000)	***YEAR 1 2012/13***	***YEAR 2 2013/14***	***YEAR 3 2014/15***	***YEAR 4 2015/16***	***YEAR 5 2016/17***	***TOTAL YR 1-5***
Management & Coordination	€ 200	€ 280	€ 400	€ 520	€ 520	€ 1920
Data Conversion	€ 392	€ 548	€ 782	€ 1018	€ 1018	€ 3758
Data Enrichment	€ 326	€ 457	€ 653	€ 848	€ 848	€ 3132
Linking Resources	€ 435	€ 609	€ 870	€ 1131	€ 1131	€ 4176
Pattern recognition	€ 413	€ 579	€ 827	€ 1074	€ 1074	€ 3967
Visualisation & Presentation	€ 348	€ 487	€ 696	€ 905	€ 905	€ 3341
Long-term preservation	€ 261	€ 365	€ 522	€ 679	€ 679	€ 2506
Dissemination & Outreach	€ 125	€ 175	€ 250	€ 325	€ 325	€ 1200
Total / year	**€ 2500**	**€ 3500**	**€ 5000**	**€ 6500**	**€ 6500**	**€ 24000**

Over the first five years, the budget for the six technical work packages totals about M€ 20. Linking resources, Pattern recognition and Data conversion are the biggest work packages, taking up on average between 18-20% of this amount. Data enrichment and Visualisation and presentation consume 15-16% of the budget; finally, Long-term preservation takes 12% of the total budget.

5.2 Non-NWO contributions

It is *not practical* to give an overview of the various non-NWO contributions in the tabular format and categories of the proposal form. We distinguish here the main investment flows and grants acquired for the various components that are being integrated into the CLARIAH infrastructure: CLARIN-NL, CLARIN-EU, DARIAH-EU, Clio-Infra, HSN, and eHumanities. Moreover, it is not always feasible to disentangle the national component from EC grants and investments (especially for CLARIN-EU and DARIAH-EU). We draw a distinction here between the period 2008-2011 and 2012-2016. The estimated investments in the period 2008-2011 in CLARIAH constituents was M€ 11, and in 2012-2016 non-NWO contributions will amount to at least M€ 20, excluding overheads. The overhead costs of CLARIAH will in addition be covered by the participating organizations; these costs include office and workspace costs, office automation, and administrative support.

Period 2008-2011

- CLARIN-NL: in 2009 and 2010 CLARIN-NL spent M€ 1.4. In 2011 the CLARIN expenditures are budgeted at M€ 3.1, bringing the total for 2009-2011 at M€ 4,5.
- CLARIN-EU: The European CLARIN-prep project ran from 2008 to July 2011 and had a budget of M€ 4.1, K€ 841 of which was assigned to the Netherlands. The total investment for local CLARIN activities in other countries than the Netherlands amounted to M€ 5.3. In addition, M€ 4.9 was invested in CLARIN-related projects.
- DARIAH-EU: EC-Funding (M€ 2,5) from Brussels for the preparation of DARIAH-EU in the period 2008-2011. DARIAH members contributed an additional 1,0 M€ from other means during this period. An estimated K€ 750 of this amount was spent for the Dutch coordination by DANS. Two European DARIAH-related infrastructure projects were granted: European Holocaust Research Infrastructure (EHRI) and Data Service Infrastructure for the Social Sciences and Humanities (DASISH), in which also CLARIN participates. The majority of the budget of these projects is however for the period 2012-2016, and the amounts are mentioned there.
- Clio-Infra: K€ 735 from NWO Large investments is spent in 2011, the remainder in 2012-2016. Van Zanden also invested the lion's share of his Spinoza prize 2003 (M€ 2,5) in the development of Clio-Infra.
- HSN: The direct investment of IISG over 2008-2011 adds up to K€ 300 (excluding overheads). The grants received for HSN-related projects since 2008 total M€ 2.6, of which M€ 1.2 million was spent in the period 2008-2011.

- eHumanities: The KNAW invested M€ 1.5 in the development of Alfalab (2009-2011), including M€ 0.7 by the participating institutes. The additional investment in eHumanities will largely be spent in the next years and are mentioned below.

Period 2012-2016

- CLARIN-NL: The budget for CLARIN-NL allotted from the National Roadmap 2008 for the period 2012-2014 is M€ 4.6.
- CLARIN-EU: The Dutch CLARIN-EU contribution is part of the CLARIN-NL budget and will amount to M€ 1.3 over the period 2012-2016 (ca. K€ 250-270 per year). The CLARIN related project EUropean DATa (EUDAT, Oct 2011 – Oct 2014), has a budget of M€ 9.3, of which M€ 1.1 has been assigned to the Netherlands.
- DARIAH-EU: The DARIAH-EU construction phase is expected to start in 2012 with a total budget of around M€ 4 annually, plus M€ 2 for community engagement projects. The Dutch participation is estimated to amount to K€ 408 yearly, of which KNAW and NWO Humanities division contribute 50 K€ per year for a three-year period. The DARIAH-related infrastructure project EHRI has a budget of M€ 7.0, M€ 2.5 of which is being spent in The Netherlands. DASISH has a total budget of M€ 8.1 for 2012-2014, K€ 400 of which will be spent in The Netherlands.
- Clio-Infra: M€ 3.2 of the NWO Large investment will be spent in the period 2012-2016.
- HSN: direct investment of IISG over 2012-2016 will amount to K€ 330 (excluding overheads). M€ 1.3 on HSN-related projects will be spent in 2012-2016.
- At the beginning of 2011, KNAW made available M€ 4 for the establishment of a KNAW-wide e-Humanities Group, which is responsible for the implementation of the KNAW's computational humanities programme. Projects in computational humanities are the result of cooperations between KNAW institutes and universities, and also involve in-kind contributions by all.
- CLARIAH aims to make use of the existing and future national eScience infrastructures, such as BiG Grid and its successor eInfrastructure. The storage and processing costs of CLARIAH are estimated at K€ 250 on average over the first five years (K€ 1250 for five years). In the unlikely case the national eInfrastructure is not taking off or cannot be used by CLARIAH, the CLARIAH budget needs to accommodate the additional cost by reducing other budget lines.

5.3 Requested NWO Financing

The requested NWO funding is equal to the total budget, because the contributions from other sources have not been incorporated in the total budget due to their heterogeneous nature and because they formally consist of separate grants with distinct responsibilities.

6 Technical case

6.1 Technical feasibility and challenges

CLARIAH builds upon the infrastructure that is being developed in CLARIN and DARIAH. The European CLARIN preparatory project has set a lot of design decisions and has implemented parts of the infrastructure. In CLARIN-NL the development of this infrastructure has been continued and is being extended. The CLARIN centres in the Netherlands, which are cooperating to implement this, are all participants of the CLARIAH consortium. The national CLARIN project in Germany has just started and will contribute, and very likely other national projects will follow soon. At the European level, the EUDAT project with participation of SARA and MPI from the Netherlands will focus on professional and robust common infrastructures services.

This gives CLARIAH a head start, and it is clear that CLARIAH's activities are built on and embedded in a lot of other projects working on related topics, thus reducing the risks of technical challenges. Initial versions for a lot of the basic infrastructural services are already in place and in use. They include AAI functionality with Single-Sign-On via SAML2 based Federated Identity Management; services for persistent identifiers (PID) assignment and resolution; a

metadata framework (CMDI), with a registry and various editors; and the Data Category Registry ISOCAT for semantic interoperability.

An initial version of the metadata browsing functionality has been implemented in the Virtual Language Observatory[25], and functionality for searching in metadata and the actual data, distributed across multiple centres, is under development in CLARIN-NL. Several projects in Europe are working with systems in which web services are combined in flexibly adaptable work flows, inter alia in the CLARIN-NL TTNWW project (cooperating with Flanders).

The technical infrastructure will be implemented in selected CLARIAH-centres. Preferably use will be made of the emerging nationally integrated e-Infrastructure as a successor to BiG Grid as proposed by SURF. The storage and processing of large data collections on this e-Infrastructure creates opportunities for interoperable applications implemented within a service-oriented architecture, and may also be a solution for transporting and processing huge data amounts (e.g. processing audio-visual data).

The PID infrastructures in use by CLARIN and DARIAH and other partners (mainly Handle, DOI and URN:NBN based solutions) will be further integrated, offering means for all types of PIDs to be translated into the appropriate location.

Interoperability, and especially semantic interoperability, are key to the success of the Common Lab, and will therefore be central in the whole project. The CLARIN-NL project has already shown that achieving semantic interoperability is not easy but also that solutions can be found for the problems. Applying data category mapping from real data and tools used by researchers has already led to refinements, better guidelines, and the conception and set up of an additional registry RELCAT to represent relations between data categories.

A variety of demonstrators, developed in the context of the DARIAH preparation phase project, CLARIN-EU, CLARIN-NL, the KNAW Alfalab project, Clio-Infra and HSN offer building blocks for the integrated access to the resources in CLARIAH. Where appropriate these demonstrators will need to be extended to fit the wider common infrastructure. The demonstrators showcase how CLARIAH will provide an open system, which is modular enough to work in different research environments.

To prove the functionality of the infrastructure and to promote its use by Arts and Humanities researchers a mix of demonstrator and data curation projects will be carried out, solicited via open calls and/or calls for tender and evaluated by independent experts. In this way the functionality of the infrastructure is tested against data and tools in real use by humanities researchers, which can be incrementally adapted to suit their demand.

In CLARIAH new types of data and tools will be part of the infrastructure and they stem from a broader range of humanities disciplines. This will surely require extensions of the infrastructure and the standards it supports. The structured data and associated tools with their focus on the history domain are a clear example: the generic technical infrastructure common to all structured data hubs, such as tools for data annotation, documentation, storage, sharing and access, as well as communication facilities for discussion, will be integrated in the Common Lab. The existing software for analysing and disseminating data will be upgraded. The upgrade and integration will very likely require both an extension of initial CLARIAH standards and adaptations and extensions of the existing data, their metadata and the associated tools to ensure seamless interoperability.

The Common Lab will offer workspaces for (groups of) researchers as described in section 1 (Science Case). This requires authentication and authorization procedures in line with those developed within CLARIN and using SURFNet Federated Identity Management. Particular attention needs to be paid to the issue of user delegation, which is currently being addressed through projects such as GEANT and BiG Grid. For browsing and searching the facilities under construction in CLARIN will be used with additional attention to authorization aspects relevant for virtual workspaces. Here searching through content of heterogeneous workspace data remains an issue that needs to be addressed.

[25] www.clarin.eu/vlo

Management and operation of web services poses a challenge. There is a lot of expertise for management and use of large quantities of data, but doing this for web services especially with a view to efficient data transport, long-term persistency and security requires significant attention in the project.

IPR and Ethical issues must be taken into account. Though CLARIAH promotes open access to data and open source software, data and tools are available under different conditions. The facility will provide the means to properly deal with IPR and ethical restrictions, which involves organisational matters but also technical implementation (access rights, digital signing of license, etc.). Expertise among the CLARIAH partners is in place to support this.

6.2 Risk Analysis

Event	***L***	***S***	***R***	***Preventive actions (PA) and contingency plan (CP)***
Underestimation of effort needed for integration of data or services	4	2	8	PA: early sensing problems by checking deliverables and milestones; resettling allocations and priorities. CP: reallocation of tasks and resettling priorities by the board
Interoperability problems	4	4	16	PA: early sensing by continuous interactions through a dedicated coordinator; determining the reason of the problems and finding best practice solutions. Taking care of an up-to-date standard registry to guide implementations. CP: implement temporary best practice solutions and check whether priorities need to be re-established
IPR and ethical issues	4	3	12	PA: systematic attempts to arrange IPR and ethical issues properly (already successful twice in CLARIN-NL) CP: implement protocols and processes to properly handle IPR and Ethical Issues
Creating technical solutions with insufficient under-standing of the require-ments, and constructing large systems of the type 'one size fits none'	2	2	4	PA: Close cooperation between users and developers from the start; usage tests in early stages in the project; approach based on components (not monolithic); adoption of agile development methodologies CP: Analyse the problem; reassign tasks
Overtaken by new technologies	4	2	8	PA: Component-based approach reduces this risk to individual components CP: Replace the obsolete components
Heterogeneous require-ments of key enabling tech-nologies across partners and disciplines	4	4	16	PA: Strict prioritizing of technologies to apply, based on user demands CP: Reassign the priorities and stick to them
Deployment of key enabling technologies financially not feasible	4	2	8	PA: Monitor development and budget closely CP: Prioritise which technologies to implement; reduce ambitions
Availability of key enabling technologies	4	4	8	PA: Monitor availability and progress of technologies CP: Develop alternative strategies at an early stage, reduce ambitions
National ICT Infrastructure (e-Infra) for storage and processing does not emerge	3	4	12	PA: Actively support proposal for national e-Infrastructure; prepare for storage capacity and processing power elsewhere CP: Find additional funds to cover additional storage costs in partner budgets

Risks identified at the start of the project

The table above assesses the risk level of an event that may have a negative effect on the CLARIAH objectives by considering its likelihood and its impact. It also proposes preventive actions to avoid the risk and a contingency plan to minimise the negative impact, taking into account an assessments of the costs for the control measures.

For each major risk, the table describes the event, its likelihood (L), its impact (S), its risk level (R), its preventive actions and its contingency plan. Impact and likelihood range over the values 1 (low) through 5 (high); effect and

likelihood together result in the risk level by multiplying their values, resulting in a range of values between 1 and 25. Scores 13 and higher constitute serious threats to the project and should be monitored continuously by the project management.

A more detailed list of risks and control measures will be drawn up in the first phase of the project, when the detailed operational plan is prepared. However, the process of identifying and analysing risks, and then deciding on the appropriate course of action in order to minimise them, is a continuous process.

7 Possible focus for the Netherlands

The Netherlands is a leading partner in the infrastructures and collaborative projects directly relevant to CLARIAH. We provide three examples to support this.

First, Utrecht University and MPI play the leading role in CLARIN-EU and are recognized as such by the consortium partners and all network participants in the CLARIN initiative. The Netherlands will host the CLARIN ERIC. It therefore has an excellent opportunity to determine the CLARIN-EU profile. CLARIN-NL has enabled us to consolidate and further extend this unique driving position of the Netherlands in the European context.

Second, DANS, the Dutch national centre focusing on permanent access to data in the arts and humanities, initiated by KNAW and NWO in 2005, takes an internationally leading position in its field. The coordinator of DARIAH-EU in the preparation phase, it enabled the Netherlands to raise its European profile also in this domain. The Netherlands is currently at the forefront of repository developments (also cf. SURF with its digital academic repositories; KB with its E-depot).

Third, the Netherlands is the internationally recognized leader worldwide in the two historical research domains that play an important role in CLARIAH (research on family patterns and life courses, and research on social and economic history), esp. via IISG, one of the world's largest documentary and research institutions in the field of social history.

For research on family patterns and life courses, the HSN is worldwide recognized as one of the most important databases in its field, setting 'best practices', with no parallel of its size and quality in the world. The HSN involves an interdisciplinary collaborative effort of researchers in the humanities and social sciences. The Dutch research community has already invested substantially in the HSN facility in order to enable the kind of research that is high on the agenda of the Dutch scientific programme.

Concerning the research on social and economic history, Clio-Infra is a collaborative effort to create large global social-economic databases involving 10 partners. The Netherlands plays a leading role in it via several institutes (Groningen, Utrecht) but especially via IISG.

CLARIAH will strengthen the position of the Netherlands even more in all these domains, and will enable our country to take on a unique driving role in the transformation process that the arts and humanities research is currently going through.

8 Critical mass

It is a well-known challenge that the vast efforts of the last two decades regarding digitization and tool-development for the humanities have led to a highly scattered environment, seriously diminishing the accessibility and usability of available resources for scholars, students and general users. In fact, what was lacking until recently was the critical mass to combine forces on a truly national scale. But times are rapidly changing, and the awareness that joining forces is not only highly desired but a sheer necessity in this digital era is spectacularly gaining momentum.

With this proposal, an unequalled critical mass has been rallied on an institutional level – see the partners jointly participating in the application, section 4, Partnership case. Through these partners, the whole gamut of available

digital humanities and eScience expertise in the Netherlands is involved, re-using and further developing the important results of previous programmes such as CATCH, and maximizing the power needed to tackle the technical challenges posed by the CLARIAH initiative.

As to the users of the proposed Common Lab, through the affiliation of academic and other institutions, the critical mass of outreach is equally at a full maximum, potentially covering the entire humanities community. It is estimated that in the Netherlands, some 1300 scholars are active in arts and humanities research, and some 30.000 students. Internationally, the number of potential scholarly and other professional users is likely to amount to several tens of thousands.

The cumulative knowledge and skills brought together in the federation of co-applicants is highly competitive on an international scale. For the design, construction and exploitation of digital repositories MPI, DANS, KB, ISSG (with HSN and Clio-Infra), Meertens, Huygens ING, INL, B&G, the university libraries and of course the NA are internationally recognized experts, who have built up and demonstrated their expertise in these areas in a variety of national and international projects. 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. DANS plays a central role in the DARIAH initiative, bringing together institutes that offer their services to the humanities across Europe.

Not only this nationally available expertise and these data-sets will be brought together in the CLARIAH infrastructure, the Common Lab will also connect with and so incorporate the many parallel efforts of other countries through its continuously expanding networks and platforms of international collaboration, first and foremost through CLARIN and DARIAH.

Thus, the digital resources of the partners in CLARIAH could cater to the data needs of thousands of researchers. Social and economic history, together with linguistics and archaeology, are the most intense users of computational approaches within the humanities. Scholars in other fields of the humanities are also starting to use the new opportunities of digital data and tools for processing them for their research, and scholars are increasingly becoming aware of the endless new possibilities to tackle old questions, or to pose new ones. CLARIAH will further strengthen this by making available the Common Lab with all its data and user-friendly tools and by its concomitant education and outreach package.

The Netherlands have an excellent starting position for raising such a new generation of scholars. The faculties of humanities of the participating universities and institutions are breeding grounds of scholars, either individuals or well-organized groups, with strong reputations in their fields. For example, The HSN is in terms of size of the database, the investments made, and the research community surrounding it, one of the biggest in social and economic history in the Netherlands. Over the past fifteen years the HSN has provided data and inspiration for research in a wide variety of disciplines, as reflected by the list of research schools (Posthumus Institute; Interuniversity Centre for Social Science Theory and Methodology, ICS) and other institutions (Netherlands Interdisciplinary Demographic Institute, NIDI/KNAW; Centre for Migrant History, CGM, Data Archiving and Networked Services, DANS; International Institute for Social History, IISH; University of Nijmegen; Academica Sinica (Taiwan) with which the HSN collaborates. In this respect the HSN is a model of multidisciplinary collaboration and outreach to adjoining disciplines.

9 Embedding

On the national level, the CLARIAH is embedded in the full range of co-applicant institutions, together covering the whole of humanities in the Netherlands (see sections 1.1, 4 and 8).

It goes without saying that CLARIAH will be firmly embedded in the European-wide CLARIN infrastructure and network; in fact, the main applicant (UU) and one of the participants (MPI) are the main driving forces behind

CLARIN-EU. There are close relationships, and even overlapping participation, with other international networks such as FLaReNet (just finishing) and comparable initiatives in the US (Interop), Japan, Korea, South America, Australia and South Africa. There is additional participation in the Network of Excellence Multilingual Europe Technology Alliance META-NET and the META-SHARE initiative. Several related research programmes that currently are running or have recently finished strengthen the environment in which CLARIAH will run, and emphasize the need for the CLARIAH infrastructure. Examples are IMIX (recently finished), STEVIN, NWO Dyslexie[26], the ERC Advanced Research Grant awarded to Pieter Muysken (RU), DOBES, the Multilingualism Project (MPI & RU), and the Sign Language Project (RU et al).

Similarly, CLARIAH will be firmly embedded in the DARIAH-EU infrastructure and network. DANS, the 'preparing DARIAH' coordinator and one of the intended three central pillars for the DARIAH-EU construction phase (together with FR and DE), is an institute of KNAW and NWO, receiving stable funding from both umbrella organisations. NWO has tasked DANS to negotiate data contracts with all projects in the humanities and social sciences receiving an investment grant. With DANS being responsible for the long-term access to and preservation of Dutch arts, humanities and social science research data, there is also in this respect a sound embedding of CLARIAH activities in the Dutch research community. DANS is partner in a substantial number of digital humanities projects, both national and international, and member of the foremost European and worldwide data organisations in related fields, such CESSDA - Council of European Social Science Data Archives, IASSIST - International Association for Social Science Information Service and Technology).

The main members of the Netherlands Coalition for Digital Preservation (NCDD), KB, NA, B&G and DANS participate in CLARIAH. We are also directly represented in the board of the Alliance for Permanent Access to the Records of Science (APA) and member of the APARSEN Network of Excellence.

As there is no space to mention the national and international embedding of the many other partner organizations, we restrict ourselves here to the embedding of Clio-Infra: Clio-Infra is hosted at the IISG. Its leader, Prof Jan Luiten van Zanden, is President of the International Economic History Association. Clio-Infra is embedded in a number of other networks and international projects, such as the ESF Research Networking Programme 'Globalizing Europe Economic History Network'; the Economic History Initiative of the CEPR, 'Unifying the European Experience', a Marie Curie Research Training Network; the ESF COST Actions 'Gender and Well-Being: Work, Family and Public Policies' and 'Programme for the Study of European Rural Societies'; The ESF EUROCORES project, 'Inventing Europe: Technology and the making of Europe, 1850 to the Present'; the research initiative of the Centre for the Evolution of the Global Economy devoted to 'Growth and Inequality: Globalization, Shocks, and Response'; the 'Asian Historical Statistics' project, which is collecting, and publishing data on long-term GDP-, labour input- and standard of living-indicators for East Asia, India and Russia; the 'Historical Statistics of Latin America' project.

10 Proven willingness to collaborate

The consortium partners have collaborated in different, often overlapping, constellations in various projects. We provide examples to provide evidence for the willingness to collaborate.

In CLARIN-NL the number of partners grew from 11 (15 units) to 20 (25 units). They all signed the CLARIN-NL consortium agreement.[27] The partners work closely together in a range of subprojects: subprojects to implement the technical infrastructure in the CLARIN-NL project, data curation and demonstrator projects. There is also close collaboration with Flanders to adapt existing tools and data for the Dutch language and integrate them in a seamlessly working environment for workflows of web services.

As described in the previous section (9. Embedding), in Clio-Infra the strong desire of scholars to cooperate in order to create the right research infrastructure for history and the social sciences has been clearly demonstrated. It

[26] http://www.nwo.nl/dyslexie
[27] http://www.clarin.nl/node/72

builds on national research programming within the framework of the Posthumus Institute, internationally it is part of the activities of the IEHA, and participates in many international networks and initiatives aimed at putting together large datasets and intensifying research cooperation.

The IISG is the leading partner in the development of the HSN. The institute has recognized the scientific and strategic interest of the HSN and its potential to become an important resource for social and economic history. The IISG is currently investing in the development of data hubs in order to foster large-scale quantitative and qualitative historical research. Moreover, the HSN has initiated over twenty projects, all geared towards the continued development of the data collection, in cooperation with researchers from Dutch universities and abroad. It has developed close ties with sister institutions in the Netherlands (Posthumus Institute, CBS, NIDI, DANS, etc.) and abroad (DDB Umeå, IPUMS Minneapolis and ICPSR, etc.). In the project *Towards a Global Life Course* the HSN cooperated with ICPSR in Ann Arbor and the Demographic Database of Umea University to create a network for software sharing. This initiative was continued by way of the *European Historical Population Sample network* (EHPS). This ESF-granted network organizes about 30 databases from 20 European countries and another 8 databases from outside Europe and will run over the period 2011-2016.

Five KNAW institutes have successfully worked together in the Alfalab project, inter alia providing a solid basis for the Text Kit of the Common Lab. Within the e-Humanities Group of the KNAW, participants from five KNAW institutes and six universities started working together in 2011 on a five-year programme of research addressing fundamental issues within computational humanities. The e-Humanities Group also works together in an EU FP7 project with university partners from seven countries, about the implications of digitisation for academic review processes. The e-Humanities Group worked on several small projects with Oxford University about the use of information resources by humanities scholars.

The excellent and increasing cooperation in CLARIN and DARIAH at the European level has been described in section 4 (Partnership case). CLARIN is to be continued by a consortium of national projects united at the European level in the so-called CLARIN ERIC[28], hosted by the Netherlands, for which the formal submission is currently being evaluated. DARIAH is likewise well on track to become an ERIC.

CLARIAH now unites all these forces, continuing and intensifying the close collaboration between these research groups involved, both at the national and at the international level, as illustrated inter alia by the NEERI conference and the SDH conferences[29] collaboratively organized by CLARIN and DARIAH.

Summarizing, from the range of national and international projects mentioned above, it is evident that there is a strong and proven willingness to collaborate.

It has become clear, inter alia through the CLARIN and DARIAH projects, that infrastructures for humanities share a lot of properties and encounter similar problems. A common effort is therefore in order. CLARIAH unites, for the first time, all humanities researchers, in one infrastructure project, and this is the added value of the cooperation envisaged in the CLARIAH project over the cooperation in the CLARIN and DARIAH projects. By doing this, CLARIAH maximizes the potential for synergy, creates mass and avoids fragmentation and duplication of effort.

11 Reflection of social trends

CLARIAH contributes to issues of the modern information society in multiple ways. First, it directly contributes to digital longevity in the information society. It ensures that important data and tools are preserved. CLARIAH also offers opportunities to extract information from these data that may be crucial for analysing social issues, thus creating an evidence base for government policy actions. The CLARIAH infrastructure encompasses both

[28] ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric

[29] http://www.csc.fi/english/pages/neeri09; http://ztwweb.trans.univie.ac.at/sdh2010; for 2011 a new SDH is planned: http://cst.ku.dk/sdh2011/

unstructured and structured data, and tools for analysing these data. Societal developments can be detected by analysing both types of data.

11.1 Digital longevity in the information society

Digital longevity and access to digital information (government information, electronic publications, research data) has become an important social and political issue. Digital data has become an indispensable part of our collective memory. This is not only of scientific and cultural importance, there is also a political and economic need to keep our digital memory accessible. The Dutch cabinet underlines the importance of access to science data (see section 3.1). CLARIAH will combine the best efforts on digital preservation and access in Europe, and its partners are already contributing to both the theory and the practice of digital longevity.

11.2 Societal developments detected by analysing unstructured data

In the current information society, information is ubiquitous, distributed fast via telephone and Internet connections, and it comes in huge quantities. CLARIAH brings together a range of textual and audio-visual tools and data and harmonizes them so that they can seamlessly operate together. This makes CLARIAH an excellent infrastructure for addressing the challenges posed by this overflow of information, e.g. by filtering irrelevant information and summarizing crucial information, thus turning these challenges into opportunities.

One example is the issue of *Social Climate Change.* 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. CLARIAH 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.

CLARIAH will also contribute to important societal issues 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. Just like information that could have prevented the murder of Theo van Gogh was available, but not used, because there was no adequate technology for analysing and prioritizing documents. The CLARIAH 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 challenges, 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. CLARIAH will therefore, in various ways, also contribute to these innovation domains.

Finally, since the set-up of an infrastructure necessarily requires the development of interoperable data and services, based on a new generation of web-based services and applications, and since CLARIAH has set aside a significant sum for dissemination and training, it directly contributes to the *Digital Agenda*[30] and its pillars *Interoperability* and Standards, Enhancing *eSkills*, Research and Development, and ICT for Social Challenges.

11.3 Societal developments detected by analysing Structured Data

The social and economic challenges that the Dutch society is currently facing form part of international trends, for which international groups of scientists have to find the answers by merging their data and their expertise. An example is the linked problems of ageing and below-replacement fertility that is occurring in several European countries. Another example is related to the determinants of (productivity) growth in the recent past (and in the long run. By analysing long-term trends on a global scale on the basis of the global datasets, new insights into for example the causes and consequences of global environmental change, or the effects of globalization on global

[30] http://ec.europa.eu/information_society/digital-agenda/index_en.htm

inequality, can be realized. A thorough analysis of this huge challenge for policy makers requires detailed data on long-term historical trends, including micro-level data on the fit between fertility decisions and labour market decisions in various countries, and insight into the widening of life course options of individuals. CLARIAH will bring together disparate data on individual life courses with aggregate economic and social data on the meso and macro levels, on a European scale.

Changes in patterns of life courses and family structures have an impact on almost all domains of government policy: the labour market, social security, health and health care, housing and mobility, aging, social exclusion, income, families, and so forth. The data resources combined in CLARIAH will make it possible, for instance, to study the impact of the transformation of the Western family on children, negative effects of parental divorce during childhood, of growing up in an incomplete family, of family fragmentation and the role of quasi-kin have become prominent issues in family and health policies. The evidence base of the recent and more distant past can contradict popular misconceptions about family life, which influence social movements and government policies and programs. The results of analyses using the population databases can be used to better inform the public, media, and policy makers.

12 Timetable

Duration of the project

Planned starting date	1 September 2012
Expected completion date	1 September 2017

The table below gives a first general indication of the foreseen tasks, milestones and deliverables of CLARIAH. These will be further specified during the start up phase of the project.

Milestones and deliverables

Del. no.	Milestone or Deliverable name	Start date (Mo/Yr)	End date (Mo/Yr)
1	**Management & Coordination**		
1.1	**General Management & Coordination**		
1.1.1	Consortium Agreement	1/1	6/1
1.1.2	Installation of Governance & Management structure	1/1	6/1
1.1.3	Extension of CLARIN Office for CLARIAH	3/1	9/1
1.1.4	Coordination with CLARIN-EU and DARIAH-EU	1/1	12/5
1.1.5	Coordination of CLARIAH-related (EU) proposals for tender	Dependent on calls	
1.1.6	Reporting	Periodically	
1.2	**Technical management workspaces**		
1.2.1	Requirements, detailed work plan, incl. contingency planning, use cases	1/1	6/1
1.2.2	Development of workspace system, integration	9/1	12/5
1.2.3	Scalability and access	1/4	12/5
2	**Conversion of analogue sources**		
2.1	Specification of sources	1/1	6/1
2.2	Overview of conversion techniques	1/1	6/1
2.3	Selection of conversion technologies	6/1	9/1
2.4	Testing & fine-tuning	9/1	½
2.3	Conversion	1/2	12/5
3	**Data enrichment & annotation**		
3.1	Work plan	1/1	6/1
3.2	Use cases & specification requirements	1/1	12/1
3.3	Collection & testing of existing models and methods	1/1	12/2
3.4	Testing & implementation	1/2	12/5
3.5	Deployment	1/3	12/5
3.6	Exploitation	1/3	12/5
4	**Linking resources, interoperability & standardization**		
4.1	Detailed work plan; interlinking and publishing resources, interoperability standards	1/1	6/1
4.2	Inventory eHumanities standards and flexible standards registry, continuous governance	1/1	12/1,12/5
4.3	Status report on interoperability aspects for CLARIAH tools and services	6/1	12/5 biannual

4.4	General publication environment for data resources and virtual collections, version	6/2	6/3, 12/5 biannual
4.5	Tools and APIs for resource-part marking and identifier creation, version 1	1/2	12/2, 12/5 biannual
4.6	Facilities for semantic information extraction, export and publishing, version 1	1/3	12/3,12/5 biannual
4.7	Standardization workshops and education	1/2	12/5 2 WS/year
5	**Pattern recognition & content extraction**		
5.1	Detailed work plan, selection of pattern recognition en content extraction tasks	1/1	12/1
5.2	Collection and testing of existing methods	1/1	12/3
5.3	Implementation and testing of improved methods	1/2	12/5
5.4	Deployment of methods to humanities researchers	1/2	12/5
6	**Visualisation & presentation of information content**		
6.1	Detailed work plan, Specification of requirements	1/1	6/1
6.2	Use cases	1/1	6/1
6.3	Overview and selection of visualisation techniques	1/1	9/1
6.4	Overview and selection of visualisation technologies	1/1	9/1
6.5	Implementation	9/1	12/5
6.6	Testing	9/1	12/5
6.7	Exploitation	1/3	12/5
7	**Long-term preservation & access**		
7.1	Detailed preservation and access plan	1/1	3/1
7.2	Synchronize archiving system with CLARIN data centres, DARIAH-EU VCC 3, national eInfrastructure	3/1	9/1
7.3	Development & implementation of synchronized archiving system	9/1	3/3
7.4	Archiving and access provision to data	1/1	12/5
8	**Dissemination & Outreach**		
8.1	Detailed Dissemination & Outreach plan	1/1	6/1
8.2	CLARIAH conferences and workshops	(1/Lab + 1 joint)/Year	
8.3	CLARIAH Website (portal) design	6/1	12/1
8.4	CLARIAH Website (portal) exploitation	1/2	12/5
8.5	Development of training courses	6/1	12/1
8.6	Newsletters & publications	Periodically	

1 General information
2 Research proposal
3 Timetable
4 **Declaration/signature**

13 Declaration and signature

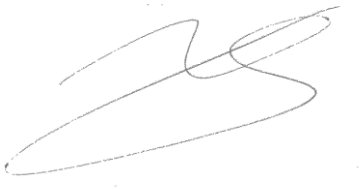
Have you requested funding for this research elsewhere?

☒ No

☐ Yes,

Declaration

☒ By submitting this form through Iris, I declare that I have completed this form truthfully and completely.



Prof. dr. J.E.J.M. Odijk

Universiteit Utrecht

SHORT PROGRESS REPORT CLARIN-NL

Project Start:	1 April 2009
Project End:	1 April 2015
Reporting Period:	1 April 2009 – 1 August 2011.

The progress report given here follows the 11 Criteria of the Roadmap Call for Proposals. More detailed yearly reports for 2009 and 2010 are available on the CLARIN-NL website.[1]

1 Science Case

The CLARIN-NL project aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent eScience working environment for researchers in the Humanities, and Linguistics in particular, who want to make use of language resources and technology for their research.

CLARIN-NL is building a technical infrastructure that is distributed in nature and accessible from anywhere via the internet. The core of the distributed network is formed by the so-called CLARIN-Centres. The four initial candidate CLARIN Centres in the Netherlands work together in a number of projects to make choices for software components and/or implement the technical infrastructure, and the fifth will join this Infrastructure Implementation Project (IIP) later this year. The project Search & Develop (S&D) implements sophisticated search facilities in metadata and data to complement the browsing functionality for which a prototype (the Virtual Language Observatory, VLO[2]) was developed in CLARIN-prep. The search facilities will have a centralized architecture for metadata (after they are harvested) and a federated one for the actual content data.

CLARIN-NL aims to incorporate existing data providers such as libraries (KB and university libraries), Beeld & Geluid, DBNL, Nationaal Archief, etc. CLARIN-NL has reserved budget and is currently developing plans with the KB, Beeld & Geluid, and DBNL to integrate these organizations in the CLARIN infrastructure. Metadata form a crucial ingredient for all data and tools in the CLARIN infrastructure. Without explicit and formalized metadata, the infrastructure would not be able to function. Special facilities for working with metadata have been created (CMDI[3]), the feasibility of these facilities has been tested and all subprojects are obliged to make CLARIN-compliant metadata for the resources (data and tools) that they work on. The approach adopted in CLARIN to achieve semantic interoperability is widely being tested and has already been refined on the basis of these tests. Achieving full semantic interoperability is not easy, but especially in the domain of metadata significant progress has been made.

A whole range (18) of (relatively small) subprojects has been started up (and partially finished (10)) in 2010 and 2011, and more are being prepared to start in 2012. These projects involve resource curation (8 projects), demonstrator projects (5), and mixed curation and demonstrator projects (5).[4] The selection of some of these subprojects was guided by the results of the user survey held in 2010.[5] In resource curation projects existing data and tools are adapted to make them CLARIN-compatible, i.e. meeting the requirements to function in the CLARIN infrastructure.[6] Demonstrator projects focus on existing software and adapt it to make it CLARIN-compatible, but they also provide a version that runs on a server of a CLARIN centre together with a demonstration scenario. These demonstrators and demonstration scenarios can be used to illustrate the potential and the benefits of the CLARIN

[1] http://www.clarin.nl/node/47

[2] www.clarin.eu/vlo/

[3] http://www/clarin.eu/cmdi

[4] For an overview see http://www.clarin.nl/node/76

[5] http://www.clarin.nl/node/64

[6] See http://trac.clarin.nl/trac/wiki/WikiStart#CLARIN-compatible for a description of the notion CLARIN-compatible.

infrastructure to other humanities researchers. Some examples of demonstrators and demonstration scenarios can be found on the CLARIN-NL website.[7] Since curation of data is crucial for a proper functioning of the infrastructure, a Data Curation Service has been set up and starts its operations on Sep 1, 2011. Even though the technical implementation of the CLARIN infrastructure is just in an initial stage, the existence of some parts of the CLARIN infrastructure have already led to new scientific insights that could not have been achieved without them.[8]

2 Talent Case

The Netherlands is clearly in the lead for the CLARIN infrastructure, both organizationally and technically. The CLARIN-NL project has strengthened this position. CLARIN-NL has organized many tutorials and workshops that attracted international participants even though they were mainly intended for researchers working in the Netherlands. The technical implementation of the CLARIN infrastructure is still in too initial a stage to have a large impact yet, nevertheless it is clear to the researchers in the Netherlands that an attractive research infrastructure is being created, parts of which they can already use now and contribute to.

3 Innovation Case

Many of the tools that are being incorporated in the CLARIN infrastructure are not only useful for humanities researchers but can be used in commercial applications as well. For example, in the CLARIN-NL WIP subproject tools are adapted and a demonstrator is being created to investigate the attitude of parliament members towards war using the parliamentary minutes. These tools can also be used in commercial applications to investigate the perception of a company, product or political party by the public in written media, and a variety of companies have expressed their interest in and support for CLARIN-NL for this reason. The tools and a large part of the data used in CLARIN are also important for the increased attention to natural language processing at the European level (witness the FLaReNet, EuroMatrix, META-NET and related projects) and CLARIN has been a major source of inspiration for the set-up of an exchange facility for language and speech technology developers (META-SHARE) and is closely collaborating with it. There are natural links with the "topgebied" Creative Industry, and initial discussions on collaboration have been initiated and are to be continued in August 2011.

4 Partnership Case

The CLARIN-NL Office has been set up and consists of the director and a project secretary. A governance structure suited to the project type has been implemented. There is an Executive Board (4 members), a Board (7 members), a national advisory panel (NAP, 17 members) and an international advisory panel (IAP, 7 members, one vacancy). The exact composition of these groups can be found on the CLARIN-NL website.[9]

The CLARIN-NL consortium does not only include universities but also libraries, data centres, KNAW institutes and other centres.

5 Business case

Total Costs The CLARIN-NL project runs from April 1 2009 to April 1 2015. For convenience we have represented the budget as if the project runs from Jan 1 2009 to Jan 1 2015. For each year we specify (1) The original budget; (2) The actualized budget; (3) Committed amount; (4) Actually spent amount. All amounts are in kEuro. For 2011, it is pretended that nothing has been spent yet. We did not budget beyond 2014, but the budget amounts of 2013 and 2014 are the expected yearly exploitation costs.

---

[7] http://www.clarin.nl/node/185

[8] E.g. Bennis 2010. See http://www.clarin.nl/node/167

[9] http://www.clarin.nl/node/6

CLARIN-NL costs (Euro*1000)	***YEAR 1 2009***	***YEAR 2 2010***	***YEAR 3 2011***	***YEAR 4 2012***	***YEAR 5 2013***	***YEAR 6 2014***	***TOTAL yr 1-6***
Original Budget	€ 1,346	€ 1,346	€ 1,904	€ 1,904	€ 1,254	€ 1,254	€ 9,010
Actualized Budget	€ 186	€ 1,180	€ 3,085	€ 2,050	€ 1,254	€ 1,254	€ 9,010
Committed	€ 982	€ 1,398	€ 3,085	€ 663	€ 344	€ 0	€ 4,091
Spent	€ 186	€ 1,180	€ 0	€ 0	€ 0	€ 0	€ 1,367

Non-NWO Contributions In a cooperation project with Flanders on tools for the (shared) Dutch language, the Dutch researchers are funded by CLARIN-NL and the Flemish researchers are funded by the Flemish government (EWI, 792k euro). In addition, CLARIN-NL supported a project independently financed by NWO (CKCC, or 'Geleerdenbrievenproject'[10]) to enable this project to use language technology in a CLARIN-compatible manner.

Requested NWO Financing None

6 Technical Case

The construction of the infrastructure is well underway. For many aspects we do not currently see or expect major technical problems. However, there are a number of areas that require special attention: (1) **semantic interoperability** Though CLARIN-prep adopted a very clearly defined approached and partially implemented this with the Data Category Registry ISOCAT[11], the actual large scale testing of this implementation led to a number of revisions, including the set-up of an additional registry RELCAT, an interface to other registries (CLAVAS) and the appointment of a person dedicated to coordinating semantic interoperability issues;(2) **Web services authentication** Currently no generally accepted way exists through which web services can authenticate and delegate the identity of a user. At this moment we have set up a pilot project with BiG-Grid that will test out a possible solution based on OAuth2; (3) **SURFNet opt-in policy** The use of Federated Identity management within the CLARIN project is hampered by the current opt-in policy from SURFnet, the Dutch NREN, which uses a set-up that is peculiar to the Netherlands. We are currently trying to approach the user organizations en block to give permissions for access to all CLARIN services. Already SURFnet has agreed to treat all CLARIN services (including the non-Dutch ones) as a single entity. But it remains problematic.

7 Possible Focus for the Netherlands

The Netherlands is clearly in the lead for the CLARIN infrastructure, both organizationally and technically, and this is recognized by all other European CLARIN partners. The CLARIN-NL project has strengthened this position. The Netherlands has (with significant legal assistance from OC&W) made a formal submission for the CLARIN ERIC. The Netherlands will also host the CLARIN ERIC.

8 Critical Mass

There are 5 candidate CLARIN Centres: MPI and INL are internationally recognized experts for the design, construction and exploitation of archives, DANS has expertise in long term preservation and also has a central role in the DARIAH infrastructure. Meertens and Huygens are institutes that already maintain large datasets and are building up expertise to be able to function as a fully qualified CLARIN-centre. Important data providers such as the KB, Beeld & Geluid, and DBNL fully participate in CLARIN-NL, and initial contacts to include university libraries have been made.

[10] http://www.clarin.nl/node/76#CKCC

[11] http://www.isocat.org

In the areas of language and speech technology, all players in the Netherlands participate in CLARIN-NL. CLARIN-NL incorporates the results from the Spoken Dutch Corpus project and the IMIX and STEVIN programmes (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. The Humanities researchers are well-covered by the participation of all humanities departments of the universities, many KNAW humanities institutes, and several other institutes (KDC, Aletta, IGTD, Veteraneninstituut). The facility being developed will be accessible to all researchers of the participating countries. Provisions for special arrangements for researchers from non-participating countries are in place.

9 Embedding

The CLARIN-NL project is part of a Europe-wide enterprise. This enterprise was initiated by the just finished CLARIN preparatory project (CLARIN-prep[12]) and is to be continued by a consortium of national projects united at the European level in the so-called CLARIN ERIC.[13] A formal request to establish the CLARIN ERIC is underway and it is expected to start early 2012. The Netherlands played an important role in CLARIN-prep, and the CLARIN ERIC will be hosted by the Netherlands.

There has been close collaboration, both at the national and at the international level with the DARIAH project, witness the NEERI conference and the collaboratively organized SDH conferences.[14] The exchange facility META-SHARE is currently being set-up in META-NET with a related but different focus and targeted at different users. There is close collaboration between CLARIN and META-SHARE: for example, META-SHARE has adopted the CLARIN metadata framework.

CLARIN-NL has organized meetings with NWO and KNAW to make sure that these organizations also require that language-related data and tools are CLARIN-compatible.[15] NWO requires this in an indirect manner via DANS, and the KNAW will include CLARIN-compatibility for the relevant resources in their policy note on resources. There is no interdependence with current or future participation of the Netherlands in other facilities.

10 Proven Willingness to Collaborate

In CLARIN-NL the number of partners grew from 11 (15 units) to 20 (25 units). They all signed the CLARIN-NL consortium agreement.[16] The partners work closely together in a range of subprojects: subprojects to implement the technical infrastructure in the CLARIN-NL project, data curation and demonstrator projects. There is also close collaboration with Flanders to adapt existing tools and data for the Dutch language. At the European level, an agreement has been reached between the European partners on the organisation and structure of the CLARIN ERIC, which resulted in a formal submission that is currently being evaluated.

The CLARIN-NL project strengthens already existing cooperation between partners, but it specifically brings together humanities researchers with technology researchers and/or providers in joint subprojects.

11 Reflection of Social Trends

Nothing specific to report over this period.

---

[12] www.clarin.eu

[13] ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric

[14] http://www.csc.fi/english/pages/neeri09; http://ztwweb.trans.univie.ac.at/sdh2010 for 2011 a new SDH is planned: http://cst.ku.dk/sdh2011/

[15] http://trac.clarin.nl/trac/wiki/WikiStart#CLARIN-compatible

[16] http://www.clarin.nl/node/72