Application form

Investment Subsidy NWO Large 2011-2012

The completed application form including attachments should be submitted electronically via the Iris system (personal account of the applicant). Please ensure that the form and attachments are saved in PDF format. Iris is accessible via the NWO website: www.iris.nwo.nl. Please use the Verdana font, size 8.5, line spacing 13 and keep to a **maximum of 30 pages A4 (incl. attachments)**.

Please read the brochure Investment Subsidy NWO Large 2011-2012 before completing this form.
The closing date for the submission of applications is 1 November 2011, 17.00 hrs. The date and time on which you upload via Iris is the valid submission date and time.

General information

Applying university/institute

University/institute	KNAW
Faculty	Meertens Instituut

Project leader/co-ordinator

Title(s)	prof.dr.	
First name	Hans	
Initials	H.J.	
Surname	Bennis	☒male ☐female
Address for correspondence	Joan Muyskenweg 25, 1096 CJ Amsterdam	
Telephone number	020-4628523	
Fax	020-4628555	
Email	Hans.Bennis@Meertens.knaw.nl	
Website (optional)	http://www.meertens.knaw.nl/cms/nl/medewerkers/142447-hansb	
Preference for correspondence		☐English ☒Dutch

Title of investment project

Nederlab - Laboratory for research on the patterns of change in the Dutch language and culture

Abstract/executive summary

Project summary

The aim of this project is to bring together all digitized texts relevant to our national heritage, the history of Dutch language and culture (c. 800 - present) in one user-friendly and tool-enriched open access web interface, allowing scholars to simultaneously search and analyse data from texts spanning the full recorded history of the Netherlands, its language and culture.

Language and culture are inherently dynamic phenomena. Scholars in the humanities – linguists, literary scholars, historians – try to understand processes of change and variation and to uncover their internal and external causes. Large quantities of data are needed to answer the research questions relevant to these fields of study. Until now, this research has necessarily consisted of case studies of limited scope, covering relatively short time periods and single disciplines. As more Dutch historical texts become available in digitized form, new research questions emerge, and systematic research into the interaction of changes in culture, society, literature and language across the full history is becoming feasible. The hypothesis is that changes in language and culture – both of them expressions of human cognition – are related, and that investigating these relationships will further our understanding of them. This research will yield novel and cross-disciplinary insights in fundamental theoretical problems such as the nature-nurture debate, the discussion on cultural identity and cultural integration, the emergence of canons, and the diffusion of knowledge, culture and language.

As yet, however, no comprehensive research corpus has been established, mainly for practical reasons. The available diachronic corpora are housed in different places, with different metadata, and cannot be searched simultaneously. Advanced software tools necessary for the analysis of the texts are available, but are in most cases specific to a particular goal rather than generally applicable. The potential of the available technological means and scientific standards of verifiability and replicability is far from being exploited.

A group of experts in diachronic research, corpus and infrastructure construction, search technologies and tool development jointly submit this application for the foundation of Nederlab. Nederlab offers a user-friendly web interface enabling a distributed search of available diachronic corpora, at text and metadata level, which will lead to better access to existing corpora, as well as quality improvement and standardization of data and metadata.

The project builds on various initiatives: for corpora Nederlab collaborates with the scientific libraries and institutions, for infrastructure with CLARIN, DARIAH (and potentially CLARIAH), for tools with eHumanities programmes such as Catch and IMPACT. Nederlab has the added value of creating a user-friendly infrastructure for researchers, which will promote cooperation and synergy as well as the formulaton of new, often interdisciplinary, research questions. During the project a great deal of attention will be paid to the dissemination of information; the infrastructure will be set up in close consultation with the research community, and will be tested in concrete research pilots.

Summary of the investment proposal for the general public

Nederlab's goal is to enable scholars in the humanities to find answers to new, longitudinal research questions. For this purpose Nederlab aims at setting up a user-friendly tool-enriched web interface, allowing researchers to simultaneously search and analyse the digital historical texts made available by scientific libraries and institutions, at text and metadata level.

Key words

eHumanities, diachronic, corpora, tools, workspace

Top 10 relevant publications

Project summary

- Barbiers, S., H.J. Bennis *et al*. (2005-2008). *Syntactic Atlas of the Dutch Dialects, Volume I, II*. Amsterdam: Amsterdam University Press.
- Coupé, Griet & Ans van Kemenade (2009). 'Grammaticalization of modals in English and Dutch: uncontingent change'. In: P. Crisma and G. Longobardi (eds.) *Historical Syntax and Linguistic Theory*. Oxford: Oxford University Press, 250-270.
- Daelemans, W. & A. van den Bosch (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.
- Hunt, L., M.C. Jacob & W. Mijnhardt (2010). *The Book That Changed Europe*. Cambridge Mass., Harvard University Press/Belknap Press.
- Kennedy, James (2008). 'Religion, Nation and European Representations of the Past'. In: Stefan Berger and Chris Lorenz (eds.), *The Contested Nation: Ethnicity, Class, Religion and Gender in National Histories*. Basingstoke: Palgrave Macmillan, 104-134.
- Leerssen, J.T. (2010). 'Viral Nationalism: Romantic intellectuals on the move in 19th-century Europe'. In: *Nations and Nationalism, 17*(2), 257-271.
- Nerbonne, John (2010). 'Measuring the Diffusion of Linguistic Change'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 3821-3828. DOI: 10.1098/rstb.2010.0048.
- Sijs, Nicoline van der (2001). *Etymologie in het digitale tijdperk. Een chronologisch woordenboek als praktijkvoorbeeld*. Doctoral dissertation Leiden.
- Steen, Gerard J., Aletta G. Dorst & J. Berenike Herrmann (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam/Philadelphia: John Benjamins.
- Stipriaan, R. van (2007). 'Words at War: the Early Years of William of Orange's Propaganda'. In: *Journal of Early Modern History*, 11, 331-349.

Investment proposal

Research field and research plan(s)

1. The goals[1]

Many Dutch and Flemish texts from the eighth century to the present day have been preserved, and many of these – both fiction and non-fiction – have by now been digitized. These texts reflect the dynamic development of Dutch language and culture, with alternating periods of acceleration and deceleration. Scholars from various disciplines – history, literature, culture, linguistics – seek to describe and analyse the historical changes in their fields, the rate at which they took place, and the forces driving them. Much, however, is not yet known, and little or no research has been done on the interaction between changes in culture, society, literature and language. Yet, there is evidence that changes in the various domains do interact: many linguistic changes are the result of language contact caused by migration, and so by changes in society; historians have shown that language can be a guiding and formative factor in political and social changes; and contemporary culture is reflected in literary works.

Nederlab will bring together the digitized historical resources made available by scientific libraries and institutions in a user-friendly and tool-enriched web interface, allowing researchers for the first time to systematically investigate changes in Dutch language and culture and their interrelations on the basis of a corpus of Dutch texts from the earliest times until the present. The hypothesis is that changes in language and culture – both of them expressions of human cognition – are interrelated, following identical or similar patterns. Using Nederlab will enable us to bring these regularities to light. The investigation will reveal which parts of the Dutch language and culture are subject to change and which remain constant. Changes will have to be explained partly as autonomous developments, and partly as the result of the influence of various (internal or external) processes. External causes may involve cultural transfer or cultural integration. Constants may be part of the inherent properties of language and culture, and increased insight into these will supply information that can be used in the nature-nurture controversy. Constants can also, however, indicate the existence of a tradition, leading to the emergence of canons, or a national cultural identity.

Research into patterns of change is not new; experienced researchers like R. Aerts, R. Bod, A. van Kemenade, J. Kennedy, J. Leerssen, W. Mijnhardt, T. Vaessens, to name but a few, have demonstrated their expertise in innovative publications in various fields. These experts play a key role within Nederlab as supervisors or as members of one of the four advisory boards (see the organization structure in 'Relation to other research groups'). They will do pioneering work in Nederlab, supported by acknowledged technical researchers such as A. van den Bosch, F. de Jong, J. Nerbonne, N. Oostdijk, L. Schomaker, by specialists in search technologies such as J. Kamps and A.P. de Vries, and by experienced corpus builders such as J. Hoeksema, G.J. Postma, P. van Reenen, M. Rem, N. van der Sijs, R. van Stipriaan and INL staff.

Owing to the lack of an extensive diachronic corpus, all existing research so far has been limited in scope, covering relatively short time periods. Nederlab will provide the means of discovering and integrating regularities of change in Dutch language and culture over long time periods: a giant leap forward for research. Nederlab will constitute a longitudinal and comprehensive research object, offering user-friendly tools to search it. It will comprise a diachronic corpus of texts from the eighth century to the present day. The texts have been digitized by various institutions, but in Nederlab they will be made available for the first time in one corpus accessible for distributed search. An important aspect of Dutch language and culture is pluriformity: both in language and in culture there has always been strong diversity – and in this diversity we will have to try and find the explanation for at least some of the changes. In order to be able to integrate all of this variety into the investigation, information on four variables, *viz*. time, place, author and text type, will be added to the corpus. The variables 'time' and 'place' add diachronic and geographical dimensions. With these four variables, both the input and the output of research questions can be

[1] The Meertens Institute, INL and the Universities of Nijmegen and Amsterdam have all contributed towards defraying the cost of developing and writing this application.

sorted out in various ways, allowing researchers to follow the spread of a particular change in the course of time in various dialects, or in certain categories of literary texts.

Since Nederlab's diachronic corpus spans more than ten centuries, new longitudinal questions can now be formulated. Examples include popular opinions about foreigners through the ages; the origin, spread and disappearance of literary genres; changing critical judgements on literary writers, scholars or historical events; or the development of the use of prepositions or conjunctions (see also the use cases).

In order to anwer such longitudinal questions, Nederlab provides a user-friendly web interface with up-to-date computer technology and instructions for non-technical researchers on how to use it. With these tools, researchers can examine the corpus or specific subsets of it. Nederlab's computer technology can also be used for automatic analysis of large quantities of text files for data mining purposes – this method will doubtless produce unexpected finds and lead to the discovery of patterns in the data that could not have emerged from manual searches, needing further clarification by researchers, or shedding new light on existing research results. We also stand to learn more on the Constant Rate Hypothesis of language changes, which states that in related linguistic contexts language changes spread at the same rate, and on the relation between linguistic changes and changes.

The discovery of regularities must of necessity be effected by translating large fundamental research questions into smaller, concrete research questions, the answers to which will contribute to an understanding of the larger ones. It is particularly for this purpose that Nederlab is extremely useful: on the one hand because all the necessary tools are brought together here, and on the other hand because, with Nederlab, researchers can successively, but also simultaneously, work on subproblems that may contribute to solving larger problems and thus further larger-scale research efforts. The data enriched by an investigator as well as the research method, will be stored in Nederlab in the form of specially adapted tools, enabling later researchers to use them. Scholars in the humanities can thus build on the results achieved by their peers: for instance, automatic name recognition developed by linguists can be developed further by historians by linking concrete biographical data to the names, while literary scholars can link authors' names to bibliographical data. Nederlab thus automatically contributes to an accelerated integration of divergent types of research in the humanities.

The instruments offered by Nederlab facilitate verifiable, reliable, objectivized, replicable, and statistically well-founded research within the humanities. For many disciplines in the humanities this is a methodological innovation that can provide a strong empirical basis for these studies: case-bound, subjective and qualitative research can now be tested quantitatively. The application of these methods can be fully expected to lead to many new insights in the dynamics of change in the Dutch language and culture, and to all kinds of new longitudinal research questions, which in turn will lead to new research methods. Nederlab's approach to research will be illustrated below by concrete use cases.

2. Use cases

Nederlab offers the instruments with which to find empirically supported answers to questions that humanities researchers working on the history of Dutch language and culture put to a diachronic corpus, allowing them to test existing hypotheses against an extensive corpus, and to come up with new hypotheses via exploratory data analysis. The instruments provided can answer an - in principle unlimited - series of specific research questions; in the course of preparing this grant application, we have consulted – orally or in writing - c. 150 researchers engaged in diachronic research and/or eHumanities (see Appendix 1) with respect to the research questions they would like to see answered with the aid of Nederlab. Based on the nature of the questions and the methodology suggested, they can be broadly subdivided into questions dealing with the detection of innovations, i.e. the onset of a change (2.1), establishing the spread of changes (2.2), finding (statistical) connections and networks (2.3), and detecting similarities and differences between texts (2.4). The answers to these questions will lead to large, coherent research themes, also referred to as 'scholarly narratives'.

In order to do justice to the variation in the corpus, the following four metadata items will be added to each individual machine-readable text:

- **time**: the year the text was written or printed, and a link to time line/periodization;
- **place**: provenance of a text, and a link to geographical index and GPS;

- **author data**: name, and a link to biographical data;
- **text type**: fiction, non-fiction, newspapers, ego documents, …; and a link to biographical data and to scans of the printed or written original.

In addition, metadata will be added on the quality of the texts, enabling researchers to weigh the reliability of the results found:

- **quality**: a. corrected text (= diplomatic transcription or corrected OCR), b. critical transcription or respelled text, c. scanned with OCR (i.e. with recognition errors).

Finally, part of the corpus will be enriched at word level, making structured searching possible. This, too, will be indicated by metadata, so that various options can be tried out simultaneously:

- **enriched in content**: annotated by reseacher, (semi-)automatically lemmatized/POS, morphologically tagged, syntactically parsed, time indications tagged, placenames tagged and georeference added, personal names tagged, information on type of language (dialect, sociolect, American Dutch, Belgian Dutch etc.) provided. The tagging of time indications, place names, personal names and information about linguistic usage agrees with the four corpus metadata, i.e. time, place, author and text type. The tagging is automatically stored via, e.g., ***FoLiA*** (Format for Linguistic Annotation), with the original reproduction of the text left unchanged. However, each individual researcher can add his own layer of enrichment or annotation.

2.1 Research questions relating to tracing the beginning of innovation in language and culture

For much of the research on the history of Dutch language and culture, it is important to record the development of change processes. The first step is to establish when and where a certain innovation appeared for the first time (2.1); after that, the spread of the innovation in time, text and space is investigated (2.2).

2.1.1 Detecting new concepts

Historians want to know how quickly technical innovations (*windmills, electricity, gaslight, telephone, photography*) reached the Low Countries from abroad. From the data it may become clear that concepts referring to a specific type of foreign innovation (e.g. concrete inventions) were introduced faster or, for that matter, more slowly than other types (such as words denoting a specific body of thought) while it may also become clear that in a certain period more foreign innovations were adopted than in other periods, or that they showed up more quickly. This in turn would tell us something about the openness or closedness of Dutch society and culture through the ages. Linguists are furthermore interested in the names these new concepts were given, while literary scholars are occupied with aspects of image formation of new concepts.

Nederlab provides previously unavailable material and tools for tracking down new concepts. Without Nederlab, a researcher has to consult several different corpora on different websites using different interfaces. Nederlab, by contrast, offers a complete diachronic corpus available for distributed search. In addition, Nederlab provides ***automatic indexation software*** for all word forms from the complete corpus, which may be consulted in various ways: as a word list, or in context (KWIC, Key Word In Context) via ***concordance building software***. Since all texts are supplied with metadata, all the forms in the corpus can be sorted chronologically, thus establishing the earliest site of each word form in a simple way. With the help of the metadata on place, author and text type, it can then be established where, by what author and in what text typr a form was first used.

This automatic service covers all forms throughout the corpus. On the basis of the quality metadata, researchers can compare forms found in corrected texts with those from texts scanned with OCR. Where the data deviate conspicuously (e.g. the word *televisie* 'television' found in an OCR-ed text from 1886), this is a reason for consulting the original source text (which shows *televisie* in 1886 to be an OCR error for *ter visie*).

The automatic indexation of forms does not create a link between the spelling and form variants of one and the same word. For many research questions, however, one would like to abstract from such variation. This, too, is provided for by tools in Nederlab:

1. ***Tokenization tools***: the input text (a sequence of characters) is converted into a sequence of tokens (= occurrences of word forms). Tokenization operates on the basis of spaces, punctuation marks, etc.; some components of the corpus have been tokenized already.

2. ***Tools for spelling normalization***: these map spelling variants, spelling mistakes, OCR mistakes etc. onto a canonical spelling form such as a headword or lemma; examples of such tools are TICCLOPS (developed at Tilburg University), tools that have been developed within the European project IMPACT, and tools that map variants of personal names (developed within the Catch-LINKS project).
3. ***PoS-tagging tools***: attribution of a Part of Speech-code to an occurrence of a word form (= token) in context; part of the corpus has already been provided with PoS information.
4. ***Lemmatization tools***: attribution of a lemma (= canonical form of an inflectional ending) to an occurrence of a word form; part of the corpus has already been lemmatized.
5. In all these processes a ***computational historical lexicon***, as developed by INL for texts from the 16th-20th centuries, is an important tool. This lexicon automatically attributes word variants to lemmas.
6. Subsequently ***indices*** can be put together on the basis of:
- the set of word-form types
- the set of lemma types
- the set of types of the word form in canonical spelling.

Using these indices will improve the precision and recall of the searches performed.

2.1.2 Detecting new word forms

Linguists want to detect not only new words and concepts, but also new word forms; they ask questions like: since when and in what dialect has the past participle with *ge-* been recorded and what type of verbs were the first to take participles with *ge-*?

Nederlab will be an indispensable new tool for historical morphology: at the moment, there are only a few historical studies of Dutch word formation, and this is mainly due to lack of material – data and tools. Using Nederlab will allow the first systematic enquiry into the way in which compounds and derivations were formed in Dutch in the course of time. A case in point may be the study of derivations, i.e. words containing affixes, an element that is not used independently, like *–te* in *diepte* 'depth' or *–ig* in *groenig* 'greenish'. The number of words that can be derived with a particular affix is in principle unlimited (although confined by specific restrictions), but the number of affixes is finite and can be established for each of the various periods of Dutch.

Researchers try to find answers to questions such as: Since when and where do the various types of derivations occur in Dutch? Why did certain types of derivations (e.g. *diepe*) disappear, to be replaced by others (*diepte*)? What were the combinatory restrictions of the various affixes in the course of time? How can we account for the rise of new types of affixes or new combinatorial restrictions? What is the relation between synonymous affixes, such as the ones indicating females in *–in* (*boerin* 'farmer's wife') and *–esse* (*secretaresse* 'secretary')?

Nederlab provides the means with which the researcher can survey and analyse all the derivations formed with a particular affix in the course of time:
1. An exhaustive ***list of Dutch affixes*** from all periods will be drawn up. At the moment, such a list exists only for the modern period.
2. ***Analysed datasets*** will be made available, consisting of words with a morphological analysis (that were or are used as training material for the development of morphological parsers).
3. Part of the corpus has been (semi-)automatically ***tagged morphologically.***
4. Several ***morphological parsers*** will be made available for modern Dutch and the 18th- 19th centuries. For the older material, parsers will have to be developed: a language-independent, learning parser can be used for this, or existing tools for medieval materials can be expanded by morphological modules such as Adelheid (developed at RU/MI for inflexion and spelling variation), and INPOLDER (for syntactic parsing).

2.1.3 Detecting new patterns

Many research questions in the humanities are not concerned with new concepts or word forms so much as with new word combinations, new patterns. Literary scholars, for example, are interested in the question what new word combinations were invented by the writers of the 'Eighties Movement', or what new rhyme words were introduced by particular poets. For linguists, the emergence of new patterns offers a way to recognize changes in word meanings. Thus, in earliest Dutch, the motion verb *varen* was combined especially with Dutch equivalents of *waggon* or *horse* – it was not until the later Middle Ages that it appeared with names of vessels. This suggests that the meaning of the word *varen* changed from 'go, travel' in general to 'travel by water'. On the basis of word patterns, linguists can also detect grammatical changes, such as the change from a transitive to an intransitive verb.

Nederlab offers tools for textual analysis (see 2.4) which detect that from a particular moment, certain words appear regularly in collocations (more or less fixed combinations), whereas they had never been linked before. In this way, new metaphors and cases of 'framing' can come to light (see 2.4.3).

2.2 Research questions concerning the spread of changes in language and culture

Many research questions in the humanities are not (only) about the beginning of a change, but about its spread through the ages. Researchers try to discover what phenomena and what types of phenomena have spread over time – from one text type to another, from one author to others, from one dialect to another, or to the standard language – and what phenomena have disappeared, and how. On the basis of a comprehensive inventory, they try to explain the existing variation from various influences including borrowing.

Nederlab enables us to detect the spread of changes in texts. This is accomplished by measuring the frequency of phenomena, as the following cases show.

2.2.1 History of ideas

The study of the change of terms denoting historical concepts such as *burgerschap* 'citizenship', *nationaliteit* 'nationality', *Nederlanderschap* 'Dutch nationality', *verzuiling* 'social compartmentalization', *moderniteit* 'modernity' is a special branch of history. In literary research, the focus is on changing ideas about, e.g., 'romanticism' or 'sentimentalism'. Such reception studies have so far been based on literature study by hand, and by intuition. Nederlab helps to systematize this research and to give it a firm statistical basis, so that empirical findings may be checked against facts, which is certain to improve research standards and to bring many new insights.

The spread of a phenomenon appears from its frequency through the ages. Nederlab offers a large number of ***search programs*** with which researchers can find the occurrence of a particular term in a particular corpus or subcorpus. In order to find, in one search session, both the spelling and form variants of a term, researchers can search a text that is only lemmatized or via a computational historical lexicon (see 2.1.1.).

Nederlab offers ***visualization*** of search results through line graphs, bar graphs, circle graphs, or scatter graphs. Thus the frequency of a particular concept through the ages can be made clear at a glance. It is also possible to include the frequencies of several related concepts in one chart, so that they can be compared. This is rather like Google's Ngram Viewer (which at present does not contain Dutch texts), see J.-B. Michel, E.L. Aiden *et al.* 'Quantitative Analysis of Culture Using Millions of Digitized Books, in: *Science* December 16, 2010. Nederlab is, however, much more sophisticated than Google Books: within Nederlab, the frequency dates may be sorted on the basis of metadata. Thus, we can recover what text types (political, legal, literary etc.) are responsible for peaks in the use of a particular term in a given period, and we can then go on to establish a connection with social developments. If, for instance, *slavernij* 'slavery' has a peak in the second half of the 19th century, we can find a connection with national and international debates about the abolition of slavery. Another difference between the methodology of Aiden *et al.* and ours is that Nederlab will be Open Access: anybody can check the results of a Nederlab search. Aiden *et al.*'s work on the other hand is based on a way to access the Google Books database that is *not* open to outsiders – some of the authors of that work are Google employees.

2.2.2 Systematic charting of language changes

Linguists have for decades been looking for an answer to the fundamental question as to the system that underlies the loss of inflections in Dutch ('deflection'). Deflection is the phenomenon whereby word endings wear off, disappear or coalesce. As a result of deflection, the functions of cases and inflected forms (i.e. synthetic constructions) are replaced by analytical paraphrases with articles, prepositions and auxiliaries. Thus, for instance, *sconinx boec* ('the kings book') changed to *het boek van de koning* ('the book of the king'). Deflection also leads to a variety of shifts in morphology and in word order at the sentence level. Signs of deflection can be found in the earliest known stages of Dutch, and it has continued to the present day. Many language changes in Dutch have been somehow related to the deflection phenomenon.

Although in the course of time, a number of details about deflection in Dutch have been uncovered, many questions still remain unanswered – questions such as: when and in what dialect did particular changes start, how did they spread from one dialect to another, what are the points of stability and variation between the various Dutch dialects? Is there a systematic relationship between particular types of deflection and word order change? Have there been cases of acceleration, deceleration or even inversion in the deflection process? How have the various changes interacted in the course of time and what (linguistic or social) factors have influenced the process?

In order to learn more about deflection in Dutch, we shall have to answer a great many component questions – such as: how, where and when did every single article, pronoun, preposition and conjunction come to be used, and what was the further development? How did morphological cases disappear in the various Dutch dialects? When and where did auxiliaries like *hebben, zijn, worden, zullen* originate, and how have they developed over the years? How have impersonal constructions like *mi lanct na di* come to change into personal ones (*ik verlang naar jou* 'I long for you'*)*.

Changes in the verb endings of the preterite second person singular may serve as an example of how language changes can be measured in the corpus. We know that 'you heard' was expressed in earlier stages of Dutch by equivalents like *du hoordes, ghi hoordet, gij hoorde, jij hoorde, u hoorde*. What we do not know, however, is when exactly the changes in form took place, in what dialects they began, and whether or not a social component played a part in the changes: were some forms considered as polite and others as rude?

To find out about this, we will have to establish the frequencies of all verb forms in different syntactic contexts, and this will have to be related to the place (the dialect) they occurred in. Such syntactic research questions cannot be answered by checking for concrete forms; both the verb forms and the pronominal forms are variable and some of the variation is unpredictable. Nederlab now offers various expedients:

1. (semi-)automatical ***syntactic parsing*** has been applied to part of the corpus, that is: grammatical information has been added.
2. Several ***syntactic parsers*** will be provided, enabling researchers to annotate relevant parts of the corpus. For Modern Dutch, there are several syntactic parsers, some of them language-independent and self-learning; for medieval texts, INPOLDER will soon become available.

With the aid of the parsed material, researchers can now gather the relevant verb forms. Using the metadata, they can establish the dialectal spread; for a visualization, Nederlab offers several ***cartographical tools***, such as the CLARIN project Gabmap.

2.3 Research questions relating to the finding of (statistical) connections and networks

For all kinds of research questions in the humanities it is essential to recognize more or less fixed word patterns – that is how we can find, e.g., literary motives oropen ac word fields of related words. By charting the links between related phenomena, such as personal names and place names, networks can be uncovered. Such networks yield insight into the spread of knowledge, culture and language. Within a diachronic corpus, shifts and developments can be established in patterns, which in turn can point to developments in language and culture. Nederlab offers a range of tools that can establish links between words, so that patterns can be automatically recognized.

2.3.1 Establishing patterns and motives

Historians will want to know, for example, how ideas about astronomy changed as more sophisticated instruments were invented. Or how people's opinions about various religions changed in the course of time, or how their opinions about foreign countries evolved. Literary scholars want to find out how immigrant inhabitants of a colony are described in (post)colonial novels, with what motives they are associated, and what the variations are for each genre, region, author and/or period. To help find answers to such research questions, Nederlab offers several ***text mining or data mining tools.*** These establish semantic word fields for large quantities of texts, on the basis of proximity relations: the frequency with which words occur in collocation. Thus, they draw up a semantic field round the concept "astronomy" with words like *ster* 'star' and *planeet*, but also the much older word *dwaelder*. A semantic field around the concept of "religion" contains words such as *Mohammedan, islam, pagan*. Nederlab provides the possibility to visualize such fields in ***word clouds***, but also to distinguish them statistically and to monitor changes. In this way, significant changes in concepts can be measured or derived. For Modern Dutch, a semantically enriched corpus has been created at the Vrije Universiteit, DutchSemCor, which can be used as an aid in text mining. This is a more precise semantic interpretation of the text enabling researchers to link meaning to the collocational context, but also to search for concepts irrespective of form (e.g. both *planeet* and *dwaelster*, *dwaelder* and *hemellichaam* 'celestial body'). Such an enrichment can also be applied to the historical corpus. This offers further possibilities to follow meaning change and concept use through the ages or to add other annotations such as subjectivity.

Literary scholars use text mining tools and topical modelling to automatically recognize motives and subjects in stories, texts and songs. Within the Tunes & Tales project, pilot research is under way, funded by a KNAW grant, partly elaborating on the NWO-sponsored of Dutch Songs Online project.

2.3.2 Establishing subjective opinions

The advantage of research on the basis of a diachronic corpus is that large quantities of data can be examined quantitatively. However, many research questions in the humanities have a qualitative aspect: they investigate, e.g., subjective opinions such as the reception of a particular author, a particular work or a particular event. Or they might wish to find out how emotions were evoked in political texts, in sermons and on-stage in the 18th century. Or with what linguistic means a political support group was created in the 20th century. Or in what type of texts, in what period, and by what authors the strongest verbal aggression is attested.

To find out about such questions, Nederlab offers tools for ***sentiment mining***, such as WEKA. This is a collection of machine-learning algorithms for data mining tasks able to determine, on the basis of adjectives used, whether the sentiment of a document is positive, negative or neutral. This is usually done by checking the words that are used in a text against a list containing the sentiment values of those words – such a list will have to be drawn up by the researchers by hand, at least for the earlier periods. With the aid of this software, positive and negative reviews of literary works, for instance, can be analysed. Researchers can then filter the data in a variety of ways and compare the results.

2.3.3 Establishing the relations between persons and places

For literary reception studies and research into processes of reputation and canon formation, it is crucial to find out how often and by whom a particular author is cited in his own time and in later years. In this way, networks of mutually influencing authors can be discovered. Such networks can be visualized, for example, by correlating the weight of lines with the frequency of mention. Establishing the relations between characters in novels or in plays can also lead to interesting new research. Historians want to find out how people have in the course of time evaluated historical figures such as rulers, politicians or scholars.

For this type of research, personal names have to be identified. Names cannot be looked up reliably in an unstructured corpus – they are not unique: several persons may carry the same name at the same time. For research purposes, it is essential that individual persons be kept apart. Also, names have to be kept distinct from words: Mr Smith is not the same man as the blacksmith in a particular village.

Nederlab offers various tools for automatically recognizing and tagging personal names and placenames, tools for ***Named Entity Recognition***, which, however, have been developed for modern Dutch only. The tagged personal names are, wherever possible, linked automatically to bibliographical and biographical files, such as the Biografisch Portaal, parlement.com, the author database of the DBNL, the bibliographical files of the KB, and genealogical data files or international data collections of names such as Wikipedia or VIAF. Earlier placenames are, wherever possible, linked to current names of municipalities, which are in turn identified by ***GPS*** coordinates, so that cartographical tools can handle them – maps can then be charted of, for instance, the birth places of the principal authors of a particular period.

Nederlab also offers so-called ***Coreference Resolution tools,*** which automatically relate references to persons by pronouns ('he', 'her') or descriptions ('the chairman', 'the trainee') to the personal name intended. The tool was developed by the Department of Information Sciences at the University of Amsterdam for Staten-Generaal Digitaal, but can also be applied to literary works and stage plays.

Thus, eventually, an huge network of interrelated data will be established within the diachronic corpus, visualizing a multitude of connections and cross-connections, as an elaboration of the results of the NWO projects Circulation of Knowledge and Learned Practices, Historical Timeline Mining and Extraction (HITIME, a tool recognizing and linking historical events, persons, organisations, professions, time references and locations), and LINKS (linking system for historical family reconstruction), which aims at reconstructing all the genealogical families in the Netherlands from the nineteenth and early twentieth centuries on the basis of the data from the digitized registers of the Registry Office in Genlias. By linking a maximal number of data, investigators can answer important research questions such as: What influence did internal and external literary and cultural movements have in the Low Countries, how did they spread and what were the long-term effects? What contacts have there been between authors in the Northern and Southern parts of the Netherlands between 1100 and today, in what periods were these contacts most frequent and in what prevailing direction did they take place? These questions are not new, but it is now becoming possible for the first time to investigate them systematically, using a full and longitudinal database, and to test old hypotheses.

2.4 Research questions based on systematic text comparison

The fact that the computer is able to compare texts automatically has opened up a wide spectrum of new research questions. Investigators use text comparison to detect similarities in texts and thus discover plagiarism, quotations or paraphrases (2.4.1), to find differences between, e.g., dialects or authors' idiosyncratic usage (2.4.2), to recognize metaphors (2.4.3) or to determine the provenance of unidentified texts (2.4.4 and 2.4.5).

2.4.1 Detecting similarities in texts: plagiarism, paraphrases, quotations

Literary scholars are interested in the quotation behaviour of authors. To find out, Nederlab offers various tools. ***Plagiarism detection tools*** recognize literally borrowed fragments and quotations from other works. Since authors often borrow by modifying rather that copying, ***paraphrase detection tools*** are also on offer: tools that recognize variations on texts, such as DAESO (Detecting And Exploiting Semantic Overlap). However, this works for modern texts only. Scanning large quantities of literary works by plagiarism and paraphrase detection tools will reveal many new and unexpected cases of authors owing tribute to others.

Paraphrase detectors or intelligent search machines can be deployed to discover where and when more or less fixed combinations such as maxims, catchphrases, proverbs or aphorisms came to be used and how they were adopted by authors and in various text genres. In the case of quotations, it is possible to check whether the (supposed) original author is mentioned and what variations are used. This may throw new light on the reputation and the popularity of quotations and writers through the ages. This topic has never been systematically investigated for lack of source material.

Quotations from often-quoted works can be tagged and linked to one specific edition. This yields interesting results for, for instance, the Bible – the most-cited and most-quoted book in both literary and non-literary texts. Books, chapter and verse references vary considerably (Gen. 10:12, Gene. X, 12, etc.), and the division into chapters and verses has changed over the centuries. Within the corpus of the DBNL, a pilot study has attempted to automatically link all Bible references to the 2005 New Bible Translation. Such linking will open up various information, such as what Bible fragments were quoted most often in what periods and what passages are most popular among literary authors or journalists.

2.4.2 Detecting differences between types of texts, authors, dialects

Nederlab offers a number of language-independent tools to compile statistical text analyses. For each text, these ***corpus analysis tools*** establish textual features on the basis of numbers of types and tokens and the ratio between them, frequencies, numbers of syllables, word length, sentence length, most frequent word combinations, etc. These tools allow investigators to identify differences between specific parts of the corpus. Thus, a corpus of business-like texts, such as advertisements from a particular period, may be compared with literary texts from the same period, and conclusions drawn about differences in usage. Naturally, the results of the analysis program will have to be weighed by the investigators, since not all points of difference are equally important. The (reviewed) results, including the investigators' judgments, can be stored in Nederlab for future investigators.

Textual analysis can provide answers to a wide range of research questions. Competence linguists may investigate the nature and extent of the differences between Dutch as used in the Netherlands and in Flanders in the 21st century, and whether these differences have increased or decreased since, e.g., 1950. Theoretical linguists may test the proposition that the language of the Authorized Bible translation is representative of Dutch as spoken or written around 1637, by comparing the text to, for instance, a corpus of contemporary literary or scholarly texts or ego documents. Historians will want to compare political usage in the 21st century with that of the second half of the 20th century, in order to establish whether the language is becoming coarser, as has often been claimed. Book historians will find it interesting to examine the relations between copies or editions of a single work over time, or the relations between original classics and their adaptations for young readers. Literary scholars will want to investigate the migration of a speech category such as free indirect discourse (‘She wondered, should she go?’) from literary to journalistic texts.

In addition, Nederlab offers tools that measure the distance between various dialects and can thus be used for ***dialectometric analysis.*** These tools have been developed for modern dialects; they measure the relative distance between the various dialects and Standard Dutch. Researchers can adapt these tools within Nederlab to make them work for earlier periods of Dutch. This will allow them to measure the differences between Old, Middle and modern Dutch dialects, and the distances between these dialects and Standard Dutch. These comparisons will help researchers to answer questions like: How did Standard Dutch develop from the Middle Dutch and Old Dutch dialects? Did the western standard language influence the dialects or did the local regional standard serve as a substrate in the development of the standard in the cities of the province of Holland? How have dialects come to diverge or converge over time, and does this coincide with political changes?

2.4.3 Finding metaphors

Researchers from all disciplines in the humanities are interested in metaphors: literary scholars and competence linguists put forward research questions like: by which author, or in which literary movement, or in which type of text, was this particular metaphor introduced? In what way are metaphors blown up from minor local expressions into full-fledged text-organizing principles that turn a poem, or even a novel, into a thrilling exercise in metaphorical and innovative thinking? Linguists want to know what words can be used metaphorically and try to gather evidence for Lakoff’s and Johnson’s ‘embodied cognition’ postulate that human cognition is predicated on physical body shape. Many hold, moreover, that metaphor is an important mechanism in the levelling of meaning in concrete utterances, which makes it an essential component in the grammaticalization process all languages go through, but which has never been examined for Dutch in this manner. Historians investigate the introduction of a new metaphor because it frequently indicates a break in thinking. For example, an idea may be first conceptualized in a newspaper (dipsomania as ‘cancer of society’) before actual changes in society or politics come about, such as raising the tax on spirits. Conceptualizations of organizations and management, of politics and government, often spring from underlying basic metaphors (society as a person or a family; an organization as a plant or a tree that can grow and be pruned), the development of which can now be traced systematically in Dutch, in a broad range of communication domains.

Metaphors can now be detected in corpora and annotated with a high degree of reliability, thanks to NWO programmes such as those implemented at the Vrije Universiteit Amsterdam. Research has shown that the distribution of metaphors is closely related to the linguistic variety – conversation, written texts, etc. - in which they are found. It has been shown that the function of metaphors in specific registers has changed considerably over time owing to evolving and diverging cultural functions of the media used. As the use of metaphor plays a fundamental role in modelling all kinds of abstract concepts, it is important to investigate the role of the use of metaphor in this process, both its cause and its effect.

For the first time, this will enable us to consider the historical development of entire metaphoric complexes as a cognitively and culturally shared means of expression for the conceptualization of all manner of emotional, mental, social and cultural phenomena, and experiences which in their turn have also gone through a marked evolution over the past thousand years. Dutch, for instance, is rich in metaphors relating to water: where do they come from and how have they shaped our thinking and the public debate on a variety of issues, including the stream/wave/flood tide/tsunami of immigrants? When did the consultation and consensus model (Dutch ‘poldermodel’) arise, and how is it related to the historical development of our ideas on democracy, participation and culture? How are these ideas used and abused in the language of the media? Such metaphors and metaphorical models sometimes become the subject of intense public debate, resulting in more – or for that very reason - less acceptable ways of speaking about specific intellectual domains, so that they eventually come to have an effect on our personal world picture, on (political and other) choices we make, and on our behaviour. Similar processes take place in the financial sector (the collapse of banks), the environment (the greenhouse effect), and health care (AIDS as a plague). Nederlab will provide reseachers with the means to tackle all such issues in Dutch language and culture on an increasingly ambitious scale.

2.4.4 Localizing and dating an unknown text

In principle, every text within the corpus has a (rough) indication as to time, place, author and text genre. There are texts, however, for which one or more of these items are as yet unknown. These texts constitute interesting research cases. Linguists want to be able to assign a particular text to a particular place or region, for instance on the basis of dialect characteristics. For linguists and literary specialists alike, it is important to date a certain text as exactly as possible on the basis of accepted formal characteristics. Literary specialists aim to classify texts in a text genre, or assign them to a specific author or translator, on the basis of formal characteristics.

Nederlab enables researchers to determine texts whose date or provenance is unknown. Let us take the case of a 15th century text of unknown provenance. In order to localize this text, the reseracher will select, on the basis of the metadata ‘place’ and ‘time’, all the texts from the 15th century of which the localization is exactly known. These can be divided by hand into dialect groups: a 15th century Limburg dialect corpus, 15th century Brabant dialect, 15th century Low Saxon dialect, 15th century Holland dialect, and so on. A variety of decisions must be made: what dialect groups can be distinguished for the 15th century: Groningen, Drente, Twente dialects or rather more widely: Low Saxon, or more narrowly: Oldenzaal dialect? What concrete places belong in a specific dialect group? The researcher records the decisions made in metadata, thus enriching the corpus.

Each dialect corpus is then analysed by a corpus-analysis tool, establishing its textual characteristics. The researcher compares the features of the various corpora with each other and weighs them, deciding what words, forms, sounds and word combinations are characteristic of a particular region. The result, then, is a description of characteristic linguistic features of the dialects in a specific period, such as the 15th century. These dialect features can be charted with cartography software. The unknown text can be analysed and associated with the most closely corresponding dialect corpus. This will allow us to determine the provenance of a text more or less accurately.

The result of the investigation is not only that an ‘orphan’ text has been localized, but also that a part of the corpus has been enriched with metadata and that an existing tool has been adapted to a new research question, which has produced new, reproducible data/criteria. Thanks to Nederlab, other researchers can gain a precise insight into the research methods used by colleagues and, moreover, replicate the investigation for the 14th and the 16th century, or for other 15th century texts. In a similar way, a researcher can date an undated text within Nederlab.

2.4.5 Text-genre recognition

Genres are interpretation frames within which judgements of texts are formed; genres and genre classifications, then, have ideological power. Genres and genre classifications change over the years. Quantitative research creates the possibility for researchers to give the intuitive knowledge of genres and genre systems an objective basis, as well as correcting them, as has been attempted in the US by Moretti *et al*.

Trends and periods in literary history (including the recording of literary history) distinguish themselves by their own special terminology, and their own themes, which are expressed in the various genres (novels, poetry, plays, book reviews, essays, historiography). The genre system (including genre indicators) and/or the hierarchy of genres may vary as well. Diachronic research on the development of particular types of texts enables researchers to detect changes without losing sight of historical diversity and synchronicity.

An explanation for the changes in genres and genre systems can be sought along several axes. The influence of foreign developments, for instance, can be traced with a comparative quantitative analysis of semantic fields. Moreover, genres and genre systems change under the influence of social developments, so that the study of genre formation is a method par excellence to map the interaction between literature and society. One example is the girls' novel, whose plot structure appears to have changed, from the first wave of feminism, under the influence of the changing position of women in society. Or, to mention a recent example: the so-called 'literary thriller' is a newcomer in the literary genre system. When does this genre first appear? Was it invented by publishers? Does this mean that the domain of literature is coming to be defined, more than in the past, by commercial parties, at the cost of critics and literary scholars? Such research fits in with the KNAW-eHumanities project 'The Riddle of Literary Quality', where the object of research is the formal characteristics of the literary genre.

Description and motivation of the investment

Motivation

A great deal of public funding has been - and is being - put into digitizing corpora and developing tools for searching, analysing and editing them. Applying statistical methods of analysis, in combination with the introduction of scientific standards of verifiability and recursion, has for some time been heralding a radical innovation of humanities research. There is, however, an obvious gap between the available means and the actual research strategies of the average humanities scholar, so that the result falls short of the promise. It is Nederlab's ambition to resolve these problems in one fell swoop.

The problem of the diachronic corpora[2] is that they are located in different places and reside in different institutions. DBNL, Huygens ING, INL, KB and the Meertens Institute all house large corpora, while smaller corpora are provided on websites of universities or individual researchers. These corpora can now only be searched and analysed separately, not simultaneously and comprehensively. Moreover, the search interfaces and search possibilities vary from one corpus to the next, and there are considerable differences in quality between the various corpora, while each institution adds its own metadata. It is imperative that these separate corpora should be searchable as a single entity in order to be able to answer longitudinal and cross-disciplinary research questions. Nederlab's greatest challenge will be to draw up an inventory of the available corpora and make sure that all historical corpora allow distributed searches at text and metadata levels (with the help of a metadata harvester). The addition of uniform metadata will quickly ensure improved access to existing corpora, which will boost the need for quality improvement and standardization. The issue of quality improvement is made all the more pressing by the danger of Google becoming the leading player in the field of corpus quality standards. However, the OCR results delivered by Google up to this point fall well short of the standards required for scholarly research, making them unfit for statistical and analytical research. Furthermore, Google's efforts are at this moment only semi-open: although the scanned books can be searched in various ways, as predefined by Google, researchers cannot access the corpus directly to find aswers to their own questions. It is important for students and researchers that Nederlab should provide a reliable corpus for research and the possibility to correct and annotate doubtful texts (or to have it done).

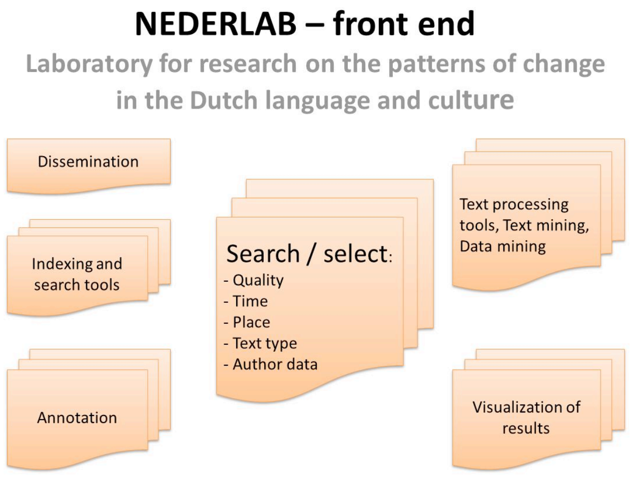
The tools pose a further problem. At the moment, many different tools are being developed which eventually should become available through CLARIN-NL. However, these tools are often suitable only in the context for which they were developed (e.g. for morphological enrichment of fourteenth-century texts that have been pre-processed in a specific way). Nederlab – in close consultation with CLARIN - aims at gathering together existing tools in a central location and adapting user interfaces in such a way that, within the available infrastructure, they can straightaway be used by technically inexperienced researchers and can be applied to diachronic texts. To attain this objective, Nederlab will subscribe to CLARIN's standards of formats and metadata.

In doing so, Nederlab also builds on the many initiatives regarding corpus construction and tool development set in motion by the Dutch government and research institutions such as KNAW and NWO, but adds an important dimension and added value: a user-friendly infrastructure for students and researchers, boosting cooperation and synergy in Dutch studies and leading to the formulation of new, mostly interdisciplinary, research questions. In this way, an immense surplus value is created by means of a comparatively small investment, in

[2] See also 'The digital drama' in *NRC Handelsblad,* Wetenschap, 2011-10-07.

keeping with the prospects outlined by Euro Commissioner Neelie Kroes in the introduction to the *Riding the wave*[3] report: "My vision is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make use of data. Data become an infrastructure that scientists can use on their way to new frontiers." Nederlab will automatically lead to standardization of metadata and formats in corpora, as will be shown below.

Technical description


NEDERLAB – front end
Laboratory for research on the patterns of change
in the Dutch language and culture
Dissemination
Indexing and search tools
Annotation
Search / select:
- Quality
- Time
- Place
- Text type
- Author data
Text processing tools, Text mining, Data mining
Visualization of results

Front end and back end

Nederlab will function as a Virtual Research Environment for diachronic research. The front end runs inside the user's web browser. It presents a transparent and efficient user interface that will be designed to meet the specific needs of end users (reseachers, students). In this respect, Nederlab differs from existing user interfaces such as those currently offered by CLARIN, which mainly focus on the supply of tools and data. The front end will present tabs or pulldown menus for the various functionalities: a tab for dissemination (with forum and help desk), a tab for navigating, searching and selecting subsets of the corpus, for result visualization, for annotation, and finally a tab for executing and monitoring text-processing tools. The back end runs on multiple servers in a distributed service environment.

Basic services

Each Nederlab user will have access to a number of basic services - depending on the type of user - and will be able to browse and search the diachronic corpus. The texts for the diachronic corpus are made available by a number of scientific libraries and institutions (see Appendix 2). Within Nederlab, texts and metadata of the various institutions will be interlinked and extended by a quality information layer. This will facilitate finding suitable texts, thus providing a wealth of added value to both Nederlab users and corpus suppliers. The quality of these collections is quite variable. While constructing the diachronic corpus, we will have to make the user aware of the full range of subcorpora, as well as the variability in quality, in order to enable the researcher to select the right materials for high level research. We shall also promote homogeneity of the data by means of interaction between data providers and users. This can be achieved by creating a separate administrative heading for each bibliographical entity (e.g. independent titles, issues of journals, dependent titles (fragments, articles)) while constructing the diachronic corpus. This will be done in the form of an xml document which will simultaneously serve as the 'container' for the digitized document. The containers will also function as the repository for metadata and tagging added by researchers at a later stage.

[3] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

A virtual workspace will be created for researchers working in Nederlab, allowing them to collect data, share data with other Nederlab users, and call up selected processes to work with the data. Researchers are given the possibility to upload their own datasets to which relevant metadata information can be added, or they may select data from the diachronic corpus. Data in individual workspaces may be shared with other researchers in order to allow collaborative work. Other researchers may propose modifications to the metadata information or may use the data for further processing. Modifications of the metadata will be subject to approval by the owner of the metadata document, i.e. the researcher who originally created the metadata for the resource. When data is further processed, the results, i.e. the secondary data, are stored in the researcher's workspace. The original document is thus retained in its original form and it should be possible to relate the secondary data to the original document using provenance data. Application of subsequent processes (e.g. lemmatization→ PoS tagging→ NER recognition) results in a series of secondary data documents, where the result of each individual step may be inspected separately by the researcher. The types of processes made available to a researcher can be customized. Nederlab will provide a number of standard services or workflows from which the researcher can select those tools that are relevant to his or her type of research. These preferences are stored as part of the workspace.

The diachronic corpus will be extended by contributions from participating researchers. Material added to the corpus is subject to the approval of an editorial team, assigned by the supervisors of the data curation track. The formal submission and approval process will be accommodated in the Nederlab environment and proper deposition arrangements will be made to ensure that new corpus data is located at suitable data-archiving centres for long-term storage.

Authentication and authorization

Authorization and authentication are key pillars for any infrastructure; workspaces, services and data must be protected from unauthorized access and researchers must be able to authorize other researchers to access their data and metadata to facilitate collaborative work. Nederlab must ensure that proper authorization procedures are upheld across all areas, for example for searching across documents. By making research results publicly available researchers may publish research results that have not (yet) been archived at participating archiving institutes.

Authorization settings are also relevant when formally submitting new research data for inclusion in the corpus: access to metadata and research data is transferred from the researcher to the editorial team in the process of preparing the materials for inclusion in the corpus. The project will build upon experience gained in projects such as CLARIN-NL, CatchPlus and SURF, which are currently running projects addressing various aspects of authorization and authentication. Authentication will take place through the SURFFederation and Single Signon features will be used as supplied by Shibboleth, although support of other schemes may be considered, if necessary.

Administration

The Administration module will provide standard administrative features such as adding and removing users or groups from the Nederlab environment. As part of the general user information, Nederlab will make it possible to assign different user roles by distinguishing different degrees of authorization in the Nederlab environment, such as researcher, editor, administrator. These authorization levels are reflected in the user's environment, providing different types of functionality to the end user.

Workspace management

Workspaces maintain information related to a user in a coherent manner. This includes personal research data, references to research data in the diachronic corpus, user preferences (e.g. tool preferences) and administrative data such as the indices. The appropriate graphical management tools will support workspace management. Users may select tools they wish to incorporate in their personalized Nederlab environment and have the possibility to upload data, organize data and manage access restrictions to data stored in their workspace. The Nederlab environment will also use the workspace to add user information to the workspace, such as indices on workspace data, to allow the smooth operation of the environment. Collaboration with SARA/BigGrid will be sought for storage of workspace data.

Metadata

All research data in Nederlab, whether residing in workspaces or as part of the diachronic corpus, will be accompanied by metadata descriptions. To provide for a larger variety of metadata descriptions, both in terms of descriptive fields and structure, a flexible and extendable metadata framework is needed in order to allow diversification of the metadata descriptions into various fields of research and projects. The metadata descriptions that are produced are also relevant outside the Nederlab environment, so appropriate publication procedures and technology must be available allowing other users from different domains to find materials relevant to their research. The CLARIN project uses a flexible metadata framework (CMDI) that provides the flexibility to produce arbitrary metadata descriptions which are semantically interoperable using the ISOcat Data Category Registry. The CLARIN infrastructure also provides the necessary technological framework enabling publication of metadata on a large scale while using OAI-PMH. Nederlab will adopt the CMDI framework for creating metadata descriptions, and it will provide editing tools tailored towards the use for Nederlab metadata profiles, to ensure that the diachronic corpus can be harvested with OAI-PMH. This ensures that the data and metadata created for the corpus are made available in a standardized manner, while cooperation with CLARIN will make the information available at a European level.

Submission process for research data into the corpus

Nederlab will adopt a procedure for accepting research data into the diachronic corpus. Any researcher may propose new data which is to become part of the corpus or propose modifications to the existing corpus data or metadata. Such a request will be subject to approval and acceptance by the editorial team and the Executive Board. Each request must contain a description of its relevance to the corpus and include the subcorpus to which the new data is to be added.

Nederlab will provide adequate feedback on the status of the proposed research data throughout the submission, evaluation and approval process and facilitate the necessary authorization settings. The editorial team will be provided with a working environment adequately supporting evaluation and validation procedures. After successful validation and approval, the submitted research data will be added to the corpus and archived at participating archiving centres.

Searching

Nederlab will offer search functionality across the diachronic corpus and researchers’ workspaces. Two search domains are distinguished: metadata search and content search, which may be presented to the end user in a unified search interface. The metadata search domain contains all metadata documents from the corpus. These will be stored in one or more CLARIN centres responsible for publishing the metadata in the wider CLARIN infrastructure. This ensures that the metadata can be harvested and are made available through CLARIN’s central Virtual Language Observatory[4] and related search engines. In addition, metadata documents in researchers’ workspaces which have been made publicly available will be collected centrally and can be made available within Nederlab with the technology of CLARIN, thus providing a metadata search domain for all publicly available metadata documents in Nederlab. For each user this metadata search domain is to be further extended by metadata documents from their private workspace and by those they are permitted to access in other researchers’ workspaces. Search indices created in this process can be stored as part of the researcher’s workspace. The metadata search domain thus covers three areas: corpus metadata, public metadata and private/shared metadata.

For content search on workspace data, the situation is complicated by the fact that the residing data are not yet part of any institute’s archive and thus can not be made accessible through any specialized content search engine. It will be important to provide a solution for Nederlab users. One possible strategy, yet to be further developed, is to relate various resource types to the various quality layers proposed for the diachronic corpus. For specific format types Nederlab can implement specialized indexing strategies providing more fine-grained access to

[4] http://catalog.clarin.eu/ds/vlo/

various enrichment levels, e.g. morphosyntactic or syntactic information. This will encourage researchers to adopt standard formats for their resources during the various stages of their work, which will result in more homogeneous contributions to the corpus. The search indices for each workspace may be stored as part of the researcher's workspace. The content search domain also covers three areas; corpus data, publicly accessible Nederlab data and private/shared data.

To help and support the end user, particular attention should be paid to user-interface design to guide the end user in the choice of the various search options that will be made available.

Tools

Nederlab will provide a range of standard tools that may be used by researchers to further process parts of the corpus or workspace data. A simple user-friendly XML-editing environment will be set up for creating new material as well as annotating existing material. Nederlab will incorporate standard tools made available through projects such as CLARIN-NL, CATCHPlus. Where necessary and possible, adaptations will be made to configure these tools to cover relevant time periods or selected use cases. Since most tools are currently delivered as web services, they can probably be combined into more complex workflows. This makes fully preconfigured production processes available to the researcher, e.g. text-to-named entities, without the need to have a deep knowledge about the full complexity of the individual processes. Tools and processes will be made available centrally to researchers who may browse and select tools/processes to be incorporated into their workspace environment. All tools and workflows will be described using CMDI metadata. Collaboration with SARA/BigGrid will be sought for delivering services through their cloud infrastructure.

Archiving

Nederlab will produce complex information structures in terms of data structures, formats and resource versions. Version control and archiving are two support activities that require special attention to guarantee full availability of the diachronic corpus data in the future. Nederlab will sign an NWO-DANS data contract with DANS. This contract serves to guarantee sustained quality and accessibility of the data in a repository that has been assigned the Data Seal of Approval. Archiving of enriched data and long-term data management will be provided by The Language Archive (TLA) of the MPI, as part of the cooperation agreement between TLA and KNAW. Consultancy on archiving quality guidelines will be obtained from DANS and SURF. Persistent identifiers will be created for data and metadata versions, which will make them uniquely identifiable and referable. Metadata will be stored separately from the ultimate source as a derivative, enriched source. The source code (converters, editing environment, user interface, etc.) will be made available as open source software.

Harmonization and standardization

There is currently a high degree of variation in the formats used, in particular in the way annotation levels of linguistic content are expressed. Although a number of standardization activities such as the LAF (Linguistic Annotation Framework; ISO/FDIS 24612), MAF (Morpho-syntactic Annotation Framework; ISO DIS 24611) and SyNAf (Syntactic Annotation Framework; ISO CD 24615) are designed to produce a framework for representing linguistic annotation of language data, these have not been widely adopted by the community or by tool builders. Recently, Linked Open Data and Open Annotation Collaboration, using Linked Open Data principles, have gained widespread attention. Nederlab proposes to combine the best practices from these approaches to come to a coherent framework for representing both manual and automatic annotations in the supported Nederlab format.

Social Relevance

The diachronic corpus of Dutch (8th century – today) is the collective textual part of our Dutch national heritage; in fact, it is for the common national good that this corpus, with its wealth of spatio-temporal information, should be made available in one location not only to scholars and students but also to anyone interested in Dutch heritage, professionals and the general public alike. Both the Dutch government and the European Union attach great value to digital longevity and access to digital information. Nederlab will facilitate the accessibility of our digital memory, conforming to the goals and action plans of the *Digital Agenda* for Europe (http://ec.europa.eu/information_ society/digital-agenda/index_en.htm). Some minor adjustments will make the corpus suitable for teaching purposes and, given its overlay of refined metadating and tagging, it can be related to end products like dictionaries, encyclopaedias and navigation systems, which will undoubtedly boost product innovation in publishing.

National context of the facility

The research questions specific to the history of Dutch language and culture, which can be answered with the help of Nederlab (cf. 1. The Goals) link up with recent developments in the humanities and present materials for fundamental theoretical issues in the KNAW and NWO cognition and eHumanities programmes. Nederlab will yield fresh insights on important issues such as the nature/nurture debate, national cultural identity, cultural integration, the origin of canons and the dissemination of knowledge, culture and language. All these subjects figure prominently in the Dutch Scientific Agenda presented by the KNAW in May 2011.

Nederlab is meant as a specific research instrument. It will be built on existing technologies and knowledge, which will be adapted specifically for Nederlab. As basic generic technical infrastructure Nederlab will utilize the technologies made available by the large (inter)national infrastructure programmes CLARIN and DARIAH (including the new roadmap proposal CLARIAH), adhering to the principles of international standards for data management and infrastructure (for a survey of the Dutch eHumanities infrastructure see Appendix 3). On this generic infrastructure, Nederlab will build its virtual user environment for a targeted research community. For research purposes Nederlab will provide tailor-made tools, which will be developed in such a way that they can be used within the CLARIN-DARIAH infrastructure. Finally, Nederlab offers researchers a diachronic corpus to analyse. The data that together form this diachronic corpus are provided by large data providers such as the National Library of the Netherlands (KB), University Libraries, DBNL and INL, who all have agreed to support Nederlab. Within Nederlab the data and metadata of the data providers will be interlinked and be made amenable to distributed search. In this way, Nederlab will allow researchers to simultaneously analyse corpora that are located on different websites, thereby increasing the quality of the data supplied by the various data providers.

Nederlab will thus occupy a special and unique position among the large (inter)national infrastructure programmes and large data providers. This unique position allows us to build specific applications for searching and organizing Dutch texts, in accordance and in cooperation with both the international infrastructure that is now being developed and the data providers, who make more and more texts available in a format relevant for research. The five prospective Dutch CLARIN-centres (DANS, Huygens ING, INL, MI, MPI) are crucially involved in Nederlab.

Nederlab is thus the first joint platform for humanities scholars, corpus providers and technicians. Researchers, students and the interested layperson will be able to sift through all existing diachronic corpora via Nederlab from a single central location. The foundation of Nederlab will have an important consequence: all the parties involved will reach agreement on standardization and harmonization. To achieve this, Nederlab cooperates with Dutch tool developers, such as the NWO-supported projects Catch and CatchPlus intended to make our cultural heritage accessible, the European IMPACT project on the improvement of optical character recognition (OCR), and with information specialists and language technologists of all Dutch universities.

Nederlab will be cast in a dynamic set-up and will naturally adopt open access and open source as a matter of principle. For copyright or IPR texts general arrangements will be made with rightful claimants or their representatives, so that data may become available for research, possibly by way of subscriptions or licences (paid for by institutions).

Nederlab claims to fulfil an important educational function for researchers and students. This is why during the grant period a great deal of attention will be paid to the dissemination of information. A first version of Nederlab will see the light as soon as possible, within a year after commencement of the grant period. An educational programme containing simple tools - 'tools light' - will be part of it. These tools have fixed basic functions which may receive layered extension and fine-tuning by researchers. Concrete examples will be provided of the operation of the tools, in the form of workflows of which the intermediary stages can be stored so that they can be used by other reseachers as well.

A helpdesk will be set up for the duration of the project, and will be maintained after the project period as part of the future maintenance plan. A digital forum for researchers to consult with each other will be incorporated in Nederlab. Master classes and annual summer schools will be organized in all partaking institutions, warranting optimal use of Nederlab by researchers, students, PhD students and postdocs. Frequent contacts with researchers will also clarify further desirables, and bottlenecks or difficulties for researchers; these can be met or solved during the grant period in mutual interaction.

Relation to other research groups/centers

The organization chart shows that Nederlab receives support from a large number of Dutch universities and KNAW institutes (see also Appendix 1).

Organization Structure

Nederlab will be governed by an **Executive Board (EB)**, supported by a national coordinator. The EB consists of representatives of the three institutes guaranteeing the permanent hosting, maintenance and curation of the project: The Meertens Institute (prof. dr. H.J. Bennis), Huygens ING (dr. H. Wals), DBNL (C.A. Klapwijk), and the national coordinator (dr. N. van der Sijs). The EB will be responsible for the day-to-day governance of the project. The board members will oversee on-going activities through regular meetings with track supervisors and chairs of the advisory boards, who, together with the EB, form the **General Board (GB)**. The GB will rule on the yearly budget and work plan; it will meet once a year.

Activities are divided into a research, technical and corpora track (for details see Management plan):
Track 1: scientific embedding: supervision The Meertens Institute/Utrecht University (prof. dr. L.C.J. Barbiers) and Uva (prof. dr. J.C. Kennedy);
Track 2a: infrastructure: supervision The Meertens Institute (user interface, workflow, general technical coordination of Nederlab; ir. M. Kemps-Snijders) and Huygens ING (work space; ing. R. Haentjes Dekker);
Track 2b: tools adaptation: supervision Radboud University (prof. dr. A.P.J. van den Bosch) and Groningen University (prof. dr. ir. J. Nerbonne);
Track 3: data curation: supervision DBNL (texts; dr. R. van Stipriaan) and INL (lexical data; lic. K. Depuydt).

Furthermore there will be four advisory boards:
1. **Scientific Advisory Board** (SAB): prof. dr. A.M.C. van Kemenade (RUN, chair); prof. dr. G.J. Dorleijn, prof. dr. J. Hoeksema, prof. dr. B.A.M. Ramakers (RUG); prof. dr. R.A.M. Aerts, prof. dr. J.B. Oosterman, dr. M. Rem (RUN); prof. dr. R. Bod, prof. dr. J.T. Leerssen, prof. dr. T.L. Vaessens, prof. dr. F.P. Weerman (UvA); prof. dr. J.L. Goedegebuure, prof. dr. T. van Haaften, prof. dr. H. te Velde, prof. dr. A. Verhagen, prof. dr. M.J. van der Wal (UL); prof. dr. G. Buelens, prof. dr. W.W. Mijnhardt, prof. dr. E. Stronks (UU); prof. dr. I. Leemans, prof. dr. B.J. Peperkamp, prof. dr. G.J. Steen (VU); dr. K.H. van Dalen-Oskam (Huygens ING); dr. G.J. Postma (Meertens Institute); prof. dr. A.P. Versloot (Fryske Akademy).
2. **Technical Advisory Board** (TAB): prof. dr. F.M.G. de Jong (UT, chair); prof. dr. G.J.M. van Noord, prof. dr. L.R.B. Schomaker (RUG); dr. N.H.J. Oostdijk (RUN); dr. M. Bouwhuis (SARA); dr. M.W.C. Reynaert (TU); prof. dr. A.P. de Vries (TUD); dr. ir. J. Kamps, dr. M. Marx (UvA); dr. J. de Kruif, prof. dr. J.E.J.M. Odijk (UU); prof. dr.

P.T.J.M. Vossen (VU); dr. P. Doorn (DANS); drs. J.J. van Zundert (Huygens ING); dr. J. de Does (INL); ir. D. Broeder (MPI).
3. **Corpus Advisory Board** (CAB): dr. J. Beeken (INL, chair), dr. W. van Bergen (UBL), dr. C. Cucchiarini (Nederlandse Taalunie/Dutch Language Union), drs. P. Doorenbosch (KB), drs. M. de Niet (DEN), drs. J.F. Oomen (Beeld en Geluid), drs. M. Slabbertje (UBU), dr. B. Zeeman (UvA).
4. **International Advisory Board** (IAB): prof. dr. F. Willaert (Antwerp, chair), drs. T. Roselaar (IVN); prof. dr. J. van Keymeulen (Ghent), prof. dr. E. Leijnse (Namen), prof. dr. J. Tollebeek (Louvain), prof. dr. W. Vandenbussche (Brussels), lic. E. Vanhoutte (Ghent); prof. dr. R. Grüttemeier (Oldenburg), prof. dr. M. Hüning (Berlin); prof. dr. J. Pekelder (Paris); dr. R. Vismans (Sheffield); dr. R. de Bies (Paramaribo), dr. K. Groeneboer (Jakarta), prof. dr. R.B. Howell (Madison WI, VS), prof. dr. D. Prinsloo (Pretoria, South Africa), prof. dr. R. Severing (Curaçao).

The Scientific, Technical and Corpus Advisory Boards will meet once a year (or more often if necessary) and will provide solicited and unsolicited advice. The TAB is concerned with activities in Track 2a (infrastructure) and Track 2b (tools adaptation). TAB members will test the infrastructure and the tools. The SAB advises on the choices to be made in Track 1 (scientific embedding). The pilot projects floated for testing and supplementing infrastructure, tools and corpora will be adjudicated by selected members of the SAB and/or TAB, in consultation with Track 1 supervisors and the SAB chairperson. The CAB will be consulted in matters concerning Track 3 (data curation). Finally, the International Advisory Board will see to it that the Dutch diachronic corpus also contains Dutch texts written outside the Netherlands and Flanders, and that the Dutch diachronic corpus links up with diachronic corpora of other, especially neighbouring, countries, thus enabling research on contacts and influences within Western Europe.

International context

Nederlab's diachronic corpus of Dutch is also of international importance. The present submission is supported by the Dutch Language Union, the International Association of Dutch Studies (IVN) – both represented in one of the advisory boards - and by CLARIN-ERIC. Flemish scholars in the humanities will concretely support this application by submitting their own proposal to the Hercules Foundation for the compilation of a corpus of ego documents from the 15th and 16th centuries in the Dutch language. Thanks to the Flemish application, the gap in the digital diachronic corpus for this period will be filled.

The diachronic corpus of Dutch is interesting to foreign researchers for its contents because it does not only contain Dutch language texts from The Netherlands and Flanders, but also from Surinam, The Antilles, America and Indonesia. Collaboration with South Africa, where a diachronic corpus of Afrikaans is being put together, has been established.

Nederlab's infrastructure may set the tone for research in this area in other parts of the world. Nowhere is there an infrastructure comparable to Nederlab. The English-speaking language area clearly leads the way in offering easily searchable and linguistically enriched diachronic corpora, which are, however, limited in scope (cf. http://www.helsinki.fi/varieng/CoRD/index.htm). Furthermore, there is the Stanford Literary Lab, founded by Franco Moretti, where pioneering work is being carried out in literature. In other countries the initiatives are mostly on a smaller scale: a limited diachronic corpus is available to subscribers in France (http://www.frantext.fr/). Germany is working on the compilation of a representative diachronic corpus of the entire historical period. All kinds of initiatives are afoot in other European countries: a survey of synchronic and diachronic corpora available in various languages can be found at http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links-en/korpora.

The most conspicuous difference between Nederlab and the initiatives abroad is that all foreign diachronic corpora contain a *selection* of available texts from various periods. Nederlab, however, offers *all* historical texts that are - or will become - digitally available. It is the explicit and unanimous wish of the Dutch research community that no restrictions should be imposed on the corpus if it is to meet all research questions from all quarters.

The start of Nederlab in 2013 might mean that it could become a blueprint for foreign research institutes; this would have the additional advantage that foreign diachronic corpora can be linked automatically to the Dutch corpus, adding a new dimension to diachronic research: language boundaries and political frontiers do not always overlap, and, diachronically speaking, Dutch must surely be viewed as a continuum of dialects going beyond national borders. Moreover, many historical and literary phenomena originated abroad and spread to the Low Countries, some of which were ‘exported’ again later.

Local context

On page 23 of the KNAW Strategic Agenda 2010-2015, the important goal of launching new information technologies for modernizing institutes in the humanities is referred to: it is the Academy’s ambition ‘to have her institutes play an initiating and course-setting role in creating and maintaining infrastructural ict facilities and the application of these facilities in scientific research’. In order for this ambition to gain form and substance, the two KNAW institutions The Meertens Institute (coordinator) and Huygens ING will be the prime movers in founding Nederlab. These two institutes, with the DBNL/Nederlandse Taalunie, have agreed to support Nederlab’s upkeep after the grant period has expired, and will provide for server room. Arrangements will be made to keep up the helpdesk that is set up during the project period. Within Nederlab’s web interface available diachronic corpora will be linked at text and metadata level, which will not only lead to better access to existing corpora, but also to quality improvement and standardization of data and metadata. Links and quality improvements are lasting results of the set-up of Nederlab that will be stored by way of persistent identifiers. Archiving of enriched data and long-term data management will be provided by The Language Archive (TLA) of the MPI, as part of the cooperation agreement between TLA and KNAW.

Budget

Requested NWO funding						Total
Item description	2013	2014	2015	2016	2017	
Track 0						
National supervisor (0.8 fte)	47,2	48,1	49,1	50,0	72,2	266,6
Helpdesk/assistance (0.8 fte)	42,3	43,1	43,9	44,8	64,7	238,8
Servers & data archiving	15,9	16,4	23,7	23,7	31,0	110,7
Track 1						
3 pilot studies closed call		90,0				90,0
6 pilot studies open call			90,0	90,0		180,0
International workshop			15,0			15,0
Edited volume					5,0	5,0
Tutorial and other dissemination					5,0	5,0
Track 2a						
Software developer (4.6/2.6/2.5/1.9/1.9 fte)	301,4	167,6	174,1	127,2	183,7	954,0
Track 2b						
Software developer (0.5 fte)	31,6	32,2	32,8	33,5	48,3	178,4
Postdocs (0.5/0.5/0.5/0.5/0.5 and 0.5/0.5/0.5/0.5/0.0 fte)	59,0	60,2	61,4	73,0	45,1	298,7
International workshop					15,0	15,0
Publication (book & software)					5,0	5,0
Track 3						
Editor (0.2/0.2/0.1/0.1/0.1 fte)	12,6	12,9	6,6	6,7	9,7	48,5
Senior copy editor (0.5/0.5/0.0/0.0/0.0 fte)	26,4	31,4				57,8
Copy editor (1.8/1.8/0.8/0.8/0.8 fte)	79,2	88,1	36,5	37,2	53,8	294,8
Postdoc (0.8 fte)	47,2	48,1	49,1	50,0	72,2	266,6
Software developer (0.4 fte)	25,3	25,8	26,3	26,8	38,7	142,9
Reprocessing data files	40,0	50,0	50,0	50,0	30,0	220,0
Total	728,1	713,9	658,5	612,9	679,4	3392,8

Co-funding						Total	By whom
Item description	2013	2014	2015	2016	2017		
Executive Board (0.3 fte)	36,3	36,3	36,3	36,3	36,3	181,5*	MI, Huygens, DBNL
Track supervisors (0.8 fte)	80,1	80,1	80,1	80,1	80,1	400,5*	Universities
Total	116,4	116,4	116,4	116,4	116,4	582,0*	

*Estimated real costs (according to NWO standards, salary scale 11, total sum would be 366,7).

Case for the requested budget

Nederlab: Working plan 2013-2017

Track	**Description**
Track 0	**General organization**
Supervision: Executive Board: prof. dr. H.J. Bennis, C.A. Klapwijk, dr. N. van der Sijs, dr. H. Wals	
National supervisor	Project management, monitoring (production and technical) coordination of the four tracks, Dissemination of information among researchers, Writing educational programme (0.8)
Assistant	Assistance in project management and practical organizational matters such as helpdesk, pr, master classes/work shops/summer schools Support for the national and track supervisors (0.8)
Track 1	**Scientific embedding**
Supervision: prof. dr. L.C.J. Barbiers, prof. dr. J. Kennedy	
	Send out closed call for three pilot studies: history, literature, linguistics (2013) Evaluation of closed call Supervision of three pilots; coordination of feedback from pilots to software developers; evaluation of pilot studies; advice on publications based on the closed call projects (2014) Send out open call 1 Evaluation and selection of three project proposals from open call 1 (in the three fields) Supervision of three pilots; coordination of feedback from pilots to software developers; evaluation of projects; advice on publications based on open call 1 projects Workshop: scholars present the results achieved on the use of the infrastructure (2015) Send out open call 2 Evaluation and selection of three project proposals from open call 2 (in the three fields) Supervision of three pilots; coordination of feedback from pilots to software developers; formulating requirements for infrastructure; advising on publications based on open call 2 (2016) Coordination of feedback from scientific projects Book publication on the project, jointly with the software developers Organization of courses, tutorials, closing workshop (2017)
Track 2a	**Infrastructure**
Supervision: ir. M. Kemps-Snijders, ing. R. Haentjes Dekker	
Software developers	Harmonization of data formats and standards to synchronize existing data formats for different annotation layers and relate these to existing standards (1.1 fte) Indexing module to manage indexing process of metadata and data. Metadata and data must be indexed at diachronic corpus level, public document level and private document level (1.1 fte) Workspace module to manage access to private workspace and data management activities (2.3 fte) User administration and authorization module to manage common user information (0.5 fte) Authorization module to manage authorization information on metadata and data resources (0.5 fte) Search/Browse module to provide search/browse capabilities on previously created indices (2.7 fte) Metadata editors for different profiles in the diachronic corpus (0.4 fte) Metadata publication to handle publication of public metadata and data documents (0.2 fte) Tools integration: Integration activities for different tools (2.2 fte) General UI: Nederlab user interface activities (2.5 fte)

Track 2b	**Tools adaptation**
Supervision: prof. dr. A.P.J. van den Bosch, prof. dr. ir. J. Nerbonne	
Software developer + postdocs	Integration of general tools for text cleaning & normalization (OCR, spelling variation), with a focus on integration and interoperability* (0.5 + 1.5 fte)
	Integration of tools on lemmatization, morphology, PoS tagging (0.5 + 1.0 fte)
	Integration of tools on syntactic analysis (0.5 + 1.0 fte)
	Other NLP integration (0.5 + 1.0 fte)
	Maintenance, improvements, feature requests (0.5 fte)
	Organization of two ‘shared task’ sessions at CLIN or at an international meeting (task postdocs)

Track 3	**Data Curation**
Supervision: dr. R. van Stipriaan, lic. K. Depuydt	
Editor	Design of record layout ‘xml containers’ & management environment° (0,2 fte)
	Quality assessment (0.5 fte)
Senior copy editor	Implementation management environment in DBNL-workflow (0.2 fte)
	Planning & control of (re)processing production (0.4 fte)
	Quality assessment (0.4 fte)
Copy editors	(Re)processing of corpora (see Appendix 2 for a survey of the material to be processed) (2.5 fte)
	(Re)processing of metadata (see Appendix 2 for a survey) (2.5 fte)
	Linking to biographical data (see Appendix 2 for a survey) (1 fte)
postdoc + software developer	Curation of Old Dutch Corpus, Corpus of CD-ROM Middle Dutch (1.0 + 0.5 fte)
	Reprocessing of lexicons: integration of historical lexicographical sources into unified diachronic lexicon and integration with modern lexicon content (3.0 + 1.5 fte)

* The integration of existing tools into the infrastructure is brought forward as much as possible, so as to avoid developing tools first (or converting and/or tuning existing tools for work with historical corpora) and integrating them afterwards, at the risk of discovering serious incompatibilities only later. Furthermore it is important to promote its scientific use (Track 1) from the earliest possible date, and to expose the system to the problems of real-world data and users. Experience has shown that genuine use often uncovers the need for special conversions or the incorporation of further (general) analysis facilities, such as facilities for particular statistical analyses, visualizations, or maps. In general we do not foresee developing any of these within Nederlab, but rather plan to import them as needed from existing open sources (e.g., from the R statistics package). Moreover, we have concentrated the work needed to support literary and historical analysis (text cleaning, normalization, vocabulary analyses, concordances, annotations of various kinds and accessibility, immediately available lemmatization and part-of-speech tagging) in the first two years, scheduling work involving more sophisticated tools (more advanced taggers, lemmatizers, named entity recognition, sentiment analysis, geo-referencing, parsing) after the first year. Thus, we hope to involve more historians and literary scholars in the effort early on, thereby creating greater interest in the more sophisticated tools.

° In order to create a unifying layer of essential metadata for the diachronic corpus, a management environment will be set up within the existing operational environment of DBNL, in which the metadata supplied will be processed and the entered digitized data will be assessed for quality by way of documentation or personal observation. With the help of newly-created ‘xml containers’, the metadata will be linked to sources in distributed storage; digital curation of the basic metadata will be done in the DBNL environment. The data will be synchronized with Nederlab at times to be determined.

While building the corpus, efforts will be made on the one hand to integrate as many bibliographical and biographical metadata as possible, gleaned from, *inter alia*, STCN, CBK, the National Library newspaper project, DBNL, INL. Files will be found, searched and enriched by way of the xml containers; furthermore various data files crucial to research will be reprocessed to make them meet the minimum requirements for reliable scientific research.

Exploitation and other costs

The costs involved in keeping up the infrastructure after the project period will be covered by the three founding institutions, The Meertens Institute, Huygens ING, and DBNL/Nederlandse Taalunie, as explained under the heading ‘Local context’ above. These institutions have agreed to provide for server room to support Nederlab’s upkeep after the grant period has expired.

Duration of the project

Planned starting date	01-01-2013
Expected completion date	31-12-2017

The facility is expected to last for several decades, since it is built on corpora and (technical) knowledge made available by robust institutions such as university libraries and (KNAW) research institutes. The virtual research environment of Nederlab will benefit, both in the realisation phase and in the exploitation phase, from additions made to the corpora by the various data suppliers, since the diachronic corpus is cast in a dynamic set-up. Furthermore, Nederlab will make use of technological innovations done during the realisation phase. The KNAW institutes, with their advanced technological departments, will ensure that tools within Nederlab are kept up-to-date.

Management plan

The table below gives a survey of the significant milestones and deliverables during the construction phase. After this phase, the exploitation will be guaranteed by the founding institutions, see above.

Nederlab: Milestones and deliverables

Track		Description of deliverables
Track 0		
	a	Dissemination programme
	b	Master classes, workshops, summer schools
Track 1		
	a	Project calls and evaluations
	b	Edited volume
	c	International workshop
	d	Tutorials and further dissemination of project results
Track 2a		
	a	General User Interface (first year: Demonstrator)
	b	Indexing module
	c	Workspace module
	d	User administration and authorization module
	e	Authorization module
	f	Search/Browse module, i.a. metadata harvester
	g	Metadata editors
	h	Metadata publication module

Track 2b

a (first year) Nederlab Tools 0.1, implementing text cleaning and normalization (e.g. TICCLops):
set up as a web service (e.g. through CLAM);
can be applied to all corpus data compliant with or converted to data

b Nederlab Tools 0.2, with diachronic text cleaning & normalization

c Technical reports describing text cleaning, normalization

d Nederlab Tools 0.3, with diachronic lemmatization, generic and domain/period-specific lemmatization, morphological analysis (e.g. MBMA), and PoS tagging (e.g. Adelheid)

e Technical reports describing lemmatization, morphological analysis, and PoS tagging

f Nederlab Tools 0.4, with diachronic syntactic analysis (e.g. INPOLDER or ALPINO) and integration of other NLP tools: co-reference, NER, geo-referencing, sentiment, shared task outcome

g Technical report describing syntactic analysis

h Technical report describing integration of other tools

i Software release: Nederlab Tools 1.0, including all modules that pass quality threshold

j Release of overall documentation on Nederlab Tools: Reference guide, overall empirical performance estimates

k Two ‘shared task’ sessions at CLIN or at international meetings

Track 3

a (first year) Integration of corpus of high quality texts and metadata within Nederlab infrastructure

b Reprocessing corpora of lower quality (made available by scientific libraries and institutions) to meet required standards, and mutually linking them

c Reprocessing metadata of lower quality to meet required standards, and mutually linking them

d Linking corpora to biographical data

e Curated Old Duch Corpus, Corpus of CD-ROM Middle Dutch

f Integrated computational historical lexicon from the 6th-20th centuries

Risk and contingency planning

The implementation of Nederlab by such a large number of partners is an ambitious enterprise. Three main parties will be jointly responsible to ensure that the activities will go forward as planned even in the event of one or other party defaulting.

There is heterogeneity in the data: a large slice of the corpus is not of the presupposed good quality, since it consists of poor OCR. In compensation, a number of representative high-quality corpora will directly be incorporated as core corpora in the demonstrator. Data curation will gradually improve the parts of the corpus of poorer quality, a necessary measure as poor quality data is difficult to access with tools. The problem can be mitigated by designing the tools in such a way that they bypass the problems of poor quality data, and by enabling researchers to compose subsets of the corpus from which such data are barred.

The speed of Nederlab’s technical implementation is determined by the combination of the forthcoming interface and the corpus. To ensure adequate progress a modular setup has been chosen: after one year a demonstrator with a core corpus of corrected texts will become available. The demonstrator will be tested at once by researchers (pilots will be floated). New components with which researchers can experiment will gradually be added to the demonstrator.

It appears that some of the researchers cannot make the most of the technical possibilities. A special instruction programme will be set up in Nederlab to bring the technology closer to the researchers.

Additional information

You can provide additional information in this section. Please keep this section as concise as possible (max 1 page A4). Institutional support letters can be added after the signature section, but must be in PDF format and must be attached to the application. Please restrict length and number of such letters to a minimum.

To this application are added three Appendices and two support letters (as pdf):
- Appendix 1. List of researchers consulted with regard to the set-up of Nederlab
- Appendix 2. Survey of the data and metadata to be processed
- Appendix 3. Survey of infrastructure and tools programmes mentioned in this application, with short descriptions
- Support letter by dr. Theo Mulder, director KNAW
- Support letter by the constituent parties: prof. dr. H. Bennis, director of The Meertens Institute; drs. L. van den Bosch, secretary of the Nederlandse Taalunie; C.A. Klapwijk, director of DBNL; dr. H. Wals, director of Huygens ING

Declaration and signature

Have you requested funding for this research elsewhere?

☒ No

□ Yes,

Please include details of any additional grants you have requested for this research project

Declaration

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

prof.dr. H.J. Bennis, The Meertens Institute, Amsterdam