KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

KNAW-Agenda Grootschalige Onderzoeksfaciliteiten

Format nadere uitwerking van een ingezonden voorstel

I. VOORSTEL ALGEMEEN

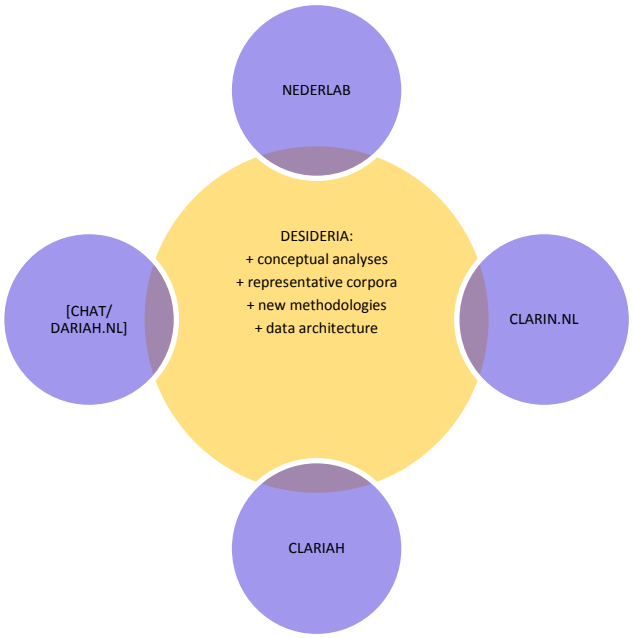
Acroniem	DESIDERIA
Naam van de infrastructuur	Dutch Extensible Semantic Infrastructure for Digital Explorative Reading and Information Architecture
Hoofdindiener Naam contactpersoon	Prof dr Els Stronks
Organisatie	Universiteit Utrecht
Functie	hoogleraar Vroegmoderne Nederlandse letterkunde
Adres	Trans 10
Telefoon	030-2538249
Email	e.stronks@uu.nl
Mede indiener(s)	
Naam contactpersoon/ mede-indiener	Prof dr Lex Heerma van Voss
Organisatie	Huygens ING
Functie	Directeur Huygens ING, PI CLARIAH
Adres	Postbus 90754 2509 LT Den Haag
Telefoon	070-331 5800
Email	lex.heermavanvoss@huygens.knaw.nl
Overige mede indieners en uitvoerders subprojecten	Prof dr Wim van Anrooij (Universiteit Leiden) Prof dr Hans Bennis (Meertens Instituut) Dr Bart Besamusca (Universiteit Utrecht) Prof dr Rens Bod (Universiteit van Amsterdam) Ir Daan Broeder (Meertens Instituut) Dr André Bouwman (Universiteit Leiden) Dr Antske Fokkens (Vrije Universiteit) Prof dr Inger Leemans (Vrije Universiteit) Matt McBurton, PhD (University of Pittsburgh) Prof dr Maarten Mous (Universiteit Leiden) Prof dr Nicoline van der Sijs (RU/Meertens Instituut) Prof dr Ingrid Tieken-Boon van Ostade (Universiteit Leiden) Dr Henk Wals (Internationaal Instituut voor Sociale Geschiedenis) Drs Joris van Zundert (Huygens ING) The consortium also consulted: prof dr J. van Eijnatten (UU/Translantis), prof dr Franciska de Jong (EUR/UU/CLARIN ERIC), prof dr Tom Koole (Universiteit Groningen), dr Marco Streefkerk (DEN), prof dr Jan Odijk (UU/CLARIAH), prof dr Piek Vossen (VU), prof dr Sally Wyatt (Universiteit Maastricht/KNAW).

Samenvatting
Geeft korte samenvatting van deze faciliteit in termen van werking, wetenschappelijke voordelen etc. (max 350 woorden).
DESIDERIA aims to offer a program that brings a new level of textual analysis to digital scholarship. It consists of three subprojects targeted at the improvement of data and tools, and the development of a related architecture. As a program, it will enable diachronic semantic analysis of cultural and political concepts like democracy and social inequality to study changing views on these notions from synchronic and diachronic perspectives. This type of semantic research is of crucial value for the humanities and will be valuable for other sciences. The three subprojects of the DESIDERIA program will tackle three current problems in semantic analyses: 1. non-representative textual data sets and lack of techniques to draw conclusions from partial digital sources; 2. insufficient quality of existing digital tools; 3. Insufficient infrastructures. These shortcomings are limiting semantic conceptual analysis of texts in all languages, including Dutch. To solve the identified problems, DESIDERIA will build on a number of successful projects developed in the Netherlands (particularly CLARIN, NEDERLAB, CLARIAH). From an international perspective, these projects provide access to unique corpora of digitized texts from all phases in the history of Dutch (as a case study for other languages), and they enable simple analyses of syntactic text structures (a necessary pre-condition for the analysis of semantic underlying conceptual structures). With the existing digital Dutch scholarly infrastructure the basis for the next step into deeper semantic analyses envisioned in DESIDERIA is in place. The three subprojects will help to accomplish DESIDERIA's main goal of supporting diachronic semantic the next level of textual analysis. Subproject 1 describes how with strategic data production and new methodological techniques suitably annotated digital text corpora can be developed. Subproject 2 offers the design of digital tools that will enable digital analysis of cultural concepts across time using linguistic features. Subproject 3 outlines an architecture that ensures long-term access and functioning of tools and textual data for digital scholarship. DESIDERIA's intended infrastructure is flexible. It is designed in such a way that it can be expanded with huge volumes of digitized text (in Dutch, and other languages) and newly developed tools. A user-centric, iterative, agile development is foreseen. Once DESIDERIA is active, scholars from a wide variety of disciplines will profit from access to an infrastructure that produces more reliable interpretations of textual data.

Kernwoorden
Geef maximaal 8 kernwoorden die de faciliteit typeren.
conceptual textual analysis; strategic digital data production; semantic tools; data science for the humanities

VOORSTEL INHOUDELIJKE UITWERKING

A. SCIENCE AND TECHNICAL CASE
Volledig nieuwe faciliteit of verbetering van reeds bestaande
Beschrijf in hoeverre het hier een geheel nieuw idee betreft of een verbetering of opvolging van een reeds bestaande faciliteit.



* DESIDERIA will expand the work done in the context of the Investment Subsidy NWO Large NEDERLAB (*Laboratory for research on the patterns of change in the Dutch language and culture,* http://www.nederlab.nl). The aim of NEDERLAB 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 - fully CLARIN-compatible - open access web interface, allowing scholars to simultaneously search and analyse data from texts spanning the full recorded history of the Netherlands. The technical ambitions of DESIDERIA surpass those of NEDERLAB in that the tools developed in the context of DESIDERIA are designed to have more semantical analytical power. Nederlab's analytical tools will be a stepping stone for the type of conceptual analyses envisioned in DESIDERIA. Also, the data sets used in research projects based on DESIDERIA will cover more data currently only available material collections (printed and handwritten collections that have not yet been digitized). NEDERLAB is not developed with the intention to actively resolve *lacunae* in research collections. In NEDERLAB, researchers only work with the textual data that happen to be available. These are the result of different, and in themselves rational, decisions to digitize important texts, like newspapers, or parliamentary debates. Conclusions based on these data sets would be more adequately supported if the available digital collections were selected and produced with the intention of representing the available genres and types of sources more strategically.

* DESIDERIA also builds on the Investment Subsidy NWO Roadmap for Large Scale Research Infrastructures for 2015-2018 CLARIAH (*Common Lab Research Infrastructure for the Arts and Humanities,* http://www.clariah.nl/; in part the successor to CLARIN-NL and DARIAH-NL) but has a wider, more longitudinal scope (for it has the ambition to include full text disclosure of also the earliest phases of Dutch).

Historical research supported by CLARIAH, is focused on social economic history, which makes intensive use of structured databases rather than textual resources. To these databases, DESIDERIA adds the dimension of research into textual evidence, with a special focus on conceptual shifts in texts, which requires semantic analysis tools. CLARIAH's WP 3 (Linguistics) and WP5 (Media studies) also works with textual data, but do not have the specific objective to provide tools for conceptual analyses.

* DESIDERIA is proposed with an eye currently developed in the context of CHAT.NL (Center for Humanities and Technology). CHAT is technically speaking not an infrastructure (with data, tools etc.), but a network of Digital Humanities Scholars, structured around a research programme and the potential cooperation with business partners interested in that programme. DESIDERIA differs from CHAT in its focus on strategic data production, but shares a common interest in textual research.

Science Case

Geef een algemene introductie van de wetenschappelijke waarde van de faciliteit.

The aim of DESIDERIA is to develop an environment that provides 'deep' perspectives on human life on the basis of conceptual analysis of longitudinal and synchronic textual corpora. Over centuries humanities researchers have built up an impressive expertise in the analysis and interpretation of everything humans make, do, think, or give meaning to, based on textual analyses. Digital resources (digitized texts, digital tools) have added a radical new dimension to this age-old expertise: scale. Over the last decades, humanities researchers, in collaboration with computer scientists, have developed various innovative strategies for digital analysis of textual data. This textual research is a welcome addition to existing longitudinal research based on numeric data, surveys, experiments – for the latter often only sheds light on developments of the last decades or years. The relative "short-termism" this research suffers from (as argued by historians Armitage and Guldi in their *History Manifesto* (Cambridge University Press, 2014)) can be complemented by longitudinal research based on digitized text corpora.

For not only does text contain massive amounts of factual statements but, more importantly, it also reflects our perspective on these statements: our emotions, opinions and views. Textual data is therefore not only 'big' but it is also 'deep', shedding light on conceptual stabilities and shifts embodied in the statements and perspective on statements in language by various groups in society and in various period. Historians and literary scholars have a longstanding tradition in working with this type of research, starting with the paradigmatic work of Arthur Oncken Lovejoy (see for instance *The Great Chain of Being: A Study of the History of an Idea* (Harvard University Press, 1936)). Conceptual history has since developed analytical approaches to unravel shifting value systems and ideologies embodied in the semantics of language (see for instance the work of Otto Brunner and Reinhart Koselleck; e.g. Reinhart Koselleck, *Begriffsgeschichten* (Suhrkamp, Frankfurt am Main, 2006)). However, these approaches typically are based upon a time intensive, traditional Humanities approach. With the methodology proposed here, this type of analysis can deal with more concepts of different types and the scale of texts covered can be increased immensely. Most importantly, the analysis becomes repeatable and reproducible, which increases scientific rigour.

This type of historical, literary research does not stand on its own, but has sought an alliance with conceptual research done in other disciplines – the social sciences as well as life sciences and legal studies. Social scientists interested in analyzing how (polarized) ideologies are expressed and produced by discourse, benefit from deeper semantic analyses. Interdisciplinary longitudinal research helps to unravel the development of concepts like 'compassion – empathy.' According to biologist Frans de Waal, empathy is an universal and innate primate function (*The Age of Empathy: Nature's Lessons for a Kinder Society* (Harmony Books, 2009)). Conceptual historians contest this vision by stating that compassion and empathy (in their view two different concepts) are not only trainable but also fluid. These concepts are

debated, construed, and taught in textual discourses. Longitudinal analysis of textual cultures therefore allow for the tracing of concept change: in Catholic societies for instance, the object of compassion often was Christ. Protestant societies started to doubt the possibilities of feeling the feelings of Jesus (the son of God). Consequently, the object of their compassion shifted towards one's suffering neighbour: the imprisoned, slaves or other socially challenged groups. These shifts can be traced in religious texts: in Catholic cultures these texts formed discourses channelled by priests, in Protestant cultures new categories of authors arose (citizens, politicians etc.) resulting in new forms of channelling. This in turn had long lasting consequences for the ideas associated with, and the organisation of poor relief in Catholic and Protestant countries. The shift is thus likely to be indicated not only by changes in frequently used words (from Jesus to neighbour, for instance), but also by changes in perspectives.

Current digital research into concepts mainly emerges from computational linguists rather than (literary) historians and other experts in conceptual research – see for instance Dirk Geeraerts (KU Leuven), Susan Fitzmaurice and Jane Hodson (University of Sheffield) and, in the Netherlands: Piek Vossen (VU) and Antal van den Bosch (RU). Computational linguists predominantly use distributional semantics to determine changes in word meaning. Distributional semantics represents the meaning of words based on their distribution in text. This form of analysis reveals which words were close in meaning at one time but have subsequently become more distant at another (and vice versa). It is successful in modelling, for instance, changes in meaning of the English world *cool*. In the literal sense, *cool* was first applicable to for instance products (as in: 'cool beer'). However, as early as the sixteenth century, *cool* changed in meaning when it was also applied in the metaphorical sense. And within this metaphorical application, new shifts occurred over the centuries and even within decades. Whereas, for instance, it mostly referred to behaviour ('a cool dude') in the 1990s, it is now also often applied to products and gadgets ('a cool App'). Such changes can be detected with much more certainty when not simply words, but concepts in all of their semantic complexity–the expertise of conceptual historians and other experts in conceptual studies– are taken as point of departure. DESIDERIA aims to explore the type of crossovers between linguistic and historical disciplines in the humanities to obtain better understanding of the way in which language and discourses feed into conceptual shifts and in doing so, bridge the gap between the humanities and other sciences interested in conceptual research.

Beschrijf de wetenschappelijke voordelen en verwachte doorbraken.

DESIDERIA is set up to further develop a digital humanities infrastructure towards deep conceptual analysis. It strengthens humanities research by uniting the expertise and capabilities currently scattered across a number of research and heritage institutes to facilitate other scientific advancements through the accumulating expertise of (digital) humanities. It consists of three sub-projects targeted at the improvement of data, modeling tools and methodologies, and architecture. In addition to the benefits expected from more advanced conceptual analyses, each of the three sub-projects have their own expected benefits beyond DESIDERIA's overarching goals:

Subproject 1 - Data
The digitization of textual data comes with challenges: current projects address issues related to Optical Character Recognition and variation in spelling. But larger issues around completeness and representation overshadow such efforts and are addressed in subproject 1: a national strategy for the digitization of presently still underrepresented texts, and new methodologies to deal with partial and incomplete textual data sets in quantitative research.

Subproject 2 - Modelling and evaluation
At present, analytic tools are more effective at syntactic rather than semantic analyses. Subproject 2 aims to achieve a breakthrough in the current situation by offering a design for semantic search infrastructure

that build upon the currently available techniques for syntactic annotations (such as Part of Speech-tagging). This subproject will develop a completely novel infrastructure to help humanities researchers obtain better insight into the correctness of their textual analyses (precision and recall) and the potential to address specific research questions (usefulness and reliability) by developing gold standard evaluation data.

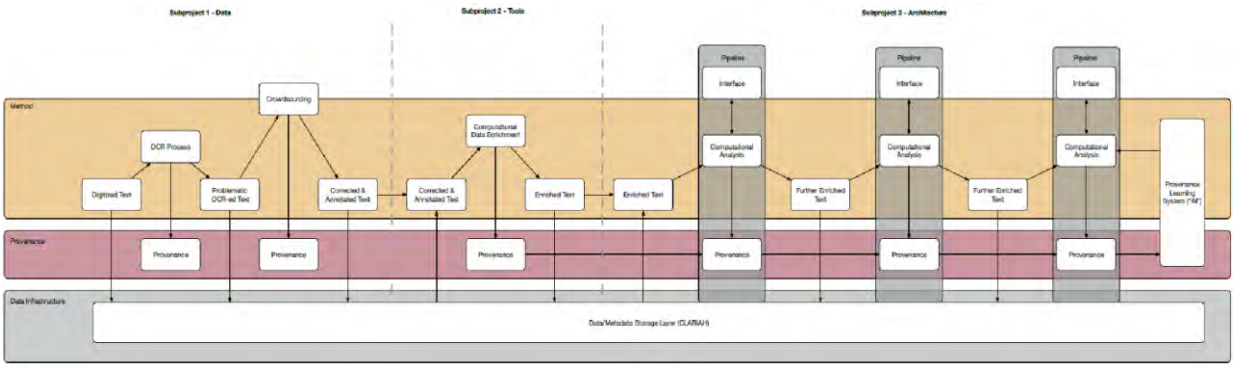
Subproject 3 - Architecture
Current textual digital infrastructures are not designed to work and rework the outcomes of textual analyses. Subproject 3 departs from the idea that digital textual scholarship not only consumes, but also produces data (as well as analyses and algorithms). If data is produced in a re-usable form, it accelerates textual digital scholarship. This subproject will develop a framework for expressing and composing textual analysis pipelines, which can be used, reused, or remixed.

Beschrijf hoe deze faciliteit zich verhoudt tot alternatieve faciliteiten/onderzoeksmethoden.

DESIDERIA introduces three innovative strands in current digital textual scholarship. Each strand is represented, in different shape and form, in each of DESIDERIA's subprojects. The three strands are:

- under-studied methodological issues
- issues of provenance
- data-infrastructural issues

This diagram shows how the three strands are interwoven in the fabric of DESIDERIA:



1. *under-studied methodological issues*: existing efforts focus on practical rather than methodological issues, and leave aside questions such as "how much should we digitize of each text genre available in order to work towards a more representative data set?" or "how appropriate is this resource for this type of research?". Yet, answering these questions is a necessary requirement for the maturation of quantitative textual research. Work in this strand will explore views on the reliability of analytic results and the degree of (un)certainty in the conclusions of quantitative textual research.

2. *issues of provenance*: textual datasets are multi-layered; containing data as well as metadata (for example, the data on how a text was mediated). In the past decade, international standards for the disclosure of textual data and metadata have come into being (such as Dublin Core), but advanced digitization and analysis introduces ever more complex issues of provenance. When new textual datasets are generated through innovative analyses, information about the original data and how the newly generated data and annotation layers were provided should be recorded.

3. *data infrastructural issues*: at present, the sharing and compiling of large textual datasets is hampered by the lack of a flexible technical structure that offers mechanisms for the expression and composition of advanced analytics addressing the *under-studied methodological issues* and promoting the evolution of an open system or ecology of data and tools. Furthermore, infrastructure should support the easy and automatic generation of provenance records to address the *issues of provenance.*

Technical case

Geef op hoofdlijnen een technische beschrijving van de faciliteit. Hoe zit de faciliteit in elkaar en hoe werkt het?

The technical infrastructure we envision enables representative datasets and techniques for working with partial data sets (subproject 1), high-quality semantic tools (subproject 2), and a new architecture for textual research (subproject 3).

Subproject 1: Strategic Data Production and New Methodological Techniques

Current digitized text collections are biased and of poor quality, and therefore do not provide the metadata and coverage needed to acceptably select representative samples. Yet, even with the strategic data production of underrepresented textual resources, coverage will always be partial and incomplete in nature (for instance: some concepts made it into written texts less often than others; some texts are lost over time, especially so in earlier periods). Analytical tools can never produce objective results; they are always subject to interpretation. Textual scholars face challenges in learning how to draw conclusions from partial sources, statistical inference, etc. We describe the activities needed to overcome both existing biases in digital collections and the methodological issues relating to data partiality and quantitative analytical tools.

Strategic Data Production

We have conducted a number of surveys to locate obstructions in data and metadata sets (§ 1.1), related these obstructions to future digitization plans of the national institutes (§ 1.2), and outline the requirements of a strategic digitization plan and crowdsourcing platform that discloses underrepresented texts (§ 1.3). *N.B: in the context of subproject 1, 'digitization' is used as a term to describe full text + high quality disclosure of scanned textual sources (i.e. transcriptions, so machine readable texts are the result) .*

1.1. Lacunae in Digitized Texts Designated by Researchers

A survey, distributed on public forums such as Neder-L and through email to humanities scholars, inventoried the needs and wants of researchers from various disciplines in the humanities (see Appendix 1). What priorities in digitization (production of both textual data and metadata) do they foresee in their current and future research, especially requirements of representativeness? The survey was completed by 59 researchers (from various disciplines in the humanities and historical periods; two-third of them are working in the Netherlands, with Dutch handwritten or printed texts).

The survey revealed a large demand for diplomatic or critical transcriptions of texts or spoken language corpora. Transcriptions produced through OCR-techniques (that typically do not result in flawless transcriptions) were found insufficient for 44 out of the 59 researchers. The other 15 researchers (specialised in texts from the 20th and 21st century), were satisfied with the quality of OCR-results. Nearly 60% of all researchers maintain that having a small set of metadata (consisting of: author, title, year and place of publication) is sufficient, but a third of researchers indicated more detailed metadata of various nature (sociolinguistic, genre classification, musicological information, etc.) is required.

Priorities vary, but most researchers do point out the need for balanced data sets (in genres, historical periods) with texts enriched with linguistic annotations, for reliable attributions (in time, in place), for reliable transcriptions, and for interfaces that allow for the selection of sub-corpora exportable to specific tools.

The survey produced a more detailed priority list:

- medieval and early modern manuscripts,
- private documents (e.g. letters and diaries),
- periodicals,
- corpora with texts in various dialects,
- data sets with more variation in genres,
- literary newspapers of the 20th century,
- printed ordonnances,
- lexical data sets,
- archives,
- non-Dutch text corpora,
- spoken language corpora,
- web texts and texts produced on social media,
- video data.

1.1.2 Institutional Digitization as a Check of the Researchers' Priority Lists

This priority list has been checked with current plans of the Royal Library and Metamorfoze[1]. It is the Royal Library's ambition to digitize its own collection of printed books, periodicals, newspapers, (medieval) manuscripts by 2030. The Royal Library houses the Metamorfoze project, which aims at digitizing periodicals and archives of interest to Dutch History.[2] In addition, 150,000 extra titles (owned by partners of the Royal Library) are planned to be digitized before 2018, and the Royal Library is also coordinating a national program to digitize newspapers. The scientific committee of Metamorfoze is in charge of the selection of these newspapers. In cooperation with ProQuest, another 18,206 volumes of manuscripts made before 1700 have been digitized, but are only available through the website of ProQuest (http://eeb.chadwyck.co.uk/geoLocSubscription.do). The 78,995 manuscripts dating from 1700 and after, also in the Royal Library's own collection, have been incorporated in current plans, but with a lower priority.

At present, these titles have been digitized and published in the context of the Royal Library's repository, Delpher (http://www.delpher.nl/):

Period	Genre	Number	Partners/project
16th-20th centuries	books	52,000	KB and other national heritage institutions
16th-19th centuries	books	170,000	Google Books project KB and UBA
1618-1995	newspapers	8,000,000 pages	KB
19th-20th centuries	periodicals	1,500,000 pages	KB
1937-1989	news bulletins	1,800,000 pages	ANP

At present, it is not possible to chart the overlap between the researchers' priorities and the plans of the

[1] In a meeting with Steven Claeyssens, Jasper Faase, Marg van der Burgh.

[2] This is done in the context of the Metamorfoze project which will scan archives from various national institutes (mostly Dutch archives, but for instance also ca. 160.000 scans of Dutch letters kept in the (British) National Archive in Kew). No metadata or transcriptions will be provided.

Royal Library and Metamorfoze at a detailed level. It is clear, however, that the digitization efforts of the Royal Library and other national heritage institutions so far have not yet resulted in a representative set of digitized texts, since the digitization of texts in the Metamorfoze programme focuses mainly on printed materials from the modern period (especially 1840-1950) that are threatened by physical deterioration. The digitization of medieval manuscript books and archival documents is mostly left to the initiative of individual heritage institutions. These materials are, moreover, unique and material relevant to Dutch language and history are kept worldwide, not just in the Netherlands and Belgium.

Furthermore, it is clear the OCR-quality of the digitized printed materials is insufficient for researchers working with texts from earlier time periods. Quality enhancement through crowdsourcing will be necessary: crowdsourcing has been shown to be an effective remedy for correcting and enriching poorly OCR'd texts. The Royal Library has expressed the intention to develop a crowdsourcing infrastructure with the KNAW as a partner (building on existing cooperation between the Royal Library and Meertens Instituut/NIOD). Digitized texts from the 20th and 21st century do not have such OCR quality issues, but do have copyright issues (see § 1.5).

1.3 Requirements for strategic digitization

To provide the kind of balanced, reliable, and representative data sets researchers need, one needs to know what was published and has been preserved and what has been lost before one can select texts to digitize from what is still available and preserved. The following steps need to be taken:

1.3.1 Compile a reliable census of all texts ever printed in the Netherlands and select strategically important materials:

- link or harvest relevant parts of existing catalogues (such as NCC, STCN, Nederlandse Bibliografie in Picarta, Nederlab catalogue), with the purpose of composing one corpus without duplicates
- compare this corpus with printed stock lists (such as Saalmink, Brinkman e.d.) and compose an extended corpus or census on the basis of the comparison; compare and extend this census further with stock lists of publishing companies, archival information.
- make this census public and have a group of researchers compile a representative selection

1.3.2 Start a project to enlist Dutch manuscript books dating from before 1501, both within and outside the Netherlands and select strategically important materials:

- manuscript books dating from ca. 1200-1500 (all surviving fragments included) are especially lacking from existing digitized text collections. They have either not been included, or are only available as digitized version of later printed editions of these publications. Prioritizing the digitization of manuscripts books is complicated, as manuscript books often contain more than one texts, and the same text may exist in various versions. These aspects of these source materials should also be incorporated in prioritizing these resources.
- complete and update the existing census of Middle Dutch manuscripts (Bibliotheca Neerlandica Manuscripta)
- have a group of researchers compile a representative selection

1.3.3 Start a project to select archival documents and private documents both within and outside the Netherlands:

- a study by DEN Enumerate has resulted in the estimation that ca. 10% of all Dutch archival materials have now been scanned. The rest should be prioritized.

- have a group of researchers compile a representative selection, based on existing overviews of archives (such as: Archiefnet, Archieven.nl, Archive Grid) and private documents (such as Catalogus Epistularum Neerlandicarum, R. Dekker e.a., *Egodocumenten van Noord-Nederlanders van de zestiende tot begin negentiende eeuw. Een chronologische lijst*. Rotterdam 1993; R. Lindeman e.a., *Reisverslagen van Noord-Nederlanders van de zestiende tot begin negentiende eeuw: een chronologische lijst*. Haarlem 1994).

Use these selections and available digitization budgets as a means for a strategic data production with all (inter)national partners. When digitization requests by individual researchers may be funded from available budgets, relate them to this selection, with the purpose of prioritizing the digitization of underexplored texts.

1.3.4 Coordinate efforts to digitize audio-visual data sets; efforts made by for instance the Investment Subsidy NWO Large LISTEN and the group that works on a plan for the KNAW- toekomstagenda (ADvanced Video ANalysis Tool (ADVANT).

1.3.5 Regulate the archiving of digital born-texts and scans with partners such as the Royal Library and National Archives (NIOD, IISG, NA, Netwerk Digitaal Erfgoed Decentraal). At present, no standards or guidelines for data formats, metadata are in place. A promising step was taking by installing the Nationale strategie digitaal erfgoed, see:. http://www.den.nl/pagina/511/netwerk-digitaal-erfgoed/. The linking of datasets and research into these datasets was seriously hampered by the lack of coordination, and efforts like these ensure a better future.

1.4 Requirements for Strategic Data Production: a Crowdsourcing Platform

Most quality issues with existing text sets relate to poor OCR-results. Crowdsourcing is often referred to as a means to resolve issues of biases in text collections. In the second survey we conducted, the project leaders of a number of large crowdsourcing projects have been consulted to see how useful crowdsourcing could be (see also Appendix 2) to resolve issues of bias. Furthermore, volunteers working in these projects have been surveyed to chart their needs and desires (see also Appendix 3). The results of these surveys have been used to describe how to enroll and engage the crowd in strategic data production (§ 1.4.1), and to describe the architecture of the crowdsourcing infrastructure needed to facilitate this crowd (§ 1.4.2, see also Appendices 4 and 5).

1.4.1 Crowdsourcing as a means to produce full text corpora?

Most KNAW institutes have experimented with crowdsourcing as a means to enhance the quality of digitized data, and found it a method with potential to be developed further. Crowdsourcing is seen as the most suitable means to enhance the quality of digitized data.

Project leaders' experiences

Among researchers, experience with crowdsourcing projects is rather scarce (13% has some experience, see Appendix 1, question 20 and 21), but the results are promising. Seven leaders of crowdsourcing projects (see Appendix 2) were asked for their experience. The projects primarily revolve around transcribing and the adding of metadata. The project leaders devoted one day per week to the supervision of volunteers, for whom they developed detailed instructions. The quality of the work of the volunteers was checked by the project leaders as well as by some specially trained volunteers. The project leaders used existing software/tools, but

experienced problems in applying them to their own projects. It is clear that volunteers need to be stimulated through the organization of social events.[3] A variety of rewards systems were used, however none of the projects used gamification elements.

All project leaders were pleased with the results, and reported low costs (budget of a few thousand euro are the standard, the largest project reserved € 50,000 for crowdsourcing as well as scanning, and data management). Within this project, € 16,000 was paid for the use of the website VeleHanden (Picturae). This project reports positive results in the application of crowdsourcing, and expects this means to be useful in other, even more large-scale projects of data production and curation, incorporating the lessons learned from this experience (provide ample supervision, trust your crowd, make it an enjoyable enterprise for volunteers).

Volunteers were mostly recruited through the personal networks of the project leaders. In the beginning of each project the number of drop-outs was high, but the volunteers that stayed were very productive. About 90% of the work is done by 10% of the volunteers. The project leaders value any contribution and see no reason to have volunteers undergo selection processes.

Crowds' experiences
In total 247 volunteers active in 5 projects organized by the Stichting Vrijwilligersnetwerk Nederlandse Taal and the Meertens Instituut have been surveyed to find out what the experiences of the crowd have been thus far (see Appendix 3).

The crowd's profile:

- average age: 63.4 (youngest volunteer=23, the oldest=99)
- 51.6% female, 48.4% male
- 80.5% has a higher education (hbo- or academic), their professional backgrounds vary
- over 37% worked on more than one project
- 66.5% of the volunteers has done other voluntary tasks of great variety
- most volunteers enlisted after reading newsletters or other media exposure
- they joined the project to: support science; gain new experiences or knowledge. Social contacts, career and appreciation are less of a factor
- 50,8% is still working on a project (some of them since 2007), 38.7% has stopped, the rest has paused (due to a lack of time, personal circumstances, lack of contact with the project, technical problems)
- most volunteers are positive about the project organisation, and not interested in having more contact with other volunteers
- contact with the project organisation takes place through email (68%), Yahoo-groups (7%) or Forum (17%)
- most volunteers think they have adequate skills to perform well, and prefer more difficult (not boring) tasks
- 56% of the volunteers would not appreciate game elements in crowdsourcing ("childish, feels like we were not taken serious"), almost 30% has no opinion on this matter and only 6.5% likes the idea. Competitive elements are not appreciated, but meetings and presents are.

[3] This conclusion is supported by the outcomes of the PhD project of Montserrat Prats Lopez, *Citizen Science: A multiple case study* (VU, to be defended in 2016). The results of an interview we had with her, are incorporated in appendixes 4 and 5.

How to use crowdsourcing to handle problematic data sets?
The volunteers could be entrusted with two types of tasks:

1) open up texts, images, spoken language by means of transcriptions and metadata sets. These volunteers could produce transcriptions, correct existing OCR results.
    - this requires a crowd of which the project organisation has detailed information on the individuals, such as education level and experience. For this information a volunteers database has to be kept. Part of this crowd can be trained for special jobs.

2) gather new data or annotations. These volunteers can add metadata, lemmas, Parts-of-Speech-tagging, syntactic parsing and Named Entity Recognition. For this, a tool with game elements has successfully been developed and used in a Groningen project used to transcribe English texts.
    - this requires a large, anonymous crowd with no specialized domain knowledge .

To enroll more volunteers, a survey should be widely circulated (based on the CLARIAH Online Survey tool) to encompass the potential and preferences of all volunteers, and to make a match between volunteers and specific projects.

1.4.2 Towards a Crowdsource Infrastructure
To enhance strategic data production, the KNAW should develop and orchestrate a professional crowdsourcing platform that specializes in high quality data for scientific projects (with higher standards of quality then existing crowdsourcing platforms (like VeleHanden)). This platform will require an investment, but would serve as a central (and thus cost efficient) clearing house for all volunteers, as well as students, and would showcase results. Technical solutions should be explored to develop this platform. Crowdsourcing should be complemented by technical solutions in order to facilitate, accelerate, and improve the data production. A promising model is provided by Transkribus (https://transkribus.eu/Transkribus/): automatic handwriting recognition is combined with manual corrections by volunteers. This must be further developed for Dutch handwritten and printed texts (also printed in the Gothic script), aided by Dutch lexica and applications such as TICCL and PICCL. See Appendix 5 for a more detailed design of this platform's architecture. In general, the platform consists of tools to:

- manage volunteers and students (enrolment, editing, communication);
- reports the volunteers' and students' progress and plan their work;
- manage tasks of volunteers and students (importing, editing, etc.);
- plug in addition software (for instance a module for semi-automatic correction of OCR);
- manage resulting data sets

All of these elements of the platform should be built on a detailed usability study and a thorough survey of existing software.

1.5 PR- and Privacy-issues
The development of IPR and privacy standard is essential to all of the DESIDERIA´s ambitions. At present, 40% of all researchers report problems in this area. Indirectly, they experience problems on an even larger scale for the large digitized collections of the Royal Library and other libraries are ruled by all kinds of IPR-restrictions. The Royal Library is making progress, a position paper is being prepared within NDE with participation from IISG, and should be supported by the KNAW and other national institutes. A special set of guidelines should be developed for the rights crowdsourcing volunteers have over the data sets they produce. Along with the development of a crowdsourcing platform, this should be arranged by the KNAW.

In the Dutch context, privacy rules are governed by the Wet Bescherming Persoonsgegevens. This law also affects the registration of volunteers involved in crowdsourcing projects, for their consent is needed when their names are published on a crowdsourcing platform or forum. Furthermore, volunteers should be asked to waive their rights on the transcriptions they make before the start of each project.[4] All of this issues should be dealt with in a protocol for crowdsource projects drawn up by the Royal Library and KNAW.

1.6. New Methodological Techniques

There are significant methodological issues that relate to partiality of textual collections and cannot solely be solved by strategic digitization: collections will always be partial in nature, and textual scholars therefore need to develop techniques to deal with these partialities.

Textual scholars also need to develop new routines ('pipelines') in the handling and enrichment of the data they produce (data consisting of digitized texts as well as analytical results). These issues are dealt with in our subproject 3, where we describe a pilot project centered around the development of new technologies for post-OCR-correction (for instance the integration of TICCL, Monk and/or Transkribus) in terms of a 'pipeline'.

Subproject 2: Enrichment, Modelling and Evaluation of Data

The overall goal of this subproject is to bring forward the possibilities of humanities studies by providing data, data enrichments, evaluation of data and enrichments alongside the means to produce new data with similar enrichments. We particularly focus on enrichments that aim at *deep semantic interpretations* of text, where we determine what perspectives are given on events, participants, values, ideas and ideologies that occur in texts. In order to achieve this, the project focuses on three main aspects: (1) data enrichment through textual analysis and linking, (2) data modelling of enrichments, and (3) evaluation of data and its enrichment. The deliverables are:

- a flexible, uniform representation of data enrichment and primary metadata of the main existing datasets in the Netherlands as well as the digitized datasets from subproject 1 (these representations will be linked to the original dataset).
- links between existing and new datasets.
- a variety of evaluation datasets that can be used to assess the qualities of identified links between sets as well as the quality of linguistic analyses. This will include both the correctness of the outcome of the analyses (precision and recall) as an evaluation for their potential (usefulness and reliability) to address specific research questions.
- an inventory of necessary new tools and required improvement of existing tools in order to apply deep semantic search to texts of various genres and from different times.
- tools and improvements of tools from the inventory. The choice of tools from the inventory that will be created or improved will depend on direct needs of humanities researchers and expected workload for creating the tool.
- a methodology of how new studies can be set up by applying linguistic analyses to textual data in the set (or to newly added textual data).

[4] This implies (in Dutch): "Iedereen heeft en houdt de vrije beschikking over het zelf ingetikte gedeelte. Medewerking houdt in dat het resultaat van het werk om niet en belangeloos beschikbaar wordt gesteld voor wetenschappelijk onderzoek, bij voorkeur op een algemeen toegankelijke website. De namen van de vrijwilligers/medewerkers worden daarbij expliciet vermeld, behalve wanneer een vrijwilliger dat niet wenst. "

We focus on textual data for now, but the overall methodology and ideas behind the approach are applicable to other types of data. This approach especially includes efforts to link textual data to other types of data such as images, statistics, film or audio data. Analysis of non-textual data lies out of scope of this project, but we will also include links to non-textual data based on metadata and textual descriptions, so as to explore possible connections for the future.

First, we introduce the overall data model and explain how this model allows us to add enrichments to data and provide new links without compromising the original data. The next section outlines the enrichments of datasets and provide examples of the technologies and focus on deep semantic enrichment. Finally, we address how these three aspects provide new forms of research relevant to several disciplines across the humanities and beyond.

2.1. Data Modelling

Good systematic data representation is essential when making data available for a large group of people over a long period of time. Researchers can lose valuable time or even miss out on information essential to their studies when they have no access to or cannot rely on information because it isn't represented in a systematic, well documented way. Researchers need information that is ordered to provide immediate and reliable evaluations of the results of the their research queries and data analysis.

In this project, we will be dealing with a wide variety of data that was mostly developed, digitized, documented and enhanced with metadata independently by different institutes and research groups. The different needs and backgrounds of these groups have resulted in a large variety in data structures with variation in data completeness and reliability. Overcoming these particularities by unifying the data structures is not possible nor is it advisable, but it is vital for researchers to be aware of these differences. Scholars must be able to trace the origin and specifics of the data they use at any time, so that they can assess the reliability of the data they are using. This is influenced by the quality of the original data and the method that was used to create this data, by conversions that may have led to loss of subtleties in the original dataset and, in case of automatically generated enrichments, by the reliability of the tools used to generate the data. This leads to the following requirements of data representation:

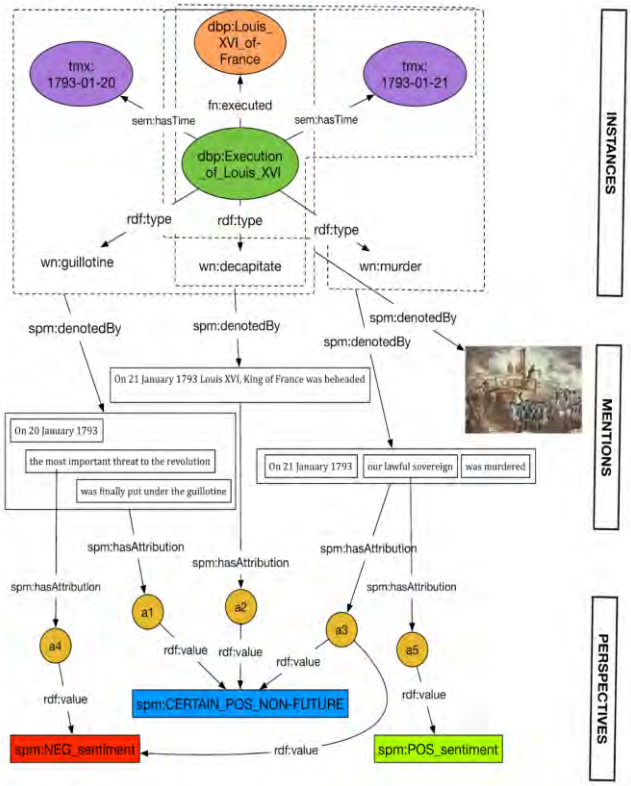
- data representations, conversions, and enhancements should never replace the original data, but be added as a separate dataset that is linked back to the original data. The user can always access the original data to verify information in derived sets.
- provenance information must be saved or added to any generated dataset. This should include information about the original data and how the newly generated data was provided as well as links to detailed documentation about the data and each step in the generation process.

The exact representation of specific information can only be determined by investigating the data included in the set. This aspect of data representation will be part of studies on establishing relatedness between sets. These details will be integrated in an overall approach for data representation through a generic *source perspective model* that is explained below.

The model consists of three layers. The first layer is an *instance* layer, which represents entities (e.g. a person or event). In this layer, various instances (including persons, organizations, events and concepts) can be represented. The second layer includes *mentions* of these entities. These mentions can be texts referring to the entity, but also pictures or other representations of the entity that are connected to relevant information. *Mentions* can be seen as references to a specific entity in a variety of sources. For each mention, we can specify information related to the perspective of the source. We can then model how certain the source is about a statement, which opinion the source expresses or what sentiment is carried by the choice of words. This information is provided in the third layer, the *perspective layer,* of our model.

The idea is illustrated in Figure 1, which represents the perspective of three sources on the execution of Louis XVI in a simplified manner. The following three expressions are modelled in this representation:

1. On 20 January 1793 the most important threat to the revolution was finally put under the guillotine.
2. On 21 January 1793 Louis XVI, King of France was beheaded.
3. On 21 January 1793 our lawful sovereign was murdered.



Alternative perspectives on the execution of Louis XVI

The top layer in Figure 1 provides the instance representation: the event itself, the main person involved and the time when the event took place. There are two different times associated with the event: January 20th is linked to the first mention, the relation between event and January 21st to the other two. The instance layer furthermore provides information about the type of event. Here, we find three interpretations: *guillotined, beheaded* and *murdered.* The first source indicates both that Louis XVI was guillotined and that he was beheaded, because we know that guillotining is a form of beheading. The third source classifies the event as *murder.* This interpretation is not given by the other two sources and the third source, in turn, does not provide any information of how the murder took place.

The links between instances and mentions already provide the first indication of alternative perspectives, because it provides insight into what information is provided by whom. Mentions can be associated with *attributions* in the perspective layer. Attributions can provide information about the truth value assigned by a source as well as the source's sentiment. In this case, all sources claim that the event took place (not denying

it, nor placing it in the past) and they do not express any doubt (they are certain). If another source would deny this ever happened or were to express doubt, this can be represented here. Sentiment can be represented about entities as well as events: Louis XVI is described negatively in statement 1 and positively in statement 3. Statement 3 also condemns the execution itself by labelling it as a 'murder'.

The example above illustrates how this standard representation allows researchers to place information from different sources next to each other directly. This can lead to more complete information when sources complement each other and insight into different perspectives when sources contradict each other. Provenance indicating the processes that led to a particular interpretation is always registered. Users can select the interpretations of the processes that fit their needs best. This has the additional advantage that we can generate new interpretations as technology improves without losing previous interpretations that may still be relevant for reasons of replicating previous experiments. In developing this type of analysis, social science literature on 'triangulation' will be helpful. Within positivist research traditions, there is a well-established way of talking about multiple perspectives in data in the field of data triangulation.

2.2. Data Enrichment

Data enrichment will consist of two main processes: linking of datasets (and thereby improving accessibility) and automatic annotations of available data. This work will build upon projects such as CLARIN, CLARIAH, DARIAH and AAA data science, but extend their efforts in the following dimensions:

- deep semantic analyses: we will integrate technologies such as event detection, semantic role labelling (including links to conceptual resources that capture events and their participants and implication of events, such as FrameNet or ESO)
- wider variation of data so that additional domains relevant to the humanities are covered
- evaluation of tools explicitly tailored towards research in the humanities

We illustrate the potential and need for these additions by elaborating a widely applicable research topic in the humanities: relatedness. Relatedness between data can apply on various levels that may be relevant to researchers:

- common subjects (datasets address the same overall topic)
- common genre/style
- common time period
- similar or same author
- shared elements occurring

These kinds of relatedness can either be identified by comparing metadata accompanying sets (mainly for subject, time period, authorship) or by analyzing the data itself. Relevant information found in the (meta)data leads to links with the data to instances. Common properties of datasets can thus be found directly through shared links (e.g. all plays are linked to the genre 'play', any text that mentions the city *Amsterdam* is linked to the instance of *Amsterdam,* etc.). Because even minor references to an entity can lead to a link, researchers may discover information that was previously discarded, because it was 'hidden' in a document the overall topic of which did not seem relevant at first sight.

We will combine NLP tools that identify events and situations together with the roles of individual participants in these situations and events. We can combine this technology with tools that establish coreference, determine syntactic dependencies between terms and can identify the meaning (word sense disambiguation) and reference (named entity recognition and disambiguation) of individual expressions as done in recent NLP and Digital Humanities projects such as NewsReader and BiographyNet. This allows for very precise analysis of how specific terms and the concepts related to them occur in various texts and periods. This technology will also provide the possibility to identify how a word was used in its very first

occurrences (which would make no sense with a distributional method, given that these occurrences are likely to be extremely sparse).

2.3 Evaluation

The development of this platform requires a twofold evaluation. We will use the term *intrinsic evaluation* to refer to evaluation of the data itself (is the dataset built correctly?) and *extrinsic evaluation* to refer to evaluations that establish whether the data lead to reliable answers to a specific research question (is the dataset right for this type of research?).

Intrinsic evaluation

The first type of evaluation involves a strategic digitization plan as well as standards for the development of metadata and an open access policy allowing for the inspection of inputs and results. All researchers should be able to perform this type of evaluation, but it requires a team of specialists (from the Royal Library and the KNAW institutes) to provide guidelines and procedures for intrinsic evaluation.

Extrinsic evaluation

The second type of evaluation requires the creation of gold standard evaluation data that can be used to evaluate the output of generation collections and tools. Whether a specific research question can be addressed reliably through the use of a given set of data cannot necessarily be answered by just looking at the outcome of the intrinsic evaluation. For instance, a research question that examines whether one newspaper writes more negatively about a subject than another can get less reliable results of a sentiment analysis tool that has an Fscore of 80%, but always misses specific negative terms that are used in one of the sources than from a system with 70% Fscore, but equally distributed misses over positive and negative remarks across both papers. In order to address this, evaluations should include:

- detailed error analysis that provide insight into the kind of errors that are made when applying the intrinsic evaluation. These errors are indicative for the extrinsic evaluation.

- hypothesis based sampling: the idea of this approach is to sample parts of the data based on the hypothesis that is addressed. This evaluation should include samples that confirm the hypothesis, samples that contradict the hypothesis and samples that provide an inconclusive answer. Each sample is then manually checked. If it indeed confirms, contradicts or leaves open the hypothesis (as it does according to the method), this provides a confirmation that the method is reliable.

Methodological techniques can be developed and tested for use cases that provide the means to set up intrinsic and extrinsic evaluation of the data. These use cases serve three purposes: first, they allow us to get an indication of the increase in quality and expanded opportunities in our analyses (What kind of humanities research questions can be addressed? What insights are gained? How reliable is the outcome of the analyses?). Second, they illustrate the methodological setup for using digitized datasets, automatic analyses, and examining issues related to representability and reliability. Third, the use cases address actual research questions of scholars across various domains.

2.4 Plan for the Tool Development

How can tools be developed? These are the steps that need to be taken for the development of:

- links to original datasets based on their metadata + descriptions
- tools for data enrichment through text mining
- evaluation sets that indicate performance of tools
  - must be varied, so that users see that a tool's reliability varies depending on genre,

time the data was created, OCR quality, etc.
    - methodologies for using datasets and tools:
        - researchers must be made aware that tools on a new data set must be re-evaluated
        - need of intrinsic and extrinsic evaluation
        - data representation (including provenance) and requirements of space when specific enrichments are applied (i.e. what is needed to ensure replicability and how much space is needed to store this)
    - use cases that illustrate possibilities of the sets and provide examples of how to use the methods outlined above
- use cases should:
    - make use of a variety of data as to clearly illustrate the influence of data on tools and methods
    - address a variety of questions (the more possibilities shown, the more they can inspire new research questions and paths of inquiry)
    - involve a variety of technologies, especially technologies that are under development, as to illustrate how we can incorporate analytical advancements

Subproject 3: Architecture

The goal of subproject 3 is to realize an architecture that ensures flexible, long-term access and function of tools and data for digital textual scholarship. Our sustainability strategy is pertinent to the tools and data developed in the context of DESIDERIA, but we propose a generalized architecture that will allow additional tools and data to be serviced and sustained on the infrastructure delivered by DESIDERIA. To this end we propose a generative IT architecture.[5] A generative architecture does not dictate a single blueprint and/or standard for any specific combination of tools and data that is tied to a particular research project or question. Rather a generative architecture implies that the infrastructure is a general platform supporting current and future development and innovation: a technical structure that offers mechanisms that are beneficial for the evolution of an open system or ecology of data and tools.

A generative architecture allows both for the addition of data and tools, for the development of new tools, as well as for the re-use of these tools and data in newly composed workflows. The architecture caters to the different type of access to tools and data for analytic processes (such as raw data access, GUI access, workflow creation, API access, and so forth). The research design of an individual project can be expressed as one or multiple *pipelines*. These combinations of tools run compartmentalized (or 'modular') in their own execution containers, pulling data transparently from a single storage layer (facilitated by CLARIAH). This architecture essentially provides any project with a separate sandbox that allows it to use or develop tools and workflows (that combine into 'pipelines') using any digital software technologies of their choice, but it guarantees the proper execution, data handling, provenance tracking, and sustainability of these sandboxes.

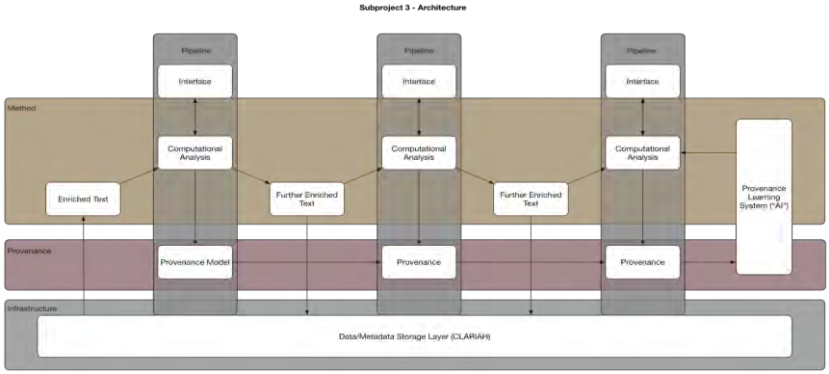
What sets this architecture apart from existing initiatives in Digital Humanities and digital textual scholarship is that it does not seek synchronization and interoperability of tools and produced data sets (as experimented with in the project WebLicht by Volker Boehlke[6]) but rather seeks synchronization through the development of pipelines and provenance for workflows. We add an automated provenance learning system that monitors data usage, tool chaining, pipeline execution, queries and query responses, and so forth. This allows the

[5] The term *generative* is drawn from Zittrain and used to mean "a technology's capacity for leverage across a range of tasks, adaptability to a range of different tasks, ease of mastery, and accessibility." Zittrain, Jonathan L. "The generative internet." *Harvard Law Review* (2006): 1974-2040.

[6] http://weblicht.sfs.uni-tuebingen.de/webservices/Volker-Boehlke-20101116_DSpinRepo_no_urls.pdf

infrastructure to infer workflow patterns and suggest typical usage scenarios, applicable data sets, and tools to users of the DESIDERIA infrastructure. The 'logging, analyzing, and learning' of workflows is aimed at enhancing the methodical explication of DH workflows, and by that at the explication, sharing, and dissemination of DH methodologies.

3.1 Pipelines and provenance

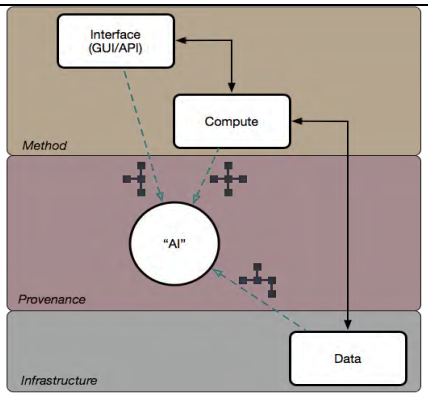


The architecture of Subproject 3

Researchers access, use, and analyze data inside a pipeline. Pipelines are "containerized" research workflows and are the fundamental unit of Subproject 3. Pipelines leverage recent advancements in virtualization and containerization of computation,[7] which compartmentalizes the details of individual pipelines for researchers and their specific needs. The design, use, or reuse of a pipeline generates a provenance record documenting the workflow. Pipelines are designed in such a way that this provenance record is automatically (or semi-automatically) generated and stored by the system. These provenance records inform a machine learning system that can provide deeper insights to computational processes contained within the pipelines.

Many researchers will use existing pipelines (made available through web based GUI interfaces), but change specific parameters to suit their specific needs. The design of these pipelines will automatically and transparently generate provenance records (with as little intervention as possible). Advanced researchers will want to design their own custom pipelines, with specific guidance (and support) for helping them build pipelines that generate the proper provenance records necessary for the "AI" subsystem.

[7] Boettiger, Carl. "An introduction to Docker for reproducible research, with examples from the R environment." *arXiv preprint arXiv:1410.0846* (2014) and Zheng, Chao, and Douglas Thain. "Integrating Containers into Workflows: A Case Study Using Makeflow, Work Queue, and Docker." (2015*). Proceedings of the 8th International Workshop on Virtualization Technologies in Distributed Computing*, ACM New York, 31/38.



A conceptual model of a pipeline

The diagram above describes the conceptual design of an individual pipeline supporting the computational analysis and use of data (primarily text) within DESIDERIA. As mentioned above, pipelines are a *generative,* open-ended architecture, that is, they enable the expression of specific computational workflows (topic modeling, full text search, semantic enrichment, etc.) within the context of a generic computational infrastructure. *Layers* are distinct abstract functional components that interconnect to form a pipeline. A pipeline has three layers:

- **Data Infrastructure Layer** - Builds upon existing data infrastructure (CLARIAH) and provides access to data and metadata.
- **Method Layer** - Provides programmatic access points (APIs) and a human-centric interfaces (GUIs) for designing, executing, and sharing pipelines.
- **Provenance Layer** - Leverages provenance descriptions of pipelines to infer patterns and suggest data collections, algorithms, and potentially applicable workflows.

3.2. Pipeline Layers

Each of these layers work in concert to support a variety of different uses, from heavy computational access, to rich interactive browsing, to long-term preservation and storage. A more in-depth description of the functions, and roles of these layers is to be found below.

Data Infrastructure Layer

The primary focus of this layer is on data storage and preservation. This layer is agnostic to the types of data, no applications or databases are running in this layer, it is just file storage. This is a "dumb" layer in that the modes of getting data in and out of this layer are standardized around a set of APIs. Accessing data through these APIs automates the production of provenance records in PROV[8] that feed into the provenance layer.

Method Layer

The primary focus of this layer is computation and analysis. Computation occurs within pipelines, which can be long running, interactive, or batch-like through re-usable and portable containers. The method layer

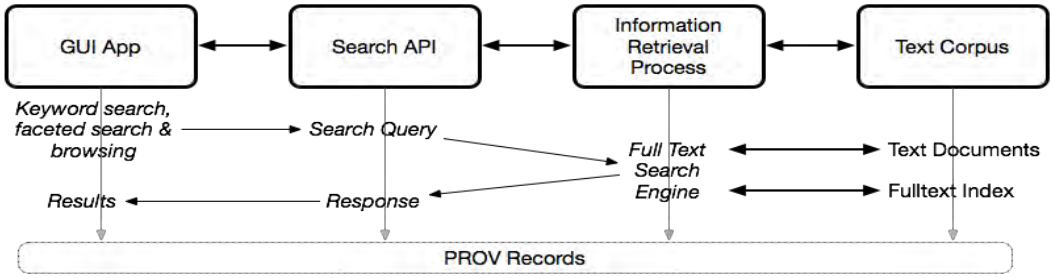
[8] http://www.w3.org/TR/prov-overview/

supports interactive and exploratory computation through command line or "notebook" based workflows (i.e. Project Jupyter[9]). It is a playground/sandbox for developing models and algorithms which can then be "productionalized" as APIs. APIs provide a mechanism by which algorithms and computational models can be "published" for a larger audience through refined access controls and provides an interface to trigger or execute parameterized pipelines enabling easy re-use. The method layer allows for rich GUI applications (built on top of APIs) allowing non-technical users to enrich and analyze data. Pipelines expressed in the method layer are designed to automatically generate provenance records that feed into the provenance layer (reducing the burden of individual users).

Provenance Layer
The provenance layer ingests behavioral data in the form of PROV records and analyzes the workflow descriptions to detect patterns in research processes, provide recommendations for pipeline designers, or assist researchers in running analyses. This layer documents and analyzes computational workflows that are expressed in the other layers of the system. This is an experimental layer that seeks to combine machine learning with workflow/provenance models to generate insight into computational analysis. For example, this layer enhances the research process by implementing a recommender system to assist researchers in the selection of data, algorithms, or computational models.

3.3. Pipeline Examples and Pilot
Each of these examples describes how a particular workflow, analysis, or function would be technically organized and how data would flow within the architecture of subproject 3.

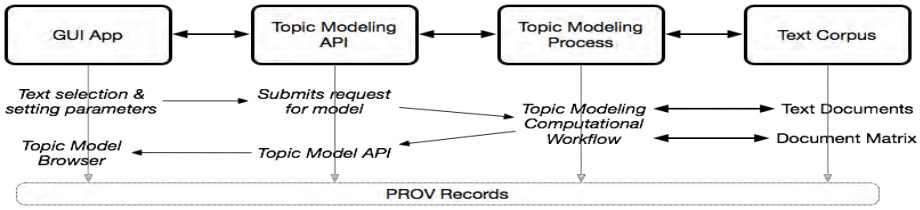
Full Text Search



A researcher can explore a textual corpus via a graphical full text search interface running as a GUI web application in the *method layer* to access texts in the *data infrastructure layer*. This provides a easy-to-use interface for keyword(s) and faceted search, and browsing for quickly finding documents of interest. The GUI application submits search queries via a full-text search API (which can be exposed to other pipelines as well). A search query submitted to the search API communicates with a full text search engine, such as Apache Solr, and to find documents matching the query. Results (in the form of document IDs or documents themselves) are sent back to the GUI App via a response through the search API. Search queries and result-sets would automatically generate PROV records for documenting the researcher's data browsing and search process for the *provenance layer*.

[9] http://jupyter.org/

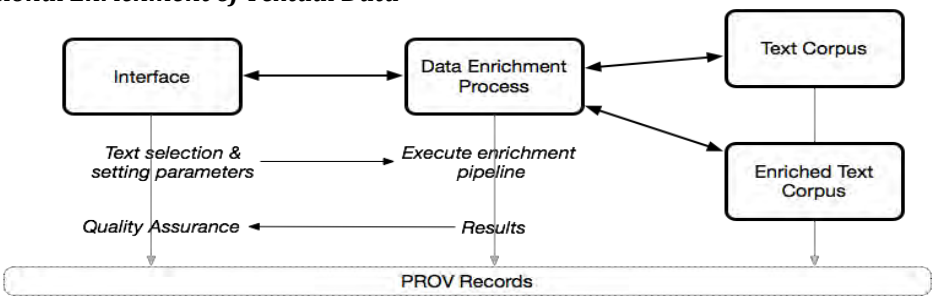
Topic Modeling



Non-technical researchers can perform "distant reading" of a large collection of text documents using a web application at the interface of a topic modeling pipeline. Researchers begin by selecting texts to be modeled via a full text search and browsing interface (as described above). A request for a topic model is submitted via an API that initiates a topic modeling process in the pipeline. The requested texts are fetched from the *data infrastructure layer,* processed into a document matrix, and a topic model is trained on those data. The trained model results (topic-word distributions and document-topic distributions) are exposed via API to the GUI application for the researcher. The topic modeling pipeline is designed to automatically express PROV records about the selection of texts, the parameters of the topic model, and execution of the processes (software versions, times, etc.) for the *provenance layer.* The provenance learning system, the "AI," can then help future users of this pipeline (or other similar topic modeling pipelines) by providing recommendations of text collections or parameters for the topic model.

Along similar lines, the pipeline of post-OCR correction described in subproject 1 will serve as an additional pilot for the further development, testing and refining of this model.

3.4 Computational Enrichment of Textual Data



For subproject 3, the research tools and workflow described in subproject 2 are an instance of a pipeline. The research team of subproject 2 can develop and reuse a number of tools that apply automated linguistic text analysis algorithms to mine entities out of the textual data access from the CLARIAH infrastructure. Deep semantic analyses are used to mine entities, events, common subjects, and sentiments related to these from the data. This information is labeled with estimations of its provenance inferred by the algorithm. This enriched data is then written back to the data layer, and thus added to the general data storage infrastructure (available to other researchers). The inferred provenance of these data are also ingested by the provenance learning system to provide future users with suggestions of data they may be looking for, but are unaware.

3.5 The "AI" Machine-learning Subsystem

The AI, machine-learning subsystem we propose to develop as part of the Huygens ING contribution to the development of DESIDERIA is a high risk/high reward part of the architecture. As researchers develop and add tools, and as they design and use research workflows from several tools, they leave traces of these activities on the infrastructure. Also the queries and analyses that the pipelines generate or represent are

monitored. These patterns of usage are stored (encrypted and anonymized obviously) to serve as data for a classifier and suggestion system that will grow better over time. In a sense, the AI subsystem regards tools and data as the 'vocabulary' used by researchers to constructs workflows that are seen as 'sentences' by the AI subsystem. In the course of time this will allow the AI component to infer one or more 'grammars' that underlies patterns of usage on the infrastructure. This 'grammar' can be used to pre-evaluate new pipelines and to assist users in creating new research designs. Given the computing intensive tasks that are involved with the execution of this "AI" component of the infrastructure, it is expected that it will be hosted on a specialized high performance supercomputing infrastructure. This will in all likelihood involve an extension of the CLARIAH infrastructure in collaboration with an appropriate hosting partner (such as for instance SARA).

Beschrijf welke onderdelen/technieken beproefd zijn en welke geheel of gedeeltelijk nieuw?

DESIDERIA aims to build on previous efforts conducted by libraries, research institutes as well as individual researchers, in the sense that it incorporates existing digital collections and tools as well as previous experiences in crowd sourcing data management and development of textual infrastructures. What we propose would not be possible without these previous efforts and experiences, and the technical standards and procedures developed in previous projects.

Uitdagingen en risico's

Beschrijf de belangrijkste technische knelpunten en geef aan hoe deze opgelost zouden kunnen worden.

- mobilizing and coordinating nationwide digitization efforts of a great number of institutes (libraries, archives, research groups etc.)
- building a technical infrastructure for crowdsourcing that is both flexible and stable, both uniform and tailor made
- developing tools that will bridge the semantic gap in textual data
- developing and implementing a nationwide infrastructural architecture for the use and re-use of textual data, analyses and software.

Beschrijf de belangrijkste risico's.

- as mentioned above, the AI machine-learning subsystem that is part of subproject 3 is high-risk, high reward, which may fail
- connecting pure digitization efforts and the development of technical and tool infrastructures.

B. INBEDDING

Hoe past dit voorstel in het (internationale) landschap van grote onderzoeksfaciliteiten?

Like CLARIAH, DARIAH.NL and CLARIN.NL, DESIDERIA seeks international cooperation and the implementation of international standards and protocols to succeed. Creating and assuring an international embedding will be one of the first steps taken when we execute DESIDERIA's plans.

Hoe wordt de nationale toegang gegarandeerd?

The KNAW Humanities Group (Huygens ING, IISG and Meertens Institute) will facilitate Open Access to DESIDERIA's program, working together with libraries and research groups that produce the digital collections and tools that DESIDERIA is built upon.

Sluit het aan op reeds bestaande faciliteiten?

See A1.

Zijn er voor zover bekend vergelijkbare ideeën (of al bestaande faciliteiten) in het buitenland? Zo ja, zou Nederland een aparte nationale faciliteit moeten hebben of betreft dit een internationale faciliteit op Europees of mondiaal niveau?
At this moment, the recently installed Permanente Commissie voor Grootschalige Wetenschappelijke Infrastructuur (http://www.nwo.nl/over-nwo/aandachtsvelden/grote+onderzoeksfaciliteiten/permanente+commissie) is mapping existing research infrastructures for textual data in Dutch. Representative textual infrastructures that allow for longitudinal research into conceptual shifts are not built in the Netherlands, nor in any other country or in any other language. Such facility would presently be unique is the world, but has an infrastructure to build upon in the Netherlands thanks to a large amount of digitized texts from most historical periods in existing programmes, and in the way the development of textual corpora and digital tools is interwoven with projects such as CLARIN-NL, CLARIAH and NEDERLAB. At this moment various Dutch research projects are already exploring possibilities for long term deep data research: AAA Data Science (UvA/VU), Translantis: America as Reference Culture (UU), and Rens Bod's NWO project on conceptual history. DESIDERIA will accelerate progress in the development of semantic tools by uniting the expertise of these groups.
Hoe past het voorstel bij de NL sterktes van onderzoek?
The Netherlands have a strong tradition in both the infrastructural and historical/linguistic expertise needed to develop the proposed infrastructure. The Dutch contribution to European programs such as CLARIN-EU, but also DARIAH-EU has been significant in the past, is presently being continued in CLARIAH and will keep its pace and momentum with the development of DESIDERIA.
Beschrijf de voordelen/belang voor NL indien zo'n faciliteit zou worden gerealiseerd. Dit mogen zowel wetenschappelijke als economische of maatschappelijke voordelen zijn.
The proposed infrastructure would secure a leading position for Dutch research groups interested in textual digital scholarship. It would also serve as a bridge between Information Sciences (to information scientists interested in conceptual frameworks and ontologies) and the Humanities and Social Sciences, and stimulate the existing expertise among Dutch groups in such a manner that they could take an international lead.
C. ORGANISATIE en FINANCIËN
Geef aan welke partijen/expertise nodig zijn voor de ontwikkeling van deze faciliteit. Geef ook aan of en zo ja hoe deze al zijn betrokken.
Already involved (either as applicants or because they were consulted in writing this application) and vital to the development of DESIDERIA: Royal Library Meertens Institute Huygens ING IISG Research groups at all Dutch universities Netwerk Digitaal Erfgoed The next would be to approach international groups and platforms to create international embedding of DESIDERIA's plans. The technical choices made in DESIDERIA (Open Access, Open Linked Data) ensure international interoperability.
Beschrijf de (mogelijke) organisatiestructuur. Geef ook aan of er al een begin van organisatievorming is.

The applicants form the current and ad hoc organisation of DESIDERIA. They have consulted a great number of stakeholders (in the Royal Library, at universities) that should also be involved in the next step towards realization of DESIDERIA, At the beginning of this next step, a more stable organisation needs to be set up, consisting of some of the applicants and well as some of the stakeholders.
Geef een globale beschrijving van de business case. Hoe zou deze faciliteit gefinancierd kunnen worden? Ga hierbij in de drie fases: ontwikkeling, bouw en exploitatie.
DESIDERIA could partly be developed using existing budgets for large scale digitization in heritage institutions such as the Royal Library (parts of subproject 1). The development of tools could be partly sponsored by the eScience institute (parts of subproject 2), and the development of the data architecture could be submitted as an NWO infrastructural application. Still, developing DESIDERIA in full would require additional funding. It is unlikely this funding would derive from IT companies, but it could be worthwhile to develop some elements of the envisioned tools with specialized companies interested in conceptual or semantic analyses.
D. VERDERE ONTWIKKELING
Beschrijf wat er moet gebeuren om deze faciliteit verder te ontwikkelen. Ga in op de belangrijkste knelpunten die opgelost moeten worden.
Resolving money issues would be the first hurdle. Once these issues are cleared, or even when part of the budget is secured, a team of developers selected from leading groups and research institute could start working along the lines of the working program outlined in this application and the seek alliances in international contexts.
Geef aan wat de ontwikkeltermijn voor deze faciliteit ongeveer zou kunnen zijn. We propose to start DESIDERIA in 2019/2020, following the completion of CLARIAH.