NEDERLAB RICHT ZICH OP GEESTESWETENSCHAPPERS

Toegangspoort tot digitaal onderzoeksparadijs

Een grote groep geesteswetenschappers wil de historische veranderingen in de Nederlandse taal en cultuur in kaart brengen. Dat kan alleen met een nieuw onderzoeksinstrumentarium. En dat is precies het doel van de oprichting van Nederlab. Nicoline van der Sijs licht het nieuwe project toe.

Nicoline van der Sijs

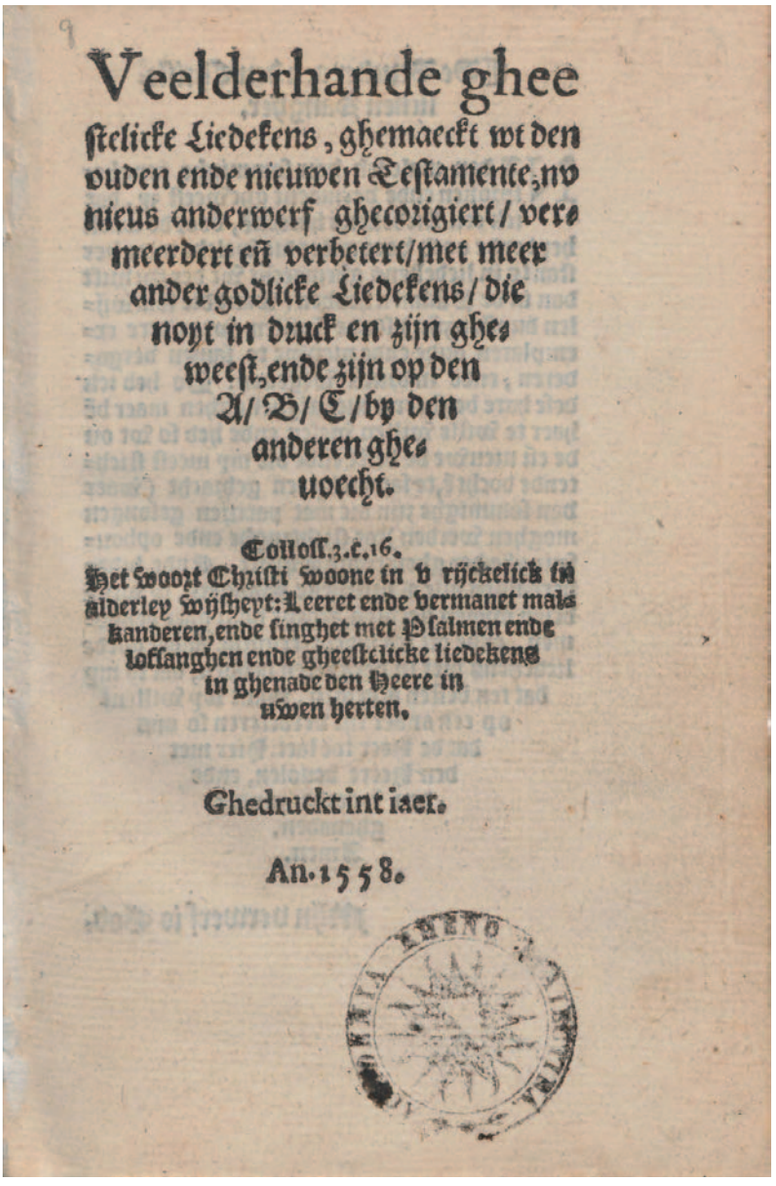
‘Nederlab wil een onderzoeksomgeving creëren waar geesteswetenschappers onderzoek kunnen doen naar de Nederlandse taal en cultuur’

Sinds de eerste erfgoedinstellingen zo’n jaar of tien geleden een eerste aarzelende schrede op het digitaliseringspad zetten, is er veel gebeurd. Hoewel veel instellingen – bibliotheken, archieven, onderzoeksinstituten, musea – nog slechts een klein deel van hun collectie gedigitaliseerd hebben, realiseren ze zich allemaal dat de toekomst ligt in de digitale wereld: gebruikers vragen om digitale toegang tot collecties, en de ervaring heeft inmiddels geleerd dat een digitale collectie veel vaker wordt geraadpleegd dan een papieren collectie. Iedere instelling worstelt met de vraag hoe de digitalisering het best, efficiëntst en goedkoopst kan worden uitgevoerd, en hoe de collectie het beste vindbaar kan worden gemaakt.

Er zijn allerlei instanties in het leven geroepen om instellingen te ondersteunen en te adviseren bij de digitalisering van tekstdata, de uniformering van de metadata en het aanbieden of ontwikkelen van tools: computerprogramma’s waarmee gebruikers hun weg door de data en metadata kunnen vinden. Stichting DEN (Digitaal Erfgoed Nederland) en de Vlaamse tegenhanger FARO geven adviezen over de beste aanpak van digitaliseren. Het infrastructuurprogramma CLARIN (met zijn beoogde opvolger CLARIAH) bevordert dat verschillende formaten en tools naadloos met elkaar samenwerken en geuniformeerd en geharmoniseerd worden. SURF en SARA leveren ict-infrastructuurdiensten aan universiteiten en hogescholen op het gebied van bijvoorbeeld clouddataopslag en beheerssystemen voor de toegang tot werkruimtes. Onder andere het KNAW/NWO-instituut DANS en het Max Planck Instituut Nijmegen zorgen voor het duurzaam opslaan van gegevens. Omvangrijke digitale tekstbestanden worden inmiddels beschikbaar gesteld door de Koninklijke Bibliotheek (KB) en de universiteitsbibliotheken. Deze bestanden zijn via massadigitalisering totstandgekomen: tientallen miljoenen pagina’s uit boeken, tijdschriften en kranten zijn gescand en vervolgens door een programma voor optische tekenherkenning (ocr) gelezen. Het nadeel van deze methode is dat de computer nog steeds veel leesfou-

Wie is...

Nicoline van der Sijs is historisch taalkundige. Ze publiceerde talloze boeken over de geschiedenis van het Nederlands. Ze is vaste medewerker van *Onze Taal* en wetenschapscolumnist bij NRC Handelsblad. Van der Sijs werkt als projectleider voor Nederlab in dienst van het Meertens Instituut.



Veelderhande ghee
ſtelicke Liedekens, ghemaeckt wt den
ouden ende nieuwen Teſtamente, nv
nieus anderwerf ghecorigiert / ver-
meerdert eñ verbetert / met meer
ander godlicke Liedekens / die
noyt in druck en zijn ghe-
weeſt, ende zijn op den
A / B / C / by den
anderen ghe-
uoecht.

Colloſſ. 3. c. 16.
Het woort Chriſti woone in v rijckelick in
alderley wijſheyt: Leeret ende vermanet mal-
kanderen, ende ſinghet met Pſalmen ende
loffanghen ende gheeſtelicke liedekens
in ghenade den Heere in
uwen herten.

Ghedruckt int iaer.

An. 1558.

Transcripties van handgeschreven en gedrukte teksten worden onderdeel van het Nederlab-corpus: een in het gotisch gedrukte titelpagina van een liedbundel uit 1558

ten maakt, bijvoorbeeld *fchip* in plaats van *schip*. Kleinere, maar nog steeds zeer substantiële digitale tekstbestanden die handmatig zijn gecorrigeerd of exact zijn getranscribeerd van het origineel, zijn vervaardigd door de Digitale Bibliotheek voor de Nederlandse Letteren (DBNL, 3 miljoen pagina's) en wetenschappelijke instituten als Huygens ING, het Instituut voor Nederlandse Lexicologie (INL) en het Meertens Instituut.

Er staan dus veel spelers op het digitaliseringsveld. De enigen die op dit speelveld nog ontbreken, is de groep van (wetenschappelijke) gebruikers. In juni 2012 hebben ook zij een stem gekregen: toen heeft namelijk het project Nederlab een omvangrijke subsidie ontvangen van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), namelijk 2.048.000 euro. Andere instellingen hebben daar nog een extra bedrag bijgelegd: de KNAW (600.000 euro), CLARIAH (250.000 euro) en CLARIN (150.000 euro). Verder matchen Meertens Instituut, Huygens ING, INL, DBNL, de Nederlandse Taalunie en de universiteiten, waarmee met de investering in totaal 4 miljoen euro is gemoeid.

Gebruiker centraal

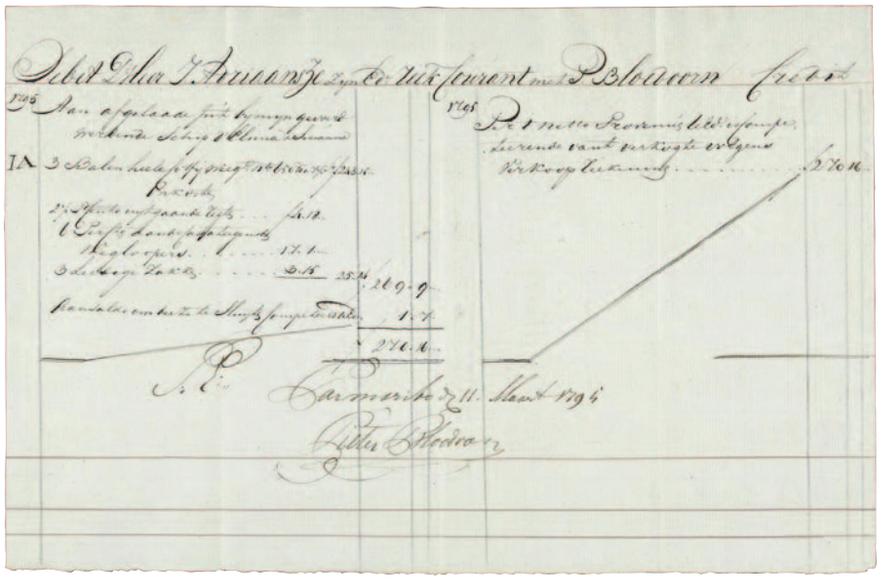
Met die subsidie wil Nederlab een onderzoeksomgeving, een laboratorium, creëren waar geesteswetenschappers onderzoek kunnen doen naar de veranderingspatronen in de Nederlandse taal en cultuur. De aanleiding tot de subsidieaanvraag was het feit dat de huidige digitale wereld voor veel soorten onderzoek nog niet geschikt is. Geesteswetenschappers – taalkundigen, letterkundigen en historici

'Aan het raadplegen van digitale teksten stellen geesteswetenschappers hoge eisen'

– zijn grootverbruikers van digitale teksten, die ze gebruiken als onderzoekscorpus: in de gedigitaliseerde teksten vinden we immers de neerslag van de Nederlandse taal en cultuur. Juist doordat geesteswetenschappers digitale teksten intensief raadplegen, stellen zij hoge eisen, waaraan momenteel nog niet wordt voldaan. In 2011 heeft daarom een aantal wetenschappelijke instituten, onder leiding van het Meertens Instituut, de koppen bij elkaar gestoken, om te bekijken op welke manier de situatie kan worden verbeterd en wat dat zou kosten aan werk en middelen. In opdracht van deze instituten heb ik 150 onderzoekers geconsulteerd over de vraag wat voor onderzoek ze willen verrichten en welke data en metadata ze daarvoor nodig hebben.

Dat leverde – het zal geen verrassing zijn – een stortvloed aan onderzoeksvragen op. Sommige onderzoekers willen weten sinds wanneer een bepaald woord (*democratie*), woordvorm (*jullie lopen* in plaats van het oudere *jullie loopt*) of woordcombinatie (*zich irriteren*) voorkomt en in welke context en betekenis. Andere onderzoekers willen weten hoe vaak een woord, woordcombinatie of naam door de eeuwen heen in teksten voorkomt: aan de hand van het aantal vermeldingen van bijvoorbeeld de auteursnamen Joost

‘Alle losse, gedigitaliseerde tekstbestanden dienen als eenheid doorzoekbaar gemaakt te worden’



Achttiende-eeuwse schuldbekentenis (afkomstig uit de documenten die Engelsen geconfisqueerd hebben op gekaapte schepen; origineel in The National Archives, Kew, Londen, kopie in Nationaal Archief Den Haag)

van den Vondel, Constantijn Huygens en Gerbrand Adriaensz. Bredero kunnen conclusies getrokken worden over de wisselende populariteit van deze drie zeventiende-eeuwse schrijvers in latere eeuwen. Nog weer andere onderzoekers willen automatische tekstvergelijkingen met de computer uitvoeren om plagiaat, citaten of parafrasen op te sporen, om metaforen te herkennen (*drankzucht* als ‘kanker van de maatschappij’) of om teksten waarvan auteur, datering of herkomstplaats onbekend zijn, te herleiden tot een specifieke auteur, periode of regio.

De beantwoording van al dit soort onderzoeksvragen levert bouwsteentjes voor het uiteindelijke, hogere doel van een grote groep geesteswetenschappers: het in kaart brengen van de historische veranderingen die binnen de Nederlandse taal en cultuur in de loop van vele eeuwen hebben plaatsgevonden, en het achterhalen welke factoren verantwoordelijk zijn voor het optreden van die veranderingen. Dat kan echter alleen met een nieuw onderzoeksinstrumentarium. En dat is precies het doel van de oprichting van Nederlab.

Breed gedragen

Om langetermijnveranderingen in de taal en de cultuur te kunnen traceren, is een heel groot corpus aan gedigitaliseerde teksten nodig, van de oudste geschreven periode (circa 800) tot heden, met allerlei soorten teksten (fictie, non-fictie et cetera), die representatief over de hele periode zijn verdeeld. Op dit moment zijn er wel veel historische teksten gedigitaliseerd, maar ze worden door een groot aantal instellingen op verschillende plaatsen aangeboden. Iedere instelling biedt zijn eigen zoekinterfaces en zoekmogelijkheden, er bestaan aanzienlijke kwaliteitsverschillen tussen de verschillende corpora, en iedere instelling voegt zijn eigen metadata toe. Het gevolg hiervan is dat al deze tekstbestanden – en hun metadata – slechts naast elkaar, en niet tegelijkertijd en samen, kunnen worden doorzocht en geanalyseerd. Voor de beantwoording van langetermijnveranderingen is het noodzakelijk dat alle losse tekstbestanden als eenheid (gedistribueerd) doorzoekbaar gemaakt worden. Dat is de vurige wens van de geesteswetenschappelijke wereld.

Die wens wordt gelukkig gedeeld door de dataleveranciers (de wetenschappelijke bibliotheken), infrastructuurorganisaties en toolontwikkelaars: ook zij zien de enorme voordelen van het aan elkaar koppelen, harmoniseren en uniformeren van omvangrijke tekstbestanden en metadata. Dat vergt echter extra inspanning van iedereen. Om die mogelijk te maken hebben enkele instellingen die onderzoek doen naar de Nederlandse taal en cultuur gezamenlijk op 1 november 2011 een subsidieaanvraag bij het programma Investeringen NWO-groot ingediend voor de oprichting van Nederlab: een gebruiksvriendelijke, algemeen toegankelijke en met tools verrijkte gebruikersomgeving, waarbinnen alle gedigitaliseerde teksten die relevant zijn voor de geschiedenis van de Nederlandse taal en cultuur zijn bijeengebracht. De Raad van Bestuur bestaat uit prof.dr. Hans Bennis (penvoerder en initiatiefnemer, Meertens Instituut), Cees Klapwijk (DBNL), dr. Nicoline van der Sijs (projectleider, Meertens Instituut) en dr. Henk Wals (Huygens ING). De aanvraag voor Nederlab wordt gesteund door de hele geesteswetenschappen: alle universiteiten zijn vertegenwoordigd in een van de vier adviesraden.

Spin in digitale onderzoeksweb

Dankzij deze subsidies wil Nederlab uitgroeien tot de spin in het wetenschappelijke digitale onderzoeksweb. En daarbij gaan we zeker niet het wiel opnieuw uitvinden. Nederlab gaat niet zelf tekstbestanden digitaliseren – die komen van de wetenschappelijke bibliotheken. Evenmin zal Nederlab nieuwe tools gaan ontwikkelen; wel worden bestaande tools aangepast zodat ze geschikt worden gemaakt binnen de infrastructuur en kunnen werken op historische teksten: de meeste tools

– voor bijvoorbeeld zoeken en datamining – zijn ontworpen voor moderne teksten. De technische infrastructuur wordt als een Ikea-kast opgebouwd uit bestaande onderdelen: technologieën voor het verlenen van toegangsrechten voor gebruikers komen van SURF, virtuele werkruimtes komen uit verschillende door NWO en KNAW gesubsidieerde toolsprogramma's, voor zoektechnologieën werken we samen met specialisten van de UvA en de Universiteit Delft, het duurzaam opslaan van de data en metadata vindt plaats in overleg met het Max Planck Instituut, het harmoniseren van de verschillende dataformaten en de verschillende standaards vindt plaats binnen CLARIN-verband. Net als bij een Ikea-kast sluiten de verschillende onderdelen lang niet altijd goed op elkaar aan en ontbreken er hier en daar onderdelen, waardoor er nog veel zal worden gevergd van het improvisatievermogen van de technici.

Nederlab wil een laag leggen boven op de portalen met data en metadata van de verschillende instellingen. In het eerste jaar wordt de Nederlab-infrastructuur neergezet, en worden de (getranscribeerde) tekstbestanden en metadata van de DBNL als onderzoekscorpus ingebracht. In de daaropvolgende jaren worden de gegevens gestructureerd uitgebreid: daarbij worden bijvoorbeeld de auteursgegevens van de KB gekoppeld aan die van de DBNL. Dat moet eenmalig met de hand gebeuren (iemand moet beslissen of Jan Janssen uit de KB dezelfde is als Jan Janssen uit de DBNL). Is de koppeling eenmaal gelegd, dan worden voortaan alle werken van Jan Janssen automatisch aan elkaar gekoppeld, ook werken die hij in de toekomst nog zal publiceren.

Betrouwbare metadata

Voor onderzoekers is het heel belangrijk dat teksten zijn voorzien van geüniformeerde en gedetailleerde metadata. Wil je bijvoorbeeld uitspraken kunnen doen over het taalgebruik in de zeventiende eeuw, dan moet een twintigste-eeuwse tekstuitgave van Vondel niet in zijn geheel tellen als twintigste-eeuwse publicatie, maar er moet een scheiding worden gemaakt tussen de oorspronkelijke – zeventiende-eeuwse – tekst en het voor- en nawerk van de twintigste-eeuwse editeur. Als je een dergelijke scheiding niet maakt, kan een naïeve onderzoeker immers concluderen dat Vondel al woorden als *tof* en *oké* gebruikte, terwijl die in werkelijkheid op het conto van de editeur geschreven moeten worden – en dus stammen uit de twintigste eeuw.

Ook moet bij teksten worden aangegeven hoe betrouwbaar de data zijn voor onderzoek: teksten die met ocr gelezen zijn, bevatten vaak veel leesfouten en zijn daardoor ongeschikt voor statistische analyses. Een van de doelstellingen van Nederlab is om te bevorderen dat de onderliggende data worden gecorrigeerd: daarvoor is de subsidie niet toereikend, maar we willen wel aanmoedigen dat tekstcorrecties uitgevoerd gaan worden door middel van crowdsourcing – waarmee het Meertens Instituut inmiddels veel en zeer positieve ervaring heeft. Daarnaast zal ook worden gewerkt aan de automatische correctie en verbetering van de zoektechnologie, waardoor leesfouten omzeild kunnen worden: daartoe werkt Nederlab samen met Nederlandse toolontwikkelaars als de door NWO-gesubsidieerde projecten Catch en CatchPlus om het culturele erfgoed digitaal te ontsluiten. Ook zullen de resultaten worden benut van het onlangs afgesloten Europese IMPACT-project dat zich bezighoudt met de verbetering van de optische tekenherkenning.

Meerwaarde

Nederlab vormt zo het eerste gemeenschappelijke platform voor geesteswetenschappers, dataleveranciers (bibliotheken) en technici. Omdat Nederlab vanuit één centraal punt alle bestaande digitale tekstbestanden tegelijkertijd doorzoekbaar maakt, zal het voor veel onderzoekers het beginpunt voor hun onderzoek worden. Iedere onderzoeker krijgt een eigen virtuele werkruimte binnen Nederlab waar hij, alleen of met andere onderzoekers, data kan verzamelen en met tools bewerken. De verwachting is dat de infrastructuur van Nederlab zal leiden tot samenwerking en synergie binnen de geesteswetenschappen en tot het stellen van nieuwe, veelal interdisciplinaire, onderzoeksvragen. Onderzoekers, studenten, promovendi en postdocs zullen zeer nauw betrokken worden bij de inrichting van Nederlab, en er zal veel tijd en energie gestoken worden in het consulteren en informeren van de onderzoekers. Tijdens de duur van het project wordt een helpdesk ingericht, en er komt een digitaal forum waar onderzoekers met elkaar kunnen overleggen.

'Met Nederlab is een investering van 4 miljoen euro gemoeid'

Op deze manier zal Nederlab leiden tot nieuwe gebruikers en nieuwe gebruiksmethoden van de data en metadata. Een ander positief effect van de oprichting van Nederlab is dat alle betrokken technische partijen zullen komen tot afspraken over standaardisering en harmonisering. Dit zal niet alleen leiden tot een betere toegankelijkheid en vindbaarheid van de bestaande corpora, maar ook tot kwaliteitsverbetering en standaardisering van de data en metadata.

Met een betrekkelijk geringe investering wordt zo een enorme meerwaarde verkregen. Dit is conform de toekomstvisie van eurocommissaris Neelie Kroes in de inleiding van het rapport *Riding the wave* (zie cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf): '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 is als nieuwe speler aan de bal: u hoort nog van ons.

www.nederlab.nl <