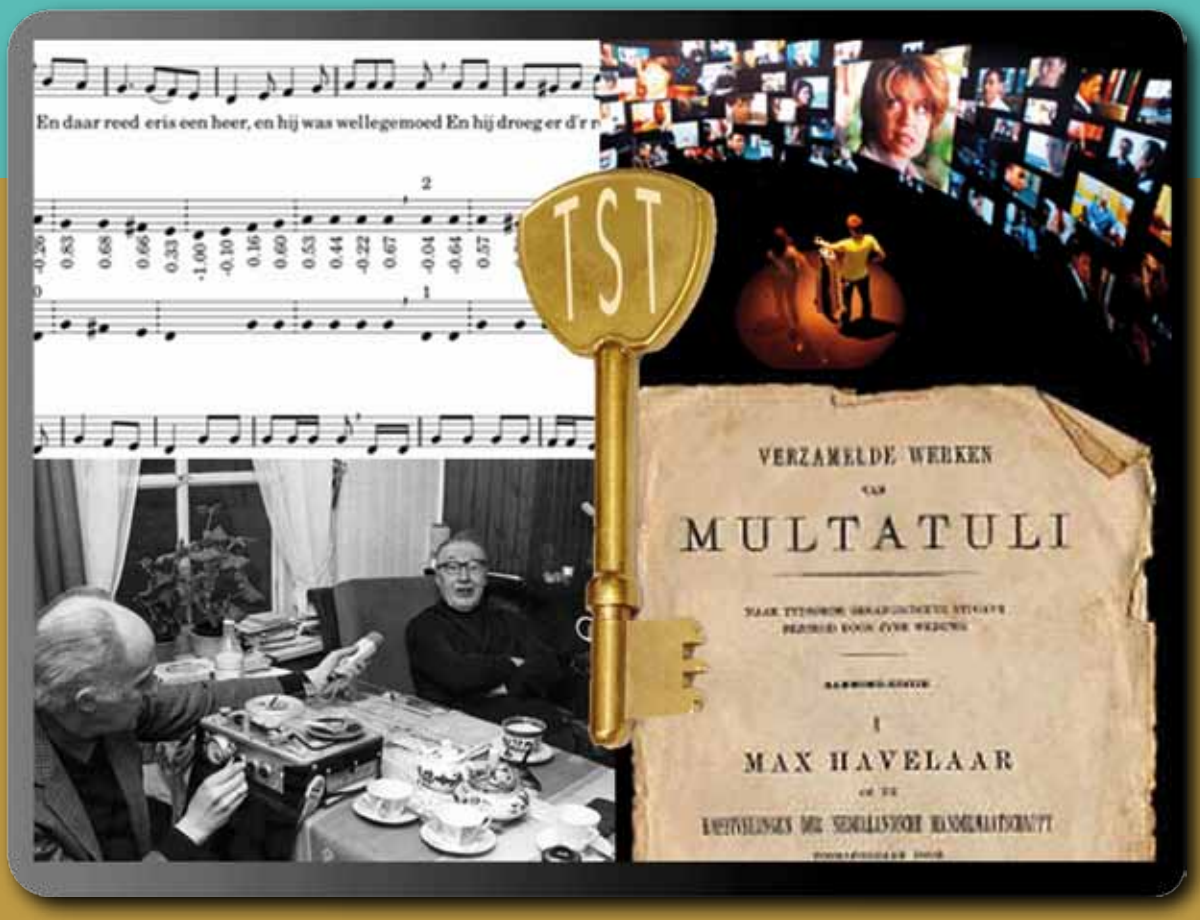
DIXIT

TIJDSCHRIFT OVER TAAL- EN SPRAAKTECHNOLOGIE

TAAL- EN SPRAAKTECHNOLOGIE voor Cultureel Erfgoed



Sleutel tot een schatkist

JAARGANG 7 NUMMER 1 DECEMBER 2010

Instituut voor Nederlandse Lexicologie

Schatkamer van de Nederlandse taal

INL
INSTITUUT VOOR
NEDERLANDSE
LEXICOLOGIE

Het Instituut voor Nederlandse Lexicologie (INL) is een Vlaams-Nederlands instituut dat Nederlandse woorden verzamelt, bestudeert, opslaat in databases, voorziet van allerlei (taalkundige) gegevens en daarmee wetenschappelijke woordenboeken maakt. Of u nu taalkundige bent of simpelweg geïnteresseerd in taal: met al uw vragen over woorden kunt u bij ons terecht.

TST-Centrale

Vlaams-Nederlandse centrale voor beheer, onderhoud en distributie van digitale taalmaterialen

Bent u op zoek naar Nederlandstalige, digitale taalmaterialen? Bij ons bent u aan het juiste adres!

De Centrale voor Taal- en Spraaktechnologie (TST-Centrale) stelt taalmaterialen beschikbaar die veelal met overheidsgeld zijn gefinancierd, waaronder materialen die op het INL ontwikkeld worden en resultaten uit het STEVIN-programma (STEVIN: Spraak- en Taaltechnologische Essentiële voorzieningen voor het Nederlands). Daarnaast ondersteunt de TST-Centrale het gebruik van de materialen door uw vragen te beantwoorden en gastcolleges en workshops te organiseren.

Beschikbare materialen:

- Historische en wetenschappelijke elektronische woordenboeken (o.a. Woordenboek der Nederlandsche Taal online)
- Gesproken, geschreven en multimediale corpora (o.a. Corpus Gesproken Nederlands)
- Mono- en bilinguale lexica (o.a. Referentiebestand (Belgisch-)Nederlands)
- Tools voor gesproken en geschreven teksten (o.a. AUTONOMATA-g2p-toolkit)

De TST-Centrale is een initiatief van en wordt gefinancierd door de Nederlandse Taalunie. De TST-Centrale is ondergebracht bij het Instituut voor Nederlandse Lexicologie met vestigingen in Leiden en Antwerpen.

www.inl.nl
www.inl.nl/tst-centrale
www.inl.nl/producten

servicedesk@inl.nl

nu | TST-Centrale

nu Nederlandse Taalunie

INHOUD

Met dank aan Folkert de Vriend (Nederlandse Taalunie) voor de illustratie op de voorzijde.

Voorwoord

Topje van de ijsberg

We weten dat onze musea en galerieën slechts een klein deel van hun collectie kunnen laten zien. Dankzij Internet kan veel meer getoond worden (de ruimte is in principe eindeloos) maar hoe vinden we wat we zoeken in de immense collecties van geschreven en gesproken archieven, zoals de miljoenen documenten die bij de Koninklijke Bibliotheek liggen of de opnames die in de kolossale kelders van het Nederlands Instituut voor Beeld en Geluid in Hilversum bewaard worden? Tot nu toe hebben zowel bezoekers als onderzoekers slechts het topje van deze ijsberg mogen zien terwijl ze weten dat er meer is. Maar hoe kom je erbij?

TST en Cultureel Erfgoed

Als NOTaS (Nederlandse Organisatie voor Taal- en Spraaktechnologie) hebben we ervoor gekozen u in deze DIXIT iets te laten proeven van de mogelijkheden die TST te bieden heeft bij het toegankelijk maken van ons culturele erfgoed. We zijn blij met de inbreng van diverse auteurs (waaronder een groot aantal nieuwe gezichten) die gezamenlijk een breed overzicht weten te schetsen van de verschillende manieren waarop dit gedaan kan worden.

De grootste audiovisuele schatkist

Voor deze nieuwe DIXIT heeft de redactie een bijzondere gasthoofdredacteur weten te strikken: Johan Oomen, Hoofd Research & Development bij het Nederlands Instituut voor Beeld en Geluid. Wij zijn hem zeer erkentelijk voor zijn waardevolle inbreng bij het samenstellen van deze fascinerende editie. Tevens wil ik een woord van dank richten aan de drie sponsoren van deze DIXIT, Digitaal Erfgoed Nederland (DEN), CLARIN-NL en het Nederlands Instituut voor Beeld en Geluid.

De toekomst

Nu het STEVIN-programma de laatste fase in is gegaan en er gesprekken gevoerd worden over het slotcongres in het najaar van 2011, zijn wij als NOTaS-bestuur, samen met verschillende andere organisaties, druk bezig met het vinden van nieuwe financieringsmogelijkheden voor zowel academisch als meer toegepast onderzoek op het gebied van TST. Alleen door continuering van onderzoeks- en ontwikkelwerkzaamheden zal onze branche de samenleving kunnen blijven voorzien van mooie, nuttige en leuke toepassingen op diverse gebieden zoals onderwijs, zorg, dienstverlening of ons culturele erfgoed.

En nu...

Duik samen met de deelnemers van NOTaS onder het ijs en beleef de immense, fascinerende en tot nu toe verborgen wereld van het Nederlands cultureel erfgoed. Ontsloten dankzij vernuftige Taal- en Spraaktechnologische toepassingen – of u nu onderzoeker, applicatiebouwer of gewoon een geïnteresseerde lezer bent.

Namens NOTaS en de DIXIT- redactie wens ik u veel leesplezier.
Debbie Kenyon-Jackson, Voorzitter

COLOFON

DIXIT: Tijdschrift over toegepaste taal- en spraaktechnologie – 7e jaargang, editie TST voor Cultureel Erfgoed. DIXIT is een uitgave van Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen. Tel. 024 - 352 88 88 – Fax 024 -354 00 90 www.notas.nl **Redactieadres:** Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen **Redactie:** Arjan van Hessen: hessen@cs.utwente.nl, Henk van den Heuvel: h.vandenheuvel@let.ru.nl, Remco van Veenendaal: remco.vanveenendaal@inl.nl, Leonoor van der Beek: leonoor.van.der.beek@q-go.com, Anne Wijnen: notas@malta-online.nl **Hoofdredactie themanummer:** Johan Oomen: joomen@beeldengeluid.nl **Advertenties:** Stichting NOTaS – Anne Wijnen: notas@malta-online.nl, 024 -352 88 88 **Abonnementen:** Voor een gratis abonnement dient u zich te wenden tot een van de NOTaS-deelnemers (zie www.notas.nl) **Druk:** Leonard Nijmegen bv **Verantwoording:** DIXIT is een uitgave van Stichting NOTaS. Overname van de artikelen is alleen toegestaan met bronvermelding en na toestemming van Stichting NOTaS. Stichting NOTaS en de bij deze uitgave betrokken redactie en medewerkers aanvaarden geen aansprakelijkheid voor mogelijke gevolgen die zouden kunnen voortvloeien uit het gebruik van de in deze uitgave opgenomen informatie.

Het belang van taal- en spraak-technologie voor cultureel erfgoed

Recentelijk schreef mediawetenschapper Clay Shirky over de toekomst van kranten in het digitale tijdperk: 'public reuse produces a kind of value that doesn't just come from publication. It comes from republication and reuse'[1]. Ook bij erfgoedinstellingen leeft het bewustzijn dat alleen hergebruik door eindgebruikers de investeringen in behoud van collecties kunnen legitimeren.

Johan Oomen
Nederlands Instituut voor Beeld en Geluid

Digitalisering en publiceren van collecties speelt hierbij een grote rol. Erfgoedinstellingen dragen door het online beschikbaar stellen van collecties bij aan een nieuw informatie-ecosysteem waarin digitale objecten en contextuele informatie onderdeel worden van een steeds omvangrijker wordend netwerk van informatiebronnen. Op het web ontstaat zo een 'Cultural Commonwealth', waarbinnen erfgoedcollecties op een revolutionair nieuwe wijze bestudeerd, gecontextualiseerd en gerepresenteerd kunnen worden[2].

Kennisinstellingen, erfgoedinstellingen en softwareontwikkelaars werken al enige jaren samen bij het efficiënt inzetten van taal- en spraaktechnologie. Zo helpen technologische innovaties bij het exploreren van deze collecties; ze maken het mogelijk om impliciete verbindingen expliciet te maken, nieuwe interacties met het publiek te faciliteren, bestaande processen voor het beschrijven van content te ondersteunen en de doorzoekbaarheid van collecties te verbeteren.

Deze speciale editie van DIXIT laat de breedte zien van de toepassingen van taal- en spraaktechnologie binnen het erfgoedveld. Naast uiteenlopende praktijkvoorbeelden uit het veld komen ook belangrijke onderzoeksprogramma's zoals MediaEval, CLARIN-NL, CATCH en STEVIN aan bod. De bijdragen schetsen een goed beeld van de toepassingen van taal- en spraaktechnologie binnen dit spannende domein en bieden aanknopingspunten voor zowel vervolgonderzoek als daadwerkelijke implementatie van de technologie. De impact van deze technologie op de manier waarop de geschiedenis bestudeerd en verteld kan worden is enorm.

[1] *Shirky, Clay. 'Let a thousand flowers bloom to replace newspapers; don't build a paywall around a public good' www.niemanlab.org/2009/09/clay-shirky-let-a-thousand-flowers-bloom-to-replace-newspapers-dont-build-a-paywall-around-a-public-good/ (laatst bezocht: 26 oktober 2010)*

[2] *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Siences. American Council of Learned Societies, 2006*

Bevrijde Woorden
Van Document- naar Persoons-gerichte ontsluiting van de Handelingen der Staten Generaal

Sinds eind 2010 zijn de volledige Handelingen der Staten-Generaal (vanaf 1814) digitaal beschikbaar. En wel gratis, vrij van rechten en vanuit elke woonkamer via www.statengeneraal-digitaal.nl, een samenwerking tussen de Koninklijke Bibliotheek en de Staten-Generaal. In dit artikel bespreken we een klein, maar wel bekend deel hiervan: de notulen van de vergaderingen in de Grote Zaal.

Maarten Markx
Universiteit van Amsterdam

Dit databestand is een uniek cultureel erfstuk dat een feest is om te lezen en in te grasduinen. Als geheel is het ook een bijzondere tijdsreeks van metingen, in kwaliteit en kwantiteit vergelijkbaar met de metingen van het KNMI. Echter, dit zijn sociale in plaats van fysische gegevens. Van elke vergadering is een woordelijk verslag gemaakt, opgeslagen volgens een vrijwel onveranderd datamodel. Er is dus een enorme geordende reeks van goed vergelijkbare meetpunten, een ideale situatie voor historisch vergelijkend onderzoek.

Omdat de Handelingen nu digitaal beschikbaar zijn wordt het mogelijk om uiteenlopende vraagstukken met behulp van computers op te lossen. Bijzonder aantrekkelijk zijn conceptueel simpele en gemakkelijk te operationaliseren vragen die nu nog alleen met een enorme hoeveelheid manuren beantwoord zouden kunnen worden.

Drie voorbeelden:

- *Verruwing.* Het taalgebruik in de Kamer zou veel ruwer geworden zijn. Is dat echt zo? Welke partijen zijn nu en waren in het verleden verantwoordelijk voor dat ruwe taalgebruik, en wat voor verschuivingen zien we daar in?
- *Glazen Plafond.* Hoe verschilt de politieke loopbaan van vrouwelijke en mannelijke Kamerleden? Zijn ze evenveel aan het woord? Groeit hun carrière even snel? Is daar verandering in gekomen sinds het aantal vrouwen in de Kamer zo sterk is toegenomen?
- *Agenda setting.* Immigratie is een vaak terugkerend onderwerp geworden in de Kamer. Het wordt nu vooral als een sociaal probleem gezien terwijl het vroeger veel meer een als een economisch onderwerp beschouwd werd. Wanneer is dat nieuwe perspectief ontstaan? Wie begon ermee? In welke volgorde namen de andere partijen dat over?

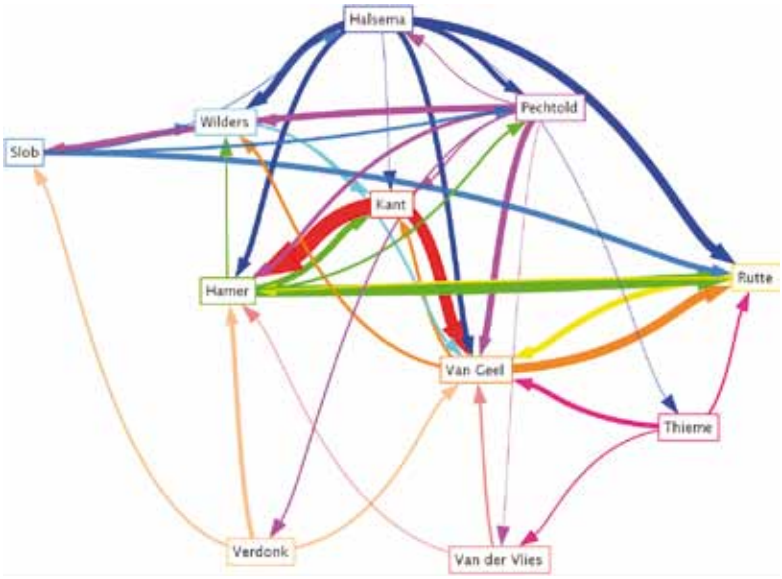
In al deze vragen spelen de tijd en de politieke actoren (politici en partijen) een cruciale rol: dat zijn de twee dimensies waarlangs de te onderzoeken variabele gevolgd wordt. De waarde op die variabele kan dan worden bepaald op basis van de tekst in de notulen. De manier waarop kan variëren van simpel woordjes tellen tot volledige zinsontleding. Hoe die waardes precies gevonden worden is voor dit artikel niet van belang. We kijken hier naar de noodzakelijke technische voorwaarden om die waardebepaling überhaupt machinaal uit kunnen te voeren.

Van document- naar persoonsgerichte ontsluiting

De Handelingen werden en worden geproduceerd om door mensen gelezen te worden en volgen een eenvoudig chronologisch datamodel. De nu digitaal beschikbare documenten wijken daar niet veel van af. Elk document bevat de notulen van één gehele dag, samen met waardes op een stuk of tien metadata velden (de datum,om welke kamer het gaat, etc.). Zowel de metadata als de tekst van de notulen is doorzoekbaar. Het antwoord op een zoekvraag is altijd een lijst documenten. We noemen dit *documentgericht* ontsluiten. De toegang tot de collectie wordt bepaald door de manier van opslaan, en er is slechts één wijze van toegang en één soort antwoord mogelijk (altijd een heel document).

Het kan ook anders. Wanneer de notulen nader bekeken worden, valt er meteen iets

op: van elk woord is bekend wie het gezegd heeft, van welke partij diegene is, in welke rol gesproken wordt, en zelfs vanaf welke plek het gezegd is (de spreekstoel, de regeringstafel, of de interruptiemicrofoon). Deze informatie is jammer genoeg alleen impliciet, verstopt in opmaak en conventies, aanwezig. Voordat je een computer kan vragen om 'alle zinnen gesproken door Ina Brouwer als CPN Kamerlid' zal die impliciete structuur eerst expliciet gemaakt moeten worden. En wat zou het leuk zijn als je die vraag verder kan vernauwen tot die woorden die ze richtte tot bijvoorbeeld Ruud Lubbers[1].
Dit is wat we bedoelen met *persoonsgerichte* ontsluiting. Dezelfde data kan nu opeens vanuit een heel ander perspectief bekeken worden. Het antwoord op een zoekvraag bij documentgericht ontsluiten is altijd een selectie van een of meer documenten. Het antwoord bij persoonsgericht ontsluiten is veel specifieker. We kunnen nu vragen naar de betogen van bepaalde politici, of de interrupties van politicus X op politicus Y. Als we eenmaal elk woord aan de juiste spreker hebben gekoppeld, en de sprekers aan een externe biografische database (bijvoorbeeld die van www.parlement.com), dan kunnen we beginnen met ons onderzoek.



Figuur 1: Visualisatie van de interrupties tijdens de Algemene Beschouwingen in de Tweede Kamer op 17 september 2008

Een mooi voorbeeld van wat dan mogelijk wordt is de zogenaamde *aanvalsgraaf* (Figuur 1) gemaakt door Rianne Kaptein[2]. De graaf geeft een overzicht van de interrupties tijdens één dag Algemene Beschouwingen. Deze graaf heeft de vorm van een sociaal netwerk. De knopen zijn de sprekers op die dag. Er loopt een pijl van A naar B, als A persoon B heeft geïnterrumpeerd. Hoe vaker A persoon B heeft onderbroken, hoe dikker de pijl. Let wel, op deze dag zijn er honderden sprekerswisselingen geweest. Het is mogelijk dit met de hand te maken, maar dat is een vervelende en arbeidsintensieve klus. Figuur 1 is volledig automatisch gegenereerd; hetzelfde algoritme kan toegepast worden op elke andere willekeurige vergadering.

In landen met een districtenstelsel is het natuurlijker om politieke informatie persoonsgericht te ontsluiten. De website www.theyworkforyou.com
doet dit op een zeer inzichtelijke manier voor het Verenigd Koninkrijk. Een voorbeeld dichter bij huis is www.pentapolitica.nl. Hier krijgt men voor elke Nederlandse politicus een overzicht van al zijn of haar sociale media uitingen (blogposts, Tweets, Youtube filmpjes, etc).

In een samenwerking tussen de Universiteit van Amsterdam, de Koninklijke Bibliotheek en de Dienst Informatievoorziening van de Tweede Kamer wordt deze transformatie naar persoonsgerichte ontsluiting van de notulen gemaakt en zal eind 2010 afgerond zijn. Daarmee wordt het beantwoorden van genoemde en andere onderzoeksvragen mogelijk, makkelijk, en hopelijk ook een stuk leuker om te doen.

Meer informatie over dit en andere projecten waarin politieke data makkelijk toegankelijk wordt gemaakt is te vinden op www.politicalmashup.nl.

[1] *Wikipedia vermeldt dat Lubbers haar wel eens mevrouw Bakker heeft genoemd en dat heeft moeten bekopen met een doos bonbons. Wat is er toen precies gezegd? (http://nl.wikipedia.org/wiki/Ina_Brouwer)*

[2] *Rianne Kaptein, Maarten Marx, Jaap Kamps: Who said what to whom?: Capturing the structure of debates. Proceedings SIGIR 2009: 831-832.*

Archieven Linken met Semantische Zoekmachines

In toenemende mate worden grootschalige archieven toegankelijk gemaakt voor een breed publiek. Prominente voorbeelden worden gegeven door de archieven van landelijke dagbladen, nationale archieven, overheidsarchieven, archieven onder beheer van de Koninklijke Bibliotheek, televisiearchieven zoals beheerd door het Nationaal Instituut voor Beeld en Geluid en, meer algemeen, door archieven van erfgoedinstellingen.

Maarten de Rijke, Krisztian Balog, Marc Bron, Jiyin He, Bouke Huurnink, Valentin Jijkoun, Fons Laan, Edgar Meij, Manos Tsagkias, Andrei Vishneuski, Wouter Weerkamp
Universiteit van Amsterdam

Een archief is geen eiland. Gebeurtenissen beschreven in een nieuwsarchief krijgen een extra dimensie als zij gekoppeld worden aan beeldmateriaal. Historisch televisiemateriaal wint aan betekenis als het gekoppeld wordt aan contemporaine commentaren en nieuwsmateriaal uit de gedrukte pers. En meer specialistische of technisch georiënteerde archieven winnen aan bruikbaarheid als ze gekoppeld zijn aan achtergrondinformatie.

Onderzoek wijst uit dat eindgebruikers er bij gebaat zijn als koppelingen tussen archieven betekenisvol zijn en bijvoorkeur langs semantische lijnen lopen, met een sterke oriëntatie op entiteiten (mensen, locaties, organisaties, artefacten, etc.), op thema's (zoals 'stadsleven,' 'festiviteiten' of 'consumentencultuur') en op gebeurtenissen (zoals 'Praagse lente,' 'Opening van de Kanaaltunnel' of 'Marathon Amsterdam'). Betekenisvolle ontsluiting van archieven komt hiermee neer op zoek- en verkenningstechnologiën rondom entiteiten, thema's en gebeurtenissen plus hun onderlinge relaties.

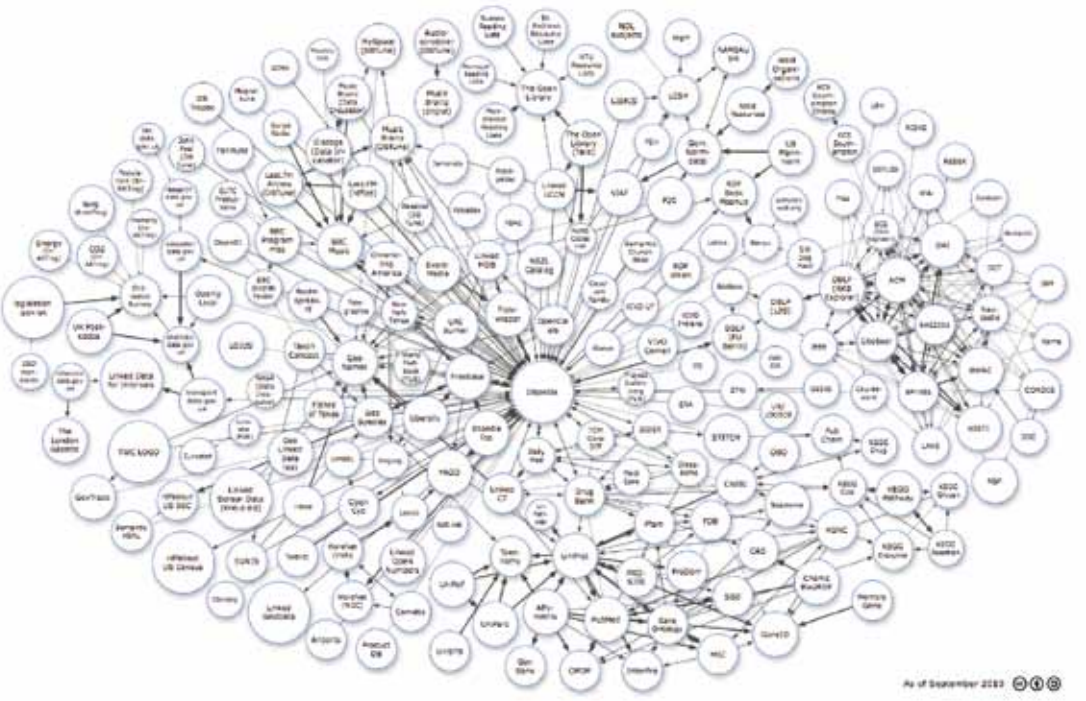
Gezien de omvang van de archieven die nu beschikbaar zijn of komen, zijn handmatige methoden om de gewenste koppelingen te leggen en om entiteiten, thema's en gebeurtenissen te identificeren in archiefobjecten eenvoudigweg niet realistisch. Een belangrijke beweging in onderzoek op het raakvlak van zoekmachinetechnologie en taaltechnologie betreft *semantisch* zoeken, waarbij de gewenste koppelingen tussen archieven langs de genoemde assen automatisch worden gelegd.

Onder de motorkap

Semantische zoekmachines stellen ons in staat om relevante entiteiten, thema's en gebeurtenissen te identificeren in grote hoeveelheden archief- of webdata. Dergelijke zoekmachines bouwen in belangrijke mate op taaltechnologie die erop gericht is om entiteiten thema's, gebeurtenissen en hun relaties in teksten te herkennen. Aan de Universiteit van Amsterdam werken we sinds 2008 aan een gedistribueerde omgeving, genaamd Fietstas, die de vereiste functionaliteit als webservice aanbiedt. Naast laag-niveau functionaliteit zoals tokeniseren en lemmatiseren biedt Fietstas semantische functionaliteiten zoals het herkennen van entiteiten en relaties, het normaliseren van entiteiten en het genereren van 'profielen' van entiteiten. Met name de laatste twee zijn interessant voor het koppelen van archieven. Het normaliseren van entiteiten betreft de volgende taak: welk object in de realiteit vormt de verwijzing van een voorkomen van een naam of beschrijving van een entiteit in een stuk tekst? Zie Figuur 1 voor een illustratie van het soort van documentannotaties dat daarbij door Fietstas gegenereerd wordt.

Parallel aan de genoemde taaltechnologie komen nieuwe algoritmes voor zoekmachi-

Figuur 1: Extractie van entiteiten.



Figuur 2: Verbonden data in de linked open data wolk.

nes op. De standaard informatie-eenheid (‘unit of retrieval’) verschuift van documenten naar scherper gedefinieerde eenheden, zoals entiteiten, netwerken van entiteiten, profielen van entiteiten, of relaties tussen entiteiten. Gebruikers van zoekmachines hebben meer nodig dan een platte lijst van links naar relevante documenten en moderne zoekmachines hebben de ambitie om hierin te voorzien middels *faceted search and exploration* waarin onderwerpsfacetten gecombineerd worden met facetten gebaseerd op entiteiten, thema’s, gebeurtenissen, locatie, tijd, genre, etc.

De twee - taaltechnologie en zoekmachinetechnologie met een nadruk op entiteiten en relaties - komen samen in een breed pallet aan taken. Bijvoorbeeld bij het automatisch genereren van hypertekstlinks, binnen Wikipedia of juist tussen specialistische documenten enerzijds en bronnen van achtergrondkennis andezijds, waarbij zowel op documentniveau als op het niveau van zoekvragen semantische structuur wordt ontdekt en benut.

Recente ontwikkelingen

Een belangrijke uitdaging voor thans vigerende algoritmes voor entity retrieval is dat zij zelden voor mensen interpreteerbare beschrijvingen weten te genereren van gevonden entiteiten of van de relaties tussen entiteiten. Linked Open Data (LOD) is een recente bijdrage van het opkomende semantische web dat de potentie heeft om de gewenste semantische informatie te leveren. De LOD-wolk bevat op dit moment miljoenen concepten uit honderden gestructureerde dataverzamelingen; zie Figuur 2. In lopend werk wordt onderzocht in hoeverre datagestuurde zoekmachine-algoritmes gecombineerd kunnen worden met gegevens uit de LOD-wolk. Voor het koppelen van archieven betekent dit werk een mogelijke verrijking van de te genereren koppelingen, zowel wat betreft hun aantal als wat betreft hun beschrijving en contextualisering.

Een tweede recente ontwikkeling die hier genoemd moet worden is het toepassen van deze koppeling van archieven in een dynamische context, bijvoorbeeld om te achterhalen hoe er in sociale media, zoals blogs, discussiefora en Twitter, gereageerd wordt op een nieuwsbericht. Een interessante uitdaging hierbij is het feit dat het taalgebruik in sociale media creatief is en niet onderworpen aan centrale redactie, waardoor te leggen koppelingen een fluïde karakter krijgen dat in hoge mate temporeel en contextueel bepaald is. Entiteiten spelen hierbij een sleutelrol, maar minstens zo belangrijk zijn de identiteit van deelnemers in online discussies en het gebruik van semantische hulpmiddelen zoals *hashtags*. Hiermee ontstaan koppelingen langs nieuwe - en voor een deel *sociale* - assen.

Conclusie

Met het steeds breder beschikbaar komen van grootschalige online archieven ontstaat

de noodzaak om deze aan elkaar en aan achtergrondkennis te koppelen. Gezien de omvang van dergelijke archieven zijn alleen automatische methoden een optie. Met behulp van moderne semantische zoekmachinetechnologiën zijn inmiddels de eerste stappen gezet om de gewenste koppelingen te genereren - van hoge kwaliteit en op grote schaal.

Dankwoord
Het type onderzoek dat hierboven beschreven wordt, is een onlosmakelijke combinatie van theoretisch, experimenteel en toegepast werk. Zonder data en use cases kan het niet uitgevoerd worden. We danken *NRC Handelsblad*, het *Nederlands Instituut voor Beeld en Geluid* en *Philips Medical Systems* voor discussies en het beschikbaar stellen van documentcollecties, zoekvragen en evaluatiemiddelen.

Het hier beschreven onderzoek wordt mede mogelijk gemaakt dankzij subsidies van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), onder projectnummers 612.066.512, 612.061.814, 612.061.815, 640.004.802, 380-70-011, het 7de Kaderprogramma van de EU, onder projectnummer 258191, het ICT Policy Support Programme van de EU, onder projectnummer 250430, het DuOMAn project dat wordt uitgevoerd binnen het STEVIN programma onder projectnummer STE-09-12, en door het Center for Creation, Content and Technology (CCCT).

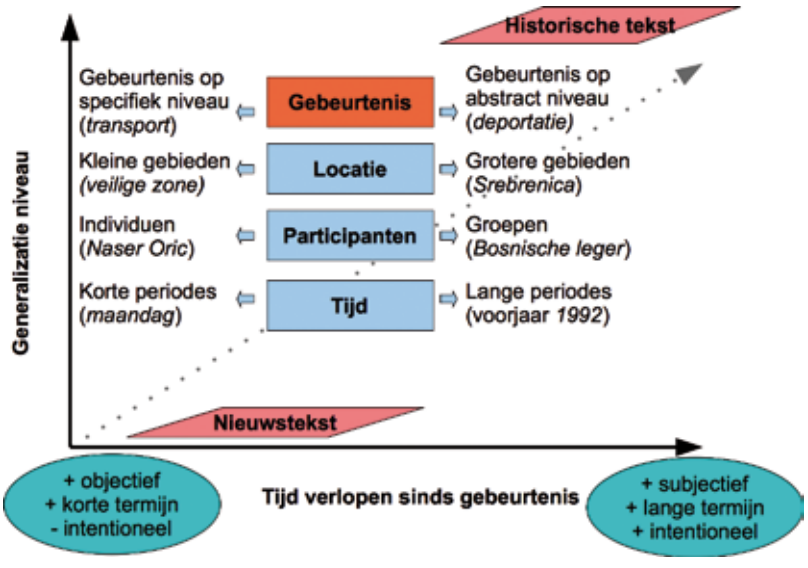
The Semantics of History
Grip op de beschrijving van historische gebeurtenissen

Geschiedenis is een retrospectieve beschrijving van veranderingen in de werkelijkheid. Historici richten zich daarbij vooral op veranderingen die van historisch belang zijn. Niet iedere gebeurtenis heeft immers historische waarde; een moord kan een historische gebeurtenis worden als blijkt dat deze moord belangrijke verschuivingen in het politieke landschap teweeg heeft gebracht. Daarbij is de interpretatie van een gebeurtenis afhankelijk van wie de gebeurtenis waarneemt of beschrijft. De één zijn werkelijkheid is niet die van de ander waardoor de geschiedenis vele malen wordt geherinterpreteerd en herschreven. We hebben daarom te maken met verschillende werkelijkheden, zowel in de periode direct na de gebeurtenis als lang daarna.

Agata Cybulska, Roxane Segers, Guus Schreiber en Piek Vossen
VU Amsterdam

De manier waarop een historische gebeurtenis wordt beschreven, zal navenant verschillen met de perceptie van de werkelijkheid en de historische verandering. Het project *Semantics Of History* heeft als doel om de tekstuele beschrijvingen van historische gebeurtenissen automatisch te analyseren en verschillen in die beschrijvingen op elkaar af te beelden. Daarbij spelen tijd en plaats een cruciale rol. Op basis van een gegeven moment in de tijd en een plaats kunnen alle beschreven gebeurtenissen met elkaar vergeleken worden. Gaat het om dezelfde gebeurtenis? Worden dezelfde participanten genoemd? Omvat de ene gebeurtenis de andere?

Srebrenica zijn gepubliceerd en de historische artikelen die later zijn geschreven. Uit deze vergelijking komt duidelijk naar voren dat naarmate de tijd verstrijkt, de gebeurtenissen anders worden beschreven, zie onderstaande schema. In krantenartikelen wordt veel gedetailleerde informatie gegeven over tijd, plaats en de participanten, terwijl historische beschrijvingen zich meer richten op hoofdlijnen. Verder blijkt duidelijk dat journalisten de gebeurtenissen niet meteen kunnen interpreteren. Naarmate de tijd verstrijkt, wordt meer bekend over wat er is gebeurd en worden intenties van de handelingen en waardeoordelen toegevoegd. Men spreekt in krantenartikelen bijvoorbeeld nog niet van *deportatie*



Zijn er oorzakelijke verbanden tussen gebeurtenissen aangegeven? Is er sprake van een waardeoordeel?

In een eerste studie hebben we een vergelijking gemaakt tussen nieuwsartikelen die direct na de gebeurtenissen rond de Val van

en *genocide* maar over *transport* en *neerschieten*. Het schema geeft aan hoe met de toename van tijdsperspectief beschrijvingen van gebeurtenissen aan de ene kant globaler worden (grotere gebeurtenissen, langere periodes, grotere gebieden, e.d.) en aan de andere kant meer subjectief en intentioneel.

De studie laat zien dat het nodig is om de 'framing' van een gebeurtenis in participant, tijd en plaats te kunnen herkennen zodat aan de ene kant alle relevante beschrijvingen kunnen worden gevonden en aan de andere kant het perspectief van die beschrijving kan worden aangegeven.

Het project gebruikt verschillende computersystemen en gegevensbestanden om de gebeurtenissen uit teksten te halen. In het project worden een historische ontologie, een lexicon en een gebeurtenismodel gebouwd waarmee historische gebeurtenissen kunnen worden verzameld, vergeleken en vervolgens met elkaar in verband kunnen worden gebracht. Het gebeurtenismodel beschrijft wat de minimale onderdelen zijn van een gebeurtenis en specificeert hoe deze onderdelen voor het computersysteem moeten worden beschreven en opgeslagen. De ontologie geeft een formele beschrijving van belangrijke concepten en de relaties tussen deze concepten in een domein. In een ontologie is bijvoorbeeld te vinden dat een Oorlog een soort Proces is met een begin- en eindpunt, dat Gevechtshandelingen een onderdeel zijn en dat een Burgeroorlog een soort oorlog is. In het historische lexicon staan alle woorden beschreven waarmee een taal deze concepten kan uitdrukken. Het lexicon en de ontologie worden vervolgens aan elkaar verbonden en kunnen dan worden gebruikt voor het automatisch annoteren van historische teksten. Met deze annotatie kunnen we verschillende soorten beschrijvingen van dezelfde gebeurtenis herkennen en analyseren. In het voorbeeld hieronder drukken de annotaties tussen rechte haken ontologische concepten uit die zijn toegekend aan beschrijvingen in de tekst. Ook al geven de fragmenten twee heel verschillende beschrijvingen van een gebeurtenis weer, door de annotatie weten we dat beide fragmenten betrekking hebben op dezelfde historische gebeurtenis of context:

- In mei 1992 [Time] heroverden [UnilateralGetting] Bosnische regeringstroepen [agent, ArmedForces] onder leiding van generaal Naser Oric [instantie van: Person] het stadje Srebrenica [instantie van:GeographicArea]
- Het Bosnische leger [agent, ArmedForces] nam al snel de controle over [UnilateralGetting] in de safe area [rol van: GeographicArea]

Binnen het project gaan we een formule definiëren voor de mate van overlap tussen historische beschrijvingen, binnen een plaats- en tijdsbegrenzing. Met dit soort annotaties kunnen we in de historische tekstcollecties gericht zoeken naar bijvoorbeeld alle beschrijvingen van legers die een locatie innemen. Om welk leger en welke locatie het precies gaat, is automatisch uit de tekst te halen. We vinden dan een hele lijst gebeurtenissen die we op basis van tijd en plaats kunnen groeperen en vervolgens als onderdeel van een grotere gebeurtenis als de Val van Srebrenica kunnen benoemen. Op die manier ontstaat een thesaurus van historische gebeurtenissen met daar aan gekoppeld de verschillende beschrijvingen en vindplaatsen van een gebeurtenis in de historische tekstcollectie. De thesaurus is geen platte lijst van historische gebeurtenissen, maar zal eruit komen te zien als een groot netwerk met tal van deel/heel- en causale relaties; De Val Srebrenica is onderdeel van de Joegoslavische Burgeroorlog en leidde indirect tot de val van het kabinet-Kok in 2002.

Semantics of History werkt samen met het CATCH project Agora dat een nieuw informatiesysteem ontwikkelt voor het Rijksmuseum en Beeld &Geluid. Met dit informatiesysteem kunnen gebruikers in een online platform zoeken naar objecten in de collectie van beide instituten op basis van historische gebeurtenissen die met de objecten geassocieerd zijn. *Semantics Of History* levert de kennis en data die door Agora gebruikt worden om die gebeurtenissen uit teksten te halen en aan de museumobjecten te koppelen.
Semantics Of History is een samenwerkingsproject van de Faculteit der Letteren en de Faculteit Exacte Wetenschappen aan de Vrije Universiteit Amsterdam, gefinancierd door het interfacultaire onderzoeksinstituut CAMeRA: www.camera.vu.nl.

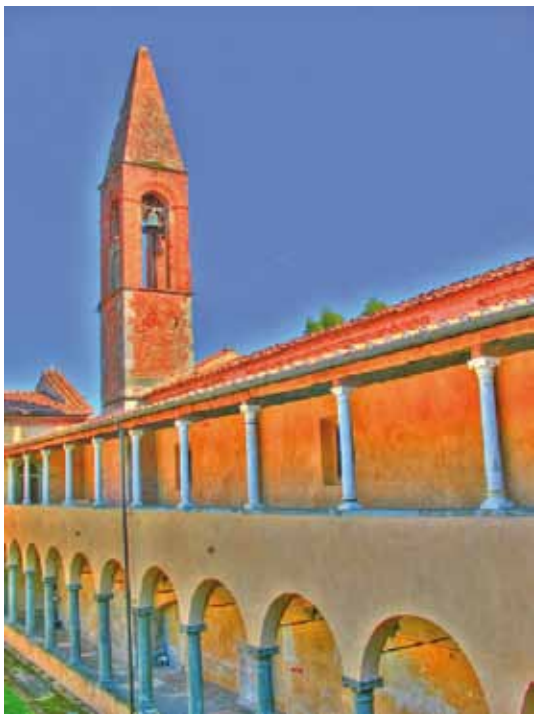
Het Rijksmuseum in Amsterdam, een van de gebruikers van het Agora-informatiesysteem.

Vele breinen maken licht werk: MediaEval brengt onderzoekers samen om multimediale erfgoed-collecties automatisch doorzoek-baar te maken

Martha Larson
Multimedia Information Retrieval Lab, TU Delft
Roeland Ordelman
HMI, Universiteit Twente

Achtergrond

MediaEval, is dat niet iets met Middeleeuwen, het ging toch over technologie? Ja inderdaad, maar hoewel 'medieval' en 'mediaeval' in het Engels refereren naar iets uit het verleden, gaat het hier om de toekomst. De toekomst van het zoeken in een schat aan multimediale erfgoed-collecties achter de muren van onze musea en archieven. 'MediaEval' is een benchmark-evaluatie die onderzoekers vanuit de hele wereld samenbrengt om na te denken over oplossingen op het gebied van multimedia-indexering en retrieval. Door praktisch met data collecties en gebruikersvragen aan de slag te gaan en onderling de resultaten te vergelijken, hopen de onderzoekers snel met praktische en bruikbare oplossingen te komen. In 2010 gebeurde dat op een wel zeer passende locatie: het middeleeuwse klooster *Santa Croce* in *Fossabanda* in Pisa, Italië.



Santa Croce in Fossabanda, Pisa, Italië

Multimediale bronnen vormen een belangrijk deel van ons cultureel erfgoed en een groot deel van deze bronnen bestaat uit audio- en video-opnamen die liggen opgeslagen bij de verschillende culturele erfgoedinstellingen zoals musea, archieven en documentatiecentra. Het Nederlands Instituut voor Beeld en Geluid bijvoorbeeld, herbergt meer dan 700.000 uur audiovisueel materiaal (dat is 80 jaar 24 uur per dag continue kijken/luisteren) en deze collectie neemt elk jaar met tienduizenden uren in omvang toe. Zonder adequate beschrijvingen zijn dit soort enorme collecties van weinig waarde. Digitaal beschikbare beschrijvingen -veelal aangeduid als 'metadata'- maken de collecties doorzoekbaar zodat materiaal daadwerkelijk bekeken, hergebruikt en/of wetenschappelijk onderzocht kan worden.

Metadata

Archivarissen hebben methoden ontwikkeld om archieven goed doorzoekbaar te maken door audiovisueel materiaal met behulp van menskracht te voorzien van hoogwaardige metadata. Deze manier van metadatering zal natuurlijk altijd belangrijk blijven, maar door de enorme hoeveelheden materiaal (zowel reeds bestaand als nieuw) lukt het eenvoudig niet om alles door mensen te laten metadateren. Bij de zoektocht naar informatie wordt steeds minder genoegen genomen met een globale indicatie waar informatie te vinden is; bij voorkeur wordt het meest geschikte fragment de gebruiker op een presenteerblaadje aangeboden (zie hier, van minuut 27 tot minuut 34 gaat het over Middeleeuwse Kloosters).

Technologie

Technologie kan archivarissen helpen bij het creëren van metadata. In MediaEval wordt er bijvoorbeeld gekeken naar het probleem om automatisch onderwerplabels (keywords) toe te kennen aan televisiemateriaal. Bij voorkeur komen die labels uit een met zorg samengestelde Thesaurus zoals de Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA), zoals in gebruik bij onder andere Beeld en Geluid en het Nationaal Archief. Voorbeelden van deze labels zijn 'Archeologie', 'Architectuur', 'Chemie', 'Dansen', 'Wetenschappelijk Onderzoek' en 'Beeldende kunst'. Omdat een Thesaurus honderdduizenden labels kan bevatten, zou het werk van de archivaris enorm versneld kunnen

worden wanneer de technologie een suggestie zou kunnen doen voor labels die het beste passen bij het materiaal. De archivaris kan dan snel de meest geschikte selecteren.

Label suggestie

Het automatisch suggereren van relevante labels gebeurt met een algoritme dat gebruik maakt van meerdere bronnen, zoals beschikbare metadata uit het productieproces, transcripties gegenereerd door middel van automatische spraakherkenning, en resultaten uit automatische analyse van het beeldmateriaal. Door statistiek in te zetten (tellen van woorden en 'co-occurence') en gebruik te maken van technieken voor 'machinaal leren' (machine learning) kunnen labels aardig goed voorspeld worden. Maar het gaat nog lang niet perfect en voor het verder verbeteren van de algoritmes is een aanhoudende gezamenlijke inspanning nodig van de onderzoekers in het veld.

Gelukkig bestaat er veel interesse onder onderzoekers om aan dit probleem van het automatisch genereren van labels te werken. Maar om iets zinnigs te kunnen bijdragen op dit terrein, moeten eerst de nodige hobbels genomen worden. Zo moeten de onderzoekers:

1. het probleem goed begrijpen en dus ook een idee hebben hoe en waarom onderwerp labels in een archief worden gebruikt;
2. de beschikking hebben over grote hoeveelheden voorbeelddata om algoritmes op uit te proberen en te testen;
3. de beschikking hebben over handmatig en/of automatisch gegenereerde bronnen voor metadata (niet elke onderzoeker heeft bijvoorbeeld een spraakherkenner tot zijn beschikking);
4. kennis hebben genomen van eerder werk op dit gebied opdat niet wordt geprobeerd het wiel opnieuw uit te vinden en ook niet om het eens met een vierkant wiel te proberen;
5. op hetzelfde testmateriaal de prestaties van hun eigen algoritmes kunnen vergelijken met de prestaties van anderen.

Het doel van een benchmark initiatief zoals MediaEval is:

- zoveel mogelijk van de bovengenoemde hobbels voor onderzoekers weg te nemen;
- een raamwerk te bieden die onderzoekers de mogelijkheid biedt zich volledig te concentreren op het probleem zelf en hun creativiteit optimaal te benutten voor het ontwikkelen van betere algoritmen.

In MediaEval gaat het om algoritmen voor het zoeken in multimedia, met speciale aandacht voor het combineren van informatie afkomstig uit spraak, tekst of van gebruikers (social tagging) met informatie uit visuele analyse (beeldherkenning) om, gegeven een zoekvraag, de relevante fragmenten te vinden.

Evaluatie

Een benchmarkevaluatie kent vaak verschillende 'tracks': taken die elk een specifiek doel nastreven waar het onderzoek dan op gericht is. Zo heeft MediaEval onder andere een 'labeling taak', een 'geo-tagging taak' (automatisch detecteren van een 'geografische label'), een 'affect taak' (automatisch detecteren van 'saaie stukjes video'!) , en een 'linking taak' ('het automatisch aan elkaar koppelen van informatie'). Een taak bestaat doorgaans uit drie onderdelen:

1. een beschrijving van het probleem dat moet worden opgelost;
2. een data set;
3. hulpmiddelen die kunnen worden gebruikt om het probleem op te lossen zoals spraaktranscripties (in MediaEval geleverd door Universiteit Twente), of video labels.

Het hele pakket van zo'n taak kan door een onderzoeker worden gedownload en gebruikt om nieuwe algoritmes te bedenken. De onderzoeker test de algoritmes op een oefenset ('development set' in onderzoeksjargon) en gebruikt de meest belovende algoritmes op een zogenaamde 'evaluatie set'. De resultaten op deze set worden naar de MediaEval organisatie gezonden die de 'score' van de verschillende deelnemers op de evaluatie set bepaalt.

Doordat alle onderzoekers dezelfde informatie hebben en dezelfde test set gebruiken kunnen de scores zinvol met elkaar worden vergeleken om te bepalen welke algoritmes het beste werken. Hoewel het voor het prestige van een onderzoeker of groep natuurlijk verre van verkeerd is om hoog te scoren, gaat het bij benchmarks vooral om de overdracht van informatie en de discussie tijdens een speciaal hiervoor georganiseerde workshop. Hier vertellen de individuele onderzoekers hoe hun algoritmes in elkaar zitten en waarom ze bepaalde keuzes hebben gemaakt. Door deze open competitie wordt samenwerking gestimuleerd en vermeden dat onderzoekers tijd verdoen met het ontwikkelen van algoritmen waarvan al is aangetoond dat ze niet goed werken.

De MediaEval 2010 workshop werd dit jaar in oktober gehouden in Pisa op een bijzonder mooie locatie: het middeleeuwse klooster met internetaansluiting 'Santa Croce in Fossabanda'. Een ideale plek om de liefde voor cultureel erfgoed te combineren met wetenschappelijke inspiratie om ons digitale culturele erfgoed toegankelijk te maken voor de toekomst.

Zo had het ook kunnen gaan
Graven in de geschiedenis van de sociale beweging met het HiTiME project

In een voormalig pakhuis in Zeeburg, Amsterdam, bewaart en onderzoekt het Internationaal Instituut voor Sociale Geschiedenis (IISG) de geschiedenis van de wereldwijde sociale beweging. Vakbonden, politieke organisaties en voormannen en -vrouwen deponeerden er hun archieven. Het IISG werkt actief aan de digitalisering van de eigen collectie. Verder denkend over nieuwe mogelijkheden om de eigen onderzoekers en het brede publiek zo ruim mogelijk toegang te verlenen, zocht het instituut aansluiting bij nieuwe ontwikkelingen in informatieontsluiting en text analytics—het overgrote deel van het materiaal van IISG is tekstueel—en vond die in het NWO-programma CATCH (Continuous Access to Cultural Heritage).

Antal van den Bosch
ILK / Tilburg centre for Cognition and Communication, Universiteit van Tilburg

Met de ILK onderzoeksgroep van het Tilburg center for Communication and Cognition, een onderzoekscentrum van de Universiteit van Tilburg, voert het IISG sinds 2009 het CATCH-project Historical Timeline Mining and Extraction (HiTiME) uit. Met taaltechnologische hulpmiddelen worden documenten van het IISG geanalyseerd op verschillende niveaus. Met named entity recognition technieken worden namen van personen, organisaties, en locaties gezocht, maar ook van beroepen, tijdsaanduidingen, en gebeurtenissen.

Het vinden van historisch significante gebeurtenissen in tekst is het meest ambitieuze doel van het HiTiME-project. Een historicus spreekt over een relevante gebeurtenis als die een merkbaar gevolg heeft gehad in de verdere loop van de geschiedenis. De letterlijke merkbaarheid van een gebeurtenis is bijvoorbeeld af te lezen aan de aandacht die eraan besteed wordt in kranten. Nieuws in dagbladen wordt door onderzoekers van sociale geschiedenis als een belangrijke bron van informatie gezien. Uit de beweging zelf kwamen toonaangevende kranten voort, zoals de socialistische krant Het Volk.

Stakingen die nooit gebeurden

Het IISG beschikt over een database met ruim 16.000 stakingen die in Nederland plaatsvonden. Volgens de samensteller, Sjaak van der Velden, mag de database compleet genoemd worden. Binnen het HiTiME-project werd een Master-scriptieonderzoek gewijd aan deze database. Martha van den Hoven, studente van de Tilburgse opleiding Human Aspects of Information Technology, combineerde de database van Van der Velden met het enorme digitale dagbladenarchief van de Koninklijke Bibliotheek in Den Haag. De beginveronderstelling van Van den Hoven was dat veel stakingen hun sporen trekken in de krant. Die veilige veronderstelling bleek te kloppen. Als op de site van de Koninklijke Bibliotheek werd gezocht met zoekopdrachten die woorden bevatten uit het record van een staking in de stakingdatabase, zoals de naam van het bedrijf en de plaats van de staking, en als er—uiteraard—werd gezocht rondom de periode dat de staking daadwerkelijk plaatsvond, dan bleek rond 50% van de gevonden artikelen inderdaad over de staking te gaan.
Van den Hoven verlegde haar aandacht vervolgens van de stakingen zelf naar de periode van onrust voorafgaand aan stakingen. Zo'n periode wordt vaak gekenmerkt door openlijke dreigementen, onderhandelingen, en protestacties die ook de krant halen. Met een kleine oogst van voorbeelden kon Van den Hoven met tekstclassificatie-technieken een classificatiemodule bouwen die in staat is een artikel te herkennen als een beschrijving van een periode van onrust, net als een spam filter een e-mail kan aanmerken als een ongewenst bericht.

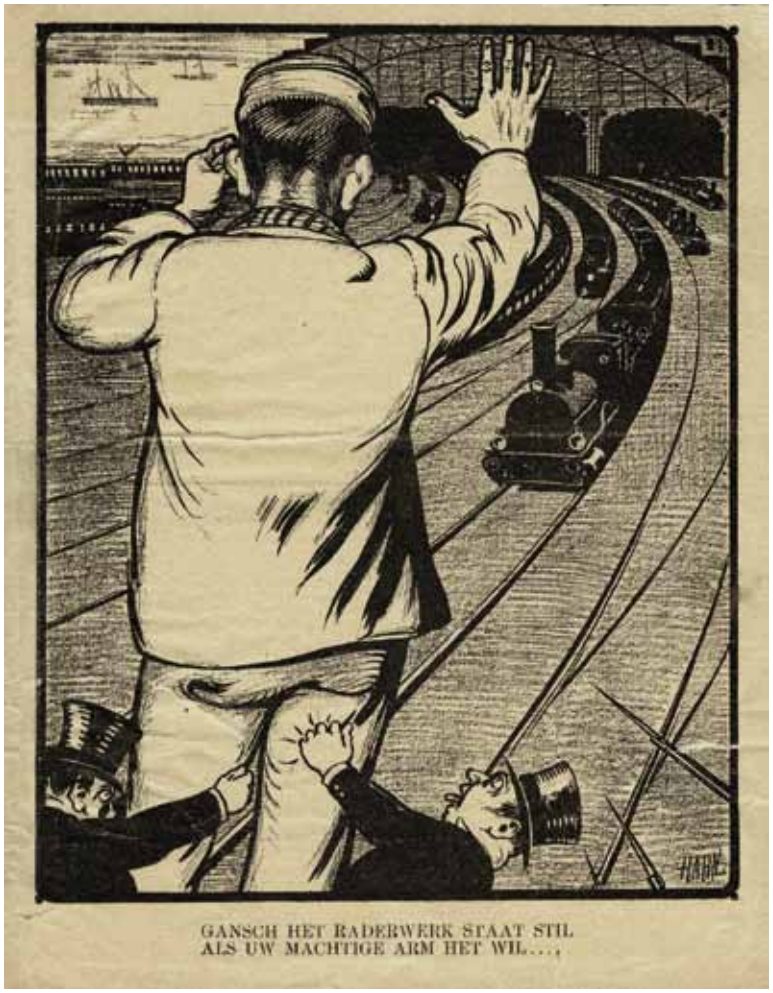
Met een eenvoudig automatisch toe te passen filter als zo'n classificatiemodule opent zich een verrassende mogelijkheid - als ditzelfde filter op de rest van de kranten van de Koninklijke Bibliotheek wordt losgelaten, zou het best eens periodes van onrust kunnen vinden die niet gevolgd werden door de stakingen waar ze op uit leken te draaien, omdat een bemiddelingspoging slaagde, of omdat de stakers in spé het niet eens werden over de te volgen acties. Noch Van der Velden noch andere onderzoekers van de sociale beweging hadden zich deze vraag

eerder gesteld: gegeven dat er stakingen zijn geweest die niet zijn doorgegaan, hoeveel zijn er dat eigenlijk geweest? En: zijn in dit soort processen patronen te ontdekken waaruit te verklaren is hoe sommige stakingsdreigingen in de kiem gesmoord werden? De scriptie van Van der Hoven beschrijft een begin van een automatisch systeem dat deze vraag deels automatisch zou kunnen beantwoorden.

Geschiedkundig onderzoek naar de ontwikkeling van gebeurtenissen in de tijd die niet plaatsvonden maar wel gebeurd hadden kunnen zijn, wordt wel counterfactual history of what-if history genoemd. Geautomatiseerd counterfactual history onderzoek blijkt, volgens de scriptie, met redelijke precisie mogelijk te zijn door het combineren van zorgvuldig samengestelde databases en massaal gedigitaliseerde primaire bronnen zoals de Koninklijke Bibliotheek die geheel vrij aanbiedt. Geschiedkundigen, die nog altijd de cruciale expertrol hebben en de rol van interpretator van de ruwe verzamelde gegevens, kunnen zich dit soort nieuwe vragen nu stellen, en besparen met de komst van deze technieken vele uren doorspitwerk in papieren archieven. De reacties op de studie van Martha van der Hoven waren dan ook enthousiast.

Vriend en vijand

Het HiTiME-project poogt in de vier jaar dat het loopt een aantal van deze automatiseringstrajecten te kunnen ontwikkelen, en tegelijk een kennisbank en -netwerk op te bouwen van belangrijke mensen, organisaties en gebeurtenissen en hun onderlinge relaties, gekoppeld aan de tijdbalk. Visualisaties van tijdbalken, sociale netwerken en geografische informatie zullen gebruikt worden om de resultaten zichtbaar te maken. Een van de lopende studies, van promovenda Matje van de Camp, richt zich op de mogelijkheid om sociale netwerken af te leiden uit biografische teksten en persoonlijke correspondentie, en daarin ook de vriend- en vijandrelaties te herkennen—en de veranderingen daarin. Een ander studieobject is de ontwikkeling van organisatiestructuren binnen de sociale beweging. Door met automatische middelen grotere patronen tussen mensen en organisaties bloot te leggen in teksten poogt HiTiME de geschiedkunde een nieuwe gereedschapskist te geven die zo dicht mogelijk tegen het ambitieniveau van de echte onderzoeksvragen aan zit.



Collectie Internationaal Instituut voor Sociale Geschiedenis, Amsterdam

Erfgoed van een Oorlog

Oorlog is een van de meest ingrijpende gebeurtenissen in een mensenleven. Angst, onzekerheid, dood en verdriet aan de ene kant en saamhorigheid, intensiteit, hoop en vreugde aan de andere kant. In veel (gesproken) getuigenissen komen deze twee facetten van een oorlogssituatie duidelijk naar voren ('De mensen waren toen nog aardig en iedereen hielp elkaar, iets wat na de oorlog snel verdween: toen was het weer ieder voor zich').

Arjan van Hessen
Universiteit Twente/ Telecats,
Michel Boedeltje
Telecats,
Petra Links en Ruben de Vos
NIOD

Tijd heeft een enorme impact op herinneringen van mensen en het kan goed zijn dat, met het verstrijken van de tijd, zowel positieve als negatieve indrukken duidelijk de boventoon gaan voeren in de herinneringen aan de oorlogssituatie. Het is daarom interessant om te zien hoe mensen zich de situatie vlak na de oorlog herinneren en hoe dat een groot aantal jaren later wellicht veranderd is. De officiële, grote geschiedenis wordt dikwijls door 'de staat' en/of 'autoriteiten' geschreven. Persoonlijke herinneringen van mensen aan 'hun oorlog' kunnen echter een enorme verrijking en nuancering betekenen van het algemene, officiële beeld en geven beter dan wat ook inzicht in de 'emotionele staat' van de betrokkenen. Oral History (het opnemen, bewaren en interpreteren van historische informatie, gebaseerd op de persoonlijke ervaringen en meningen van de spreker) is een bloeiende tak van wetenschap, mede door de dalende prijzen en stijgende geluidskwaliteit van de opnamen en de toegenomen eenvoud om de opnamen voor iedereen beschikbaar te stellen.

In dit artikel zullen we eerst kort de twee gebruikte vormen van spraakherkenning behandelen en vervolgens drie verschillende en elkaar opvolgende Oral History projecten. Ten eerste het CATCH-project CHORAL. Vervolgens bespreken we het Getuigen Verhalen project, onderdeel van het grote VWS programma 'Erfgoed van de Oorlog'. Tenslotte zullen we aandacht schenken aan het pas gestart CROME-project, waar de in Nederland ontwikkelde spraak- en ontsluitingstechnologie wordt ingezet in het nader tot elkaar brengen van de verschillende bevolkingsgroepen in de voormalige Joegoslavische deelrepubliek Kroatië.

Taal- en Spraaktechnologie (TST)

Met behulp van inmiddels bekende en veelgebruikte technieken om teksten doorzoekbaar te maken is het eenvoudig om in de transcripties van AV-materiaal te zoeken. Echter, met de huidige mogelijkheden en de trend om materiaal multimediaal op te slaan en te presenteren, zal een groot deel van de gebruikers de gezochte interviews direct willen beluisteren en/of bekijken zodat niet alleen duidelijk wordt wat er gezegd werd maar ook hoe! Men wil als het ware in de interviews zelf kunnen zoeken en beleven hoe een en ander gezegd werd. Moderne Taal- en Spraaktechnologie maakt dit mogelijk zoals hierna zal worden uitgelegd.

Herkenning Taalmodel

Het klinkt wellicht tegenstrijdig, maar om spraak goed te kunnen herkennen (zowel door de mens als de computer), moet de spraak zo voorspelbaar mogelijk zijn. Mensen die een taal slecht beheersen en dus veel minder makkelijk kunnen voorspellen welke woorden eventueel gezegd zullen worden, gegeven de al gesproken woorden, zullen a) deze taal als veel sneller ervaren en b) veel meer moeite hebben om de spraak te herkennen. Een beetje achtergrondruis of muziek lijdt dan direct tot een sterke afname van de herkenning. Dit is voor computers niet anders en we kunnen de huidige software het best vergelijken met een brugklasser die net z'n eerste lessen Frans heeft gehad. Hoe meer voorbeelden hij zal krijgen, hoe beter hij een zin zal kunnen afmaken en dus hoe beter z'n Frans zal worden.

Om de computer te laten leren welke woorden gesproken kunnen worden, worden grote hoeveelheden tekst die over dezelfde onderwerpen gaan als de spraak die herkend moet gaan worden, aan de computer aangeboden. Die leert als het ware de mogelijke woorden en de volgorde waarin deze woorden gesproken kunnen gaan worden. Hoe beter vervolgens de spraak 'past' op deze teksten, hoe beter de spraak herkend zal worden. Het resultaat van het leren van de computer is een statistisch taalmodel waarin de kans op een woord en een woordcombinatie ligt opgeslagen. Wanneer er dus interviews herkend moeten gaan worden die over de NSB, concentratiekampen, joden en Dolle Dinsdag gaan, dan moeten zoveel mogelijke teksten die hierover gaan, aan de computer worden gegeven. Hiermee moet men dan niet interviews gaan herkennen die over de aanval op Pearl Harbor of over de

Japanse bezetting van Indonesië gaan omdat hier andere woorden en woordcombinaties gebruikt zullen worden!

Spraakherkenning

Ondank al deze maatregelen, zal voorlopig de herkenning niet veel beter zijn dan 60%: dat wil zeggen dat 4 op de 10 woorden fout herkend zullen worden. Toch laten de toepassingen zien dat dit in de regel voldoende is voor het zoeken in de interviews. Wil men spraakherkenning gebruiken voor het ondertitelen dan moet men op minimaal 95% correcte herkenning zitten; iets dat anno 2010 met gewone, spontane spraak nog niet goed mogelijk is.

Van spraakherkenning naar oplijning

In veel gevallen is de spraak volledig uitgeschreven, zodat er gebruik gemaakt kan worden van oplijning: spraakherkenning waarbij vooraf bijna precies bekend is wat er gezegd gaat worden. Er is hier dus sprake van een optimale voorspelling en het enige dat de software eigenlijk moet doen is de woorden op de juiste plek (= tijd) zetten. Het resultaat is een bijna 100% correct 'herkende' tekst, zelf als binnen een zin de woordvolgorde in de opname anders is dan de woordvolgorde in de uitgeschreven tekst (of wanneer er delen van de zin missen). Dit resultaat maakt het mogelijk om de resultaten van de oplijner behalve voor het zoeken in de audiovisuele bestanden ook voor het ondertitelen ervan te gebruiken.

In de hierna te bespreken projecten wordt gebruik gemaakt van zowel oplijning als spraakherkenning.

CHORAL
Radio Oranje

In 2002 leidde een bezoek van de HMI-groep van de Universiteit Twente aan het NIOD in Amsterdam tot een initiatief om met behulp van de in Twente ontwikkelde spraaktechnologie de historische geluidsopnamen en de vooraf uitgeschreven teksten van de toespraken van Koningin Wilhelmina voor Radio Oranje (Londen, 1940-1944) te synchroniseren en via het internet beschikbaar te stellen.

Brandgrens

Na de succesvolle lancering van deze website werd besloten het Radio Oranje Project onder te brengen bij het pas gestarte CHATCH-project CHORAL (2005-2010). In het CHORAL project werd 'echte' spraakherkenning ingezet om de gesproken collecties van Radio Rijnmond te kunnen ontsluiten. Een mooi resultaat hiervan is de website over de brandgrens: de grens van het in de meidagen van '40 platgegooide deel van Rotterdam. 16 ooggetuigen vertellen over hun belevenissen voor, tijdens en na het bombardement. Alle interviews zijn met behulp van spraakherkenning doorzoekbaar gemaakt. Naast deze projecten heeft de Universiteit Twente ook meegedaan aan het ontsluiten van:

- 39 interviews met overleveraars van Buchenwald;
- 9 interviews met vrouwelijke hulpverleners uit de begindagen van de tweede feministische golf;
- De gesproken versie van het dagboek van Anne Frank.

Uiteindelijk besloot de HMI-groep, ondanks de nog steeds groeiende vraag, te stoppen met het ontsluiten van nog meer Nederlandse OH-collecties. Om de kennis niet verloren te laten gaan, werd besloten de activiteiten onder te brengen in het project Verteld Verleden: een samenwerking van verschillende Culturele Erfgoed instellingen, Beeld en Geluid en de Universiteit Twente. Ook werd de kennis beschikbaar gesteld aan verschillende MKB-bedrijven.

Getuigen Verhalen

Waarom?

Er zijn steeds minder mensen die zelf (bewust) de oorlog hebben meegemaakt: iemand die tien jaar was toen de oorlog begon, is nu al 80. Willen we de verhalen van deze mensen voor het nageslacht bewaren, dan moet er haast gemaakt worden met het optekenen. Dit is de reden geweest waarom het ministerie van VWS besloot geld beschikbaar te stellen voor het vastleggen van een groot aantal getuigenissen en waar mogelijk op het internet te publiceren. De Oral History-poot van het programma is bekend als het project Getuigen Verhalen en omvat 46 Oral History projecten. Begin 2010 vroeg het ministerie van VWS, het NIOD (Nederlands Instituut voor Oorlogsdocumentatie) de website Getuigen Verhalen op te zetten en te beheren: sinds 16 september 2010 is deze website actief.

Ontsluiting

Een vereiste van de subsidiegever was dat de opgenomen interviews door medewerkers

van de verschillende projecten zo 'letterlijk mogelijk' zouden worden uitgeschreven; deze teksten zouden dan samen met de interviews in het archief moeten worden opgenomen. Hierdoor was de tekst van de interviews beschikbaar en kon er worden opgelijnd. Omdat de Universiteit Twente dit soort grote projecten niet meer wilde doen (er was immers geen sprake meer van vernieuwend onderzoek), werd Telecats gevraagd de spraak op te lijnen. Op moment van schrijven is men hard bezig de meer dan 200 uur AV-materiaal op te lijnen. Alle eenmaal opgelijnde opnamen kunnen vervolgens op woordniveau doorzocht worden.

Na het Getuigen Verhalen project (eind 2010) zal deze oplijningstechnologie gebruikt gaan worden voor het op dezelfde wijze ontsluiten van 150 uur interviews over de late gevolgen van Sobibor.

Documenta

Het ministerie van Buitenlandse zaken stelt via het MATRA-programma (MAatschappelijke TRAnsformatie) geld beschikbaar voor projecten in landen die op de drempel staan om lid te worden van de Europese Gemeenschap. Het doel van zo'n MATRA-project is het vergemakkelijken van de overgang naar volwaardig EU-lid. Een van de landen op de drempel is Kroatië: een land dat van 1990 tot 1995 verwikkeld was in een hevige oorlog/burgeroorlog waarbij de verschillende eens zo vredig naast elkaar levende bevolkingsgroepen (Kroaten, Serviërs en Bosniërs) elkaar naar het leven stonden. Omdat na het beëindigen van de oorlogshandelingen deze bevolkingsgroepen nog steeds met elkaar door een deur moesten (niet alle Serviërs en Bosniërs waren verdreven) vatte de Kroatische mensenrechten organisatie Documenta het plan op om een grote collectie audiovisuele interviews op te nemen en met behulp van moderne taal- en spraaktechnologie te ontsluiten. Het plan lijkt dus erg op de andere twee beschreven projecten, zij het dat de Kroatische TST veel minder ontwikkeld is dan de Nederlandse. Bovendien wilde men de opgenomen interviews voor de hele wereld beschikbaar stellen. Daarom moeten de interviews dus ook vertaald worden, waarbij de synchronisatie met de oorspronkelijke spraak behouden moet blijven.

Het project, een samenwerking van de Erasmus Universiteit, de Universiteit Twente, DANS, Documenta en het Amsterdamse bedrijf Noterik, behelst het opnemen van 500 interviews met een dwarsdoorsnede van de Kroatische bevolking, het uitschrijven, oplijnen en vertalen van de interviews en het ontsluiten ervan via een website. Anders dan de oude Oral History projecten zoals hierboven beschreven, is dit project wetenschappelijk wel interessant omdat er allerlei nieuwe zaken spelen zoals meertaligheid, het ontbreken van een goede Kroatische spraakherkenner en het online toevoegen van additionele documenten en metadata. Bovendien zal onderzocht worden in hoeverre fragmenten van de verschillende interviews (semi-)automatisch aan elkaar gekoppeld kunnen worden zodat er virtuele collecties kunnen ontstaan.

Gevoeligheid

Doordat de gebeurtenissen in Kroatië nog vers in het geheugen liggen, is de emotionele lading bij de meeste interviews een stuk hoger dan bij de WOII verhalen in Nederland. De kans op serieuze uitglijers (het (ongefundeerd) beschuldigen van anderen, het verklappen van geheimen van nog levende personen) is dan ook een stuk groter. Al het materiaal moet beoordeeld worden op de mogelijke gevolgen die het publiceren via internet kan hebben voor zowel de geïnterviewden als voor anderen. Gevoelige stukken moeten onder embargo gezet worden, maar niet zomaar worden verwijderd. Dit gebeurt trouwens ook bij delen van de getuigenverhalen-collectie (bijvoorbeeld de interviews met dochters van NSB-vrouwen). Al het oorspronkelijke materiaal zal worden opgeslagen bij DANS in Den Haag en alleen de 'goedgekeurde' onderdelen van de interviews zullen via de website openbaar gemaakt worden.

Conclusie

Dankzij de inzet van Taal- en Spraaktechnologie is het nu mogelijk om niet alleen in de handmatig toegevoegde metadata, maar ook direct in de spraak van de interviews te kunnen zoeken. Eenmaal gevonden fragmenten kunnen bovendien direct beluisterd/bekeken worden. Dit zal de collecties naar verwachting een stuk interessanter maken voor zowel wetenschappelijke onderzoekers als voor geïnteresseerde leken. Deze nieuwe manier van ontsluiten vormt enerzijds een belangrijke toevoeging aan het levend houden van de getuigenissen van de oorlog en anderzijds een goed middel voor het kweken van begrip voor de 'ander'.

Audiovisueel erfgoed doorzoekbaar met spraakherkenning

Willem Melder & Roeland Ordelman
Nederlands Instituut voor Beeld en Geluid

Audiovisueel erfgoed

Beeld en geluid, zowel via radio, televisie maar ook steeds meer via het internet, speelt al decennia een belangrijke rol in de samenleving. Het Nederlands cultureel erfgoed is niet compleet zonder audiovisueel materiaal van de kroning van Beatrix, optredens van Nederlandse artiesten op Pinkpop of achtergrond reportages op Radio 1. Het Nederlands Instituut voor Beeld en Geluid in Hilversum archiveert het Nederlands audiovisueel erfgoed, voorziet het van beschrijvingen en ontsluit het materiaal voor de huidige en toekomstige generaties. Dagelijks stromen er vele uren aan radio en televisie van de publieke omroepen binnen in het archief, sinds 2006 volledig digitaal. Daarnaast beheert Beeld en Geluid collecties uit privébezit, Nederlandse films en specifieke collecties, zoals de gedurende tien jaar uitgezonden dagelijkse talkshow met Barend en van Dorp.

Bewaren, beschrijven, hergebruiken

In de depots van Beeld en Geluid liggen vele soorten dragers geconserveerd. De houdbaarheid hiervan is beperkt en daarom wordt een groot gedeelte in het Beelden voor de Toekomst project omgezet naar digitale formaten. Gedurende Beelden voor de Toekomst worden in totaal 137.200 uur video, 22.510 uur film, 123.900 uur audio en 2,9 miljoen foto's uit de archieven van de deelnemende instellingen gerestaureerd, geconserveerd, gedigitaliseerd en op verschillende manieren aan een breed publiek gepresenteerd.

Doorzoekbaarheid verbeterd met spraakherkenning

Spraakherkenningtechnologie wordt gekoppeld aan het Beeld en Geluid-archief, om audiovisuele collecties die beperkt handmatig beschreven zijn, automatisch te voorzien van beschrijvingen (metadata) die gebaseerd zijn op het gesproken woord. Dit is erg handig voor materiaal uit de depots waar geen enkele informatie over beschikbaar is. Mits het geluid aan de minimale kwaliteitseisen voldoet kan dit materiaal met spraakherkenning ontsloten worden en doorzoekbaar worden gemaakt. Wie weet welke pareltjes er op deze manier nog gevonden gaan worden?

Spraakherkenning kent ook zijn beperkingen en de transcriptie van de spraak zal niet altijd nauwkeurig zijn. Voor zoekdoeleinden is dat helemaal niet erg, mits de belangrijke woorden -zoals namen of specifieke termen- maar wel goed herkend worden. Voorwaarde is dan wel dat de spraakherkenner deze woorden kent: er kunnen geen woorden herkend worden die niet in het vocabulaire staan. Er komen voortdurend nieuwe (eigen) namen bij (nieuwe politici, bedrijven fuseren onder nieuwe naam), er ontstaan nieuwe woorden (zoals 'Damschreeuwer', 'hypotheekleed', 'kopvoddentaks' of 'gedoogakkoord') of er worden woorden geleend uit andere talen ('vuvuzela', 'boerka'). Het is nog een hele klus om het vocabulaire van de herkenner up-to-date te houden. Gelukkig kan hierbij gebruikt gemaakt worden van diverse beschikbare tekstuele bronnen, zoals meegeleverde programma-informatie en online nieuwsbronnen.

Radio 1

Door gebruik te maken van ondertiteling wordt het binnenkort mogelijk om in het archief te zoeken op gesproken woord in televisieprogramma's van de publieke omroepen. Omdat deze informatie voor radioprogramma's ontbreekt is er met spraakherkenning allereerst getest op Radio 1 programma's, zoals 'Dit is de dag', 'Langs de lijn' en 'Goedemorgen Nederland'. Zowel de automatische transcripties als de ondertitels zijn voorzien van tijdcodes, waardoor het mogelijk wordt om naar fragmenten te zoeken als naar de speld in de audiovisuele hooiberg. Media professionals, onderzoekers en het algemeen publiek kunnen gericht zoeken naar het fragment (het precieze tijdstip in de uitzending), bijvoorbeeld dat fragment waar Balkenende sprak over 'de VOC mentaliteit'.

Interviews, geknipt voor onderzoek

Onderzoekers in de letteren en in de sociale wetenschappen hebben vaak een schat aan materiaal verzameld waarvan de wereld enkel de uiteindelijke publicatie ziet. Collega-onderzoekers, maar ook het bredere publiek, zijn echter ook geïnteresseerd in de ruwe onderzoeksdata (bijvoorbeeld de opgenomen interviews) en de onderzoeksgegevens (de aantekeningen van onderzoekers bij die data). De huidige stand van zaken in de ICT biedt geweldige nieuwe mogelijkheden om via de publicatie ook de data beschikbaar te maken.

Henk van den Heuvel
CLST,
Radboud Universiteit Nijmegen

Veteran Tapes

In het project Veteran Tapes[1] is een groep onderzoekers uit verschillende disciplines met dezelfde collectie interviews aan de slag gegaan. Ieder vanuit zijn (of haar) eigen invalshoek en met zijn eigen onderzoeksvragen. Dat was nieuw voor velen van hen, want normaal gesproken verzamelen ze hun eigen data. Nu moest er relevante informatie uit een bestaande datacollectie worden gebruikt.
De collectie waar we het over hebben zijn 25 interviews die door het Veteraneninstituut zijn opgenomen. Deze maken deel uit van een veel grotere collectie van zo'n 1000 interviews. De interviews zijn afgenomen met veteranen die betrokken waren bij de Tweede Wereldoorlog tot en met de missie naar Uruzgan. Elk interview duurt gemiddeld zo'n 2,5 uur, en is geheel digitaal beschikbaar. Van de genoemde 25 interviews waren ook transcripties beschikbaar. Alle interviews zijn gearchiveerd bij DANS volgens het Datakeurmerk waarbij standaard persistent identifiers worden gebruikt[2] .

De onderzoekers hebben samen zeven artikelen geschreven gebaseerd op de informatie die zij in de 25 interviews (of een deel ervan) aantroffen. Deze artikelen zijn opgenomen in een recent verschenen boek[3].

Verrijkte publicaties

Daarnaast en daarna hebben een aantal van deze onderzoekers hun artikel uitgewerkt tot een zogenaamde verrijkte publicatie[4]. In dit geval betekent dat, dat dezelfde publicatie op internet beschikbaar komt, voorzien van allerlei links naar fragmenten uit de interviews. Deze links zijn gekoppeld aan een citaat dat voorkomt in een interview. Wanneer de link in de publicatie wordt aangeklikt wordt de lezer naar een aparte webpagina geleid waar ditzelfde citaat wordt getoond voorzien van allerlei extra informatie (meta-data) over het interview en een directe koppeling naar de audio van het betreffende fragment. De verrijkte publicatie komt beschikbaar op www.watveteranenvertellen.nl. Wat Veteranen Vertellen is ook de titel van het boek dat op 27 oktober verschenen is. In de figuur zien we links de pdf-weergave van de publicatie. Op de pagina staan twee citaten waarvan de kopregel aanklikbaar is. De lezer komt dan op de webpagina terecht die rechts van de pdf-pagina staat.

De fragmentknipper

Dergelijke fragmenten moeten natuurlijk gemaakt worden voordat ze beschikbaar kunnen worden gesteld. Hiertoe is door CLST een webapplicatie gemaakt die de onderzoeker in staat stelt om deze fragmenten gemakkelijk te selecteren en te knippen. De tool werkt stapsgewijs als volgt:

1. Ga naar de tijdcode van het gewenste interviewfragment
2. Beluister de audio van het segment
3. Wijzig de transcriptie, indien nodig (soms is anonimisering van persoonlijke gegevens op zijn plaats)
4. Voeg indien gewenst een aantekening toe
5. Selecteer de gewenste metadata
6. Controleer het fragment
7. Link het fragment aan de verrijkte publicatie

Van fragmentknipper naar annotatietool
In een vervolgproject[5] zijn er 250 interviews geselecteerd die de thema's WO II, Nederlands Indië en Nieuw-Guinea beslaan. Deze interviews werden ontsloten door middel van automatische spraakherkenning. Dit gebeurde op een andere manier dan voor de Getuigenverhalen (zie elders in dit nummer). Omdat er voor deze interviews geen uitgeschreven transcripties beschikbaar waren is er door een goede bestaande spraakherkenner een transcriptie van de audiofiles gemaakt[6]. Deze herkenner was getuned op het taalgebruik in de verschillende soorten interviews waarbij een goede herkenning van de sleutelwoorden

voorop stond. Deze sleutelwoorden waren via het Veteraneninstituut voor elk interview beschikbaar evenals een samenvatting per 10 minuten. Verder was de herkenner getuned op specifieke uitspraakverschijnselen bij de geïnterviewde door middel van sprekeradaptatie van de akoestische modellen. Om de interviews doorzoekbaar te maken werd de fragmentknipper uitgebouwd tot een annotatietool. Deze annotatietool heeft de volgende functionaliteiten:

- Inlogfunctie
- Gebruikersadministratie
- In interviews zoeken op:
  - Tijdcodes
  - Woorden (inclusief wildcards)
- Gevonden fragmenten verschijnen in lijst
- Per fragment is de metadata beschikbaar
- Gebruiker kan:
  - Grenzen van het fragment aanpassen
  - Annotaties en transcripties toevoegen
  - Files bij annotatie uploaden
  - Tabbladen raadplegen: mijn/alle transcripties/annotaties

Door de annotatietool zijn de ruwe data (de interviews) en de aanvullende gegevens (annotaties) van de onderzoekers beschikbaar gekomen voor de onderzoeksgemeenschap. Hiermee is in principe de cirkel gesloten van de in de aanhef genoemde elementen (ruwe data, onderzoeksgegevens, verrijkte publicaties).

[1] *www.dans.knaw.nl/content/categorieen/projecten/veteran-tapes-veteran-tapes-verrijkte-publicatie*
[2] *'Dataseal of Approval' (www.datasealofapproval.org)*
[3] *Van den Berg, Scagliola, Wester (eds): Wat Veteranen vertellen. Pallas Publications. AUP. Oktober 2010.*
[4] *Dit project is gesubsidieerd door de Surf Foundation: www.surffoundation.nl/nl/projecten/Pages/Veteran-Tapes-VP-Verrijkte-publicatie-op-basis-van-multidisciplinair-hergebruik-van-kwalitatieve-onderzoeksbestanden.aspx*
[5] *Het project Living Oral History Workbench, gefinancierd door het ministerie van VWS in het programma Erfgoed van de Oorlog. Zie www.rijksoverheid.nl/onderwerpen/tweede-wereldoorlog/projecten-erfgoed-van-de-oorlog/behouden-en-digitaliseren-collecties*
[6] *Het gaat hier dus niet om een oplijning met bestaande transcripties, maar om een echte spraakherkenning van nog niet getranscribeerde tekst. In de tekst van Arjan van Hessen over Getuigenverhalen is dit uitstekend beschreven onder het kopje Herkenning.*

NOTaS-Agenda

7 en 8 december 2010:
Digitaal Erfgoed-conferentie 2010
Deze 6e Digitaal Erfgoedconferentie vindt plaats in De Doelen in Rotterdam, en heeft als thema de digitale ontwikkelingen rond erfgoed en educatie, van onderwijs tot publieksbegeleiding en van kennisoverdracht tot `lifelong learning'. Cultureel erfgoed speelt een belangrijke rol in onderwijs. Het onderwijs is dan ook een van belangrijke afnemers van diensten van erfgoedinstellingen, variërend van bezoekersprogramma's tot 'serious games'. In een gevarieerd en interactief programma van lezingen, debatten en workshops passeren goede en minder goede ervaringen, beleidsdoelen en nieuwe technologieën de revue. Voor meer informatie: www.deconferentie2010.nl.

25 t/m 29 januari 2011:
Onderwijstechnologie: NOT 2011
De Nederlandse Onderwijstechnologiebeurs (NOT) vindt plaats in de jaarbeurs in Utrecht. De NOT is dé vakbeurs voor onderwijsprofessionals. Nergens vindt u zo'n compleet aanbod voor alle onderwijssegmenten. Alle sectoren van het onderwijs zijn vertegenwoordigd. Ruim 450 exposanten brengen het nieuwste van het nieuwste uit de markt. Congres en workshops zorgen voor een hoog kennisgehalte. Voor meer informatie: www.not-online.nl. De Dyslexiestraat, een initiatief gestart in 2009, krijgt in 2011 navolging. De Dyslexiestraat in hal 10 is een gezamenlijke presentatie van specialisten op het gebied van dyslexie. Een uniek initiatief dat nog niet eerder op een Europese onderwijsvakbeurs te zien was. Na de beurs van 2009 bleef dit initiatief 'in de lucht' dankzij de digitale versie www.dyslexietraat.nl. De beursvariant van de Dyslexiestraat is voor bezoekers duidelijk herkenbaar aan de grote ballon die er boven hangt en alle deelnemers hebben – toepasselijk – een straatbordje op hun stand.

11 februari 2011: CLIN:
Computational Linguistics in the Netherlands 2011
De jaarlijkse CLIN conferenties vormen dé gelegenheid voor de Nederlandse en Vlaamse computerlinguïsten en taaltechnologen om elkaar te ontmoeten en hun meest recente, vaak lopende werk te presenteren. Ook onderzoekers van buiten Nederland en Vlaanderen nemen deel aan de CLIN conferenties. CLIN wordt sinds 1990 elk jaar aan een andere universiteit gehouden; komend jaar is dat in Gent, op 11 februari 2011. Voor meer informatie: http://lt3.hogent.be/clin21/about.html.

Libratory
Het digitaal onderzoekslaboratorium voor de geesteswetenschappen

De Nederlandse wetenschappelijke bibliotheken dragen de zorg voor talloze kwetsbare, zeldzame en kostbare materialen. Deze zogenaamde bijzondere collecties bevatten oude boeken, manuscripten, prenten, tekeningen, kaarten, archivalia, brieven en foto's. Het belang van deze bijzondere collecties voor onderwijs en wetenschap is evident. De bereikbaarheid en beschikbaarheid ervan is tegelijk echter beperkt. Vanwege het unieke karakter en de kwetsbaarheid, worden de meeste bijzondere collecties alleen in eigen studiezalen en onder toezicht beschikbaar gesteld aan studenten en onderzoekers. En dat terwijl zij in deze digitale tijden hun onderzoeksbronnen overal en altijd tot hun beschikking willen hebben.

Saskia van Bergen
Universiteits Bibliotheek Leiden

Libratory

Het project Libratory moet hier verandering in brengen. Libratory is een initiatief van de universiteitsbibliotheken verenigd in de Stichting Academisch Erfgoed[1], in overleg met het samenwerkingsverband van Nederlandse universiteitsbibliotheken en de Koninklijke Bibliotheek, het UKB. Het wil de aanzet vormen tot een landelijk plan van aanpak voor de grootschalige digitalisering van de bijzondere collecties in Nederlandse wetenschappelijke bibliotheken. Het einddoel is de vorming van een zo volledig mogelijk corpus, integraal raadpleegbaar en doorzoekbaar, dat de ruggengraat vormt van een digitaal onderzoekslaboratorium voor de geesteswetenschappen.

Kosten

De kosten voor het project bedragen bijna M€ 75 totaal, dat is M€ 4,8 per jaar bij een looptijd van 15 jaar. Een aanzienlijk bedrag, maar de Libratory partners zijn van mening dat dit onderzoekslaboratorium de deeltjesversneller van het Geesteswetenschappelijk onderzoek gaat worden. Libratory is eenvoudigweg noodzakelijk om ook in de toekomst innovatief onderzoek mogelijk te maken. Hoewel het op het eerste gezicht aantrekkelijk lijkt om voor een dergelijk project met een commerciële partner in zee te gaan, is dit voor de Libratory partners geen optie. De bijzondere collecties in Nederlandse wetenschappelijke bibliotheken behoren namelijk tot het publieke domein en behoren dat ook in de digitale wereld te zijn en blijven. Daarom zal voor de financiering van Libratory een beroep worden gedaan op de Rijksoverheid.

Basisplan

In de zomer van 2010 is het basisplan voor Libratory voltooid, dat is bedoeld als werk- en discussiedocument voor de verbreding van het project in twee richtingen. Ten eerste naar alle betrokken bibliothecaire partners in Nederland, verenigd in het UKB en ten tweede naar de wetenschappelijke partners; met name KNAW als actief voortrekker van de e-humanities, en NWO als financier en stimulator van wetenschappelijk onderzoek. Uiteraard geldt: hoe groter het draagvlak, hoe groter de kansen op honorering van de claim. Maar de inbreng van alle drie partijen is tevens noodzakelijk om Libratory geheel naar de wensen en vereisten van geesteswetenschappers in te kunnen richten.

Collectie

Libratory richt zich in de eerste plaats op gedrukte teksten, die door middel van OCR machineleesbaar kunnen worden gemaakt. Centraal staat de integrale, full-text beschikbaarstelling van alle titels in de Short Title Catalogue Netherlands (STCN), de nationale bibliografie voor alle boeken die

STCN	Short Title Catalogue Netherlands	Boeken tot 1800
MMDC	Medieval Manuscripts in Dutch Collections	Middeleeuwse handschriften
ISTC	Incunabula Short Title Catalogue	Gedrukte werken tot 1501
GGC	Gemeenschappelijk Geautomatiseerd Catalogussysteem	Boeken tussen 1501 en 1540 & tussen 1800 en 1840

tot en met het jaar 1800 in Nederland zijn gedrukt, of die daarbuiten in de Nederlandse taal verschenen zijn. Dit bestand wordt aangevuld met de digitale opbrengst van enkele complementaire landelijke catalogi:

Op deze manier ontstaat één digitale Boekcollectie Nederland, die de gehele boekproductie omvat vanaf de middeleeuwen tot 1840. In totaal zullen dankzij Libratory bijna 44 miljoen pagina's - ruim 220.000 unieke titels beschikbaar worden gesteld. Tegelijk met deze teksten zullen ook duizenden pagina's aan visueel materiaal (prenten, miniaturen, atlaskaarten) beschikbaar komen voor onderwijs en onderzoek.

Portaal

Tegelijk wordt een collectieportaal ingericht dat, naast de gedigitaliseerde boeken, ook inleidingen bevat op alle academische collecties in Nederland. Hierdoor kan de gebruiker niet alleen naar en in de boeken zoeken, maar wordt ook allerlei contextinformatie aangeboden, zoals de geschiedenis van een verzameling, of de relatie tussen boeken en prenten met dezelfde herkomst. Libratory moet bovendien een aanvraagmodule gaan bevatten, waarmee onderzoekers op verzoek handgeschreven en visuele collecties kunnen laten digitaliseren. Ook deze collecties worden toegevoegd aan het collectieportaal, dat zo steeds rijker zal worden en aan betekenis zal winnen.

Conclusie

Libratory wil veel meer zijn dan alleen een digitaliseringsproject. Einddoel is een virtueel laboratorium dat, op basis van het gedigitaliseerde tekstbestand, geheel nieuwe vormen van onderzoek mogelijk maakt. Uiteraard dienen hiervoor de nieuwste diensten en instrumenten te worden ingezet. Onderzoekers moeten in de eerste plaats complexe onderzoeksvragen kunnen stellen en de resultaten langdurig kunnen opslaan, annoteren en bewerken. Denk daarbij bijvoorbeeld aan een historisch overzicht van alle spellingsvarianten van één woord, die vervolgens op een tijdsbalk worden geplaatst. Of aan de geografische verspreiding van bepaalde grammaticale constructies. Ook moeten onderzoekers in staat worden gesteld om een zoekvraag in hedendaags Nederlands te stellen, en daarmee resultaten te genereren uit achttiende- en negentiende-eeuwse teksten. En onderzoekers moeten hun onderzoeksresultaten natuurlijk kunnen exporteren naar de software op hun eigen computer, zoals een concordantieprogramma. Het doel van deze grootschalige digitalisering is dan ook geen statische database, maar een interactieve webdienst waarbinnen iedereen zijn eigen onderzoekslaboratorium op maat kan samenstellen en inrichten.

Om samenwerking tussen onderzoekers te bevorderen, dient Libratory faciliteiten te bevatten waarmee zoekresultaten kunnen worden gedeeld. Ook moet Libratory gebruikersgroepen gaan bevatten, bijvoorbeeld voor onderzoeksgroepen of collegereeksen. Hiermee wordt aan de deelnemers de mogelijkheid geboden om bijvoorbeeld in teamverband aan transcripties, edities en digitale tentoonstellingen te werken. Voor de ontwikkeling van deze diensten en instrumenten zal nauw worden samengewerkt met onderzoekers en kenniscentra, zoals DANS voor het duurzaam opslaan van zoekresultaten, het Huygens Instituut voor het maken van digitale edities en CLARIN voor het beschikbaar stellen van taalmaterialen en taaltechnologie.

[1] *www.academischerfgoed.nl*

De studie van kenniscirculatie 17e-eeuwse geleerdenbrieven

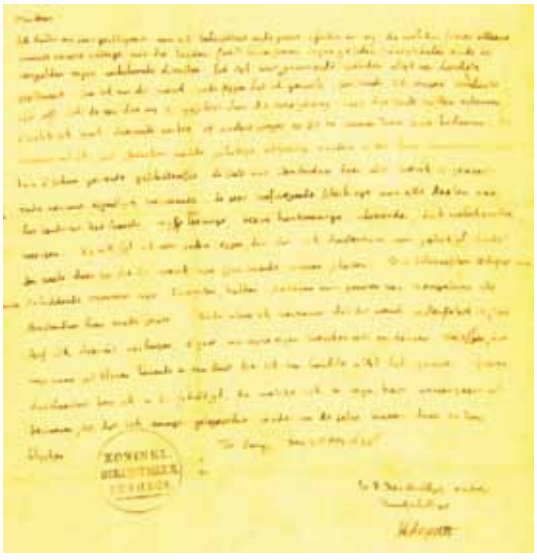
De ontwikkeling van de moderne wetenschap kent al meer dan vier eeuwen een aanhoudende groei. Ergens in de 16e en 17e eeuw begonnen drie elementen van wetenschap vruchtbaar op elkaar in te werken: theorie, experiment en toepassing. In de 17e eeuw was Nederland uitzonderlijk goed vertegenwoordigd in dat speelveld: Nederlandse geleerden correspondeerden met Europese collega's, waarvan sommigen, soms voor langere tijd, Nederland als basis kozen voor hun activiteiten.

Dirk Roorda
DANS
Data Archiving and Networked Services

Als je tegenwoordig wilt weten hoe kennis ontstaat en circuleert, vormen de wetenschappelijke tijdschriften een belangrijke vorm van informatie. In de 17e eeuw speelden brieven een vergelijkbare rol. Het leuke van brieven is, dat ze niet alleen een inhoud hebben, maar ook een afzender en een geadresseerde. En omdat brieven er op gericht zijn om op de bedoelde tijd op de bedoelde plaats aan te komen, verschaffen ze ons informatie over plaats en tijd waar de geleerden van de 17e eeuw verbleven bij het tot zich nemen van de nieuwe ideeën.

Omdat die brieven lang niet alleen over wetenschap gingen, maar ook over de toevalligheden van het dagelijks leven, is het in principe mogelijk om de gangen van en ontmoetingen tussen deze spelers in beeld te brengen als rijk verbonden netwerken.

Door netwerken van interacties uit de brieven te destilleren, kunnen verborgen verbanden zichtbaar gemaakt worden. Cats en Huygens hebben bijvoorbeeld vooral gemeenschappelijk dat ze dichters zijn, maar als je de ontmoetingen tussen Cats en Stevin vergelijkt met die tussen Huygens en Stevin, blijkt dat er een ontmoeting is geweest waarbij over waterstaatkundige oplossingen is gediscussieerd waar zowel Cats als Huygens bij aanwezig waren.

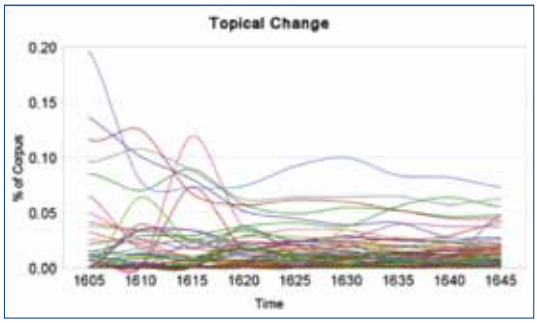


16de eeuwse geleerdenbrief

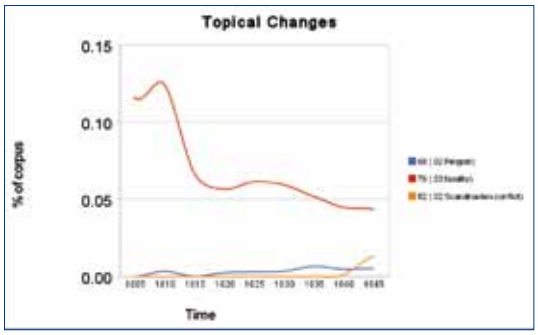
Er zijn zo'n 20.000 brieven van de belangrijkste geleerden in de Nederlandse 17e-eeuwse republiek. Die zou je allemaal kunnen lezen. Als je ze allemaal paarsgewijs wilt vergelijken moet je theoretisch zo'n 200 miljoen vergelijkingen maken. En als je 3 brieven in elke vergelijking wilt opnemen, dan kom je bij 100 miljard. De vraag is dus: hoe kun je automatisch de informatie die in het Nederlands, Frans of Latijn in die brieven opgeschreven staat, gebruiken om netwerken op te stellen die je in staat stellen om de gecombineerde informatie in die brieven op nieuwe manieren te exploreren?
Het Geleerdenbrievenproject, een NWO-middelgroot project[1], heeft dat precies als doelstelling. Partners zijn de Koninklijke Bibliotheek (om de bronnen in het goede formaat aan te leveren), het Descartescentrum van de Universiteit Utrecht, het Huygens Instituut van de KNAW, de Universiteit van Amsterdam (allen geïnteresseerd in tekstonderzoek en geschiedenis van kennisontwikkeling), de Virtual Knowledge Studio van de KNAW (om technieken aan te leveren waarmee patronen gevisualiseerd kunnen worden die relevant zijn voor de bestudering van kennisontwikkeling), en het DANS-instituut van de KNAW (om de bronnen en resultaten van het project duurzaam te archiveren en voor hergebruik beschikbaar te stellen).

Het zal u niet ontgaan zijn dat ergens in de keten tussen het oorspronkelijke brievencorpus en de daaruit resulterende netwerkvisualisaties een lastige stap bevindt: automatische taalverwerking. Nog steeds is men er niet in geslaagd software te schrijven waarmee computers dezelfde semantische en pragmatische

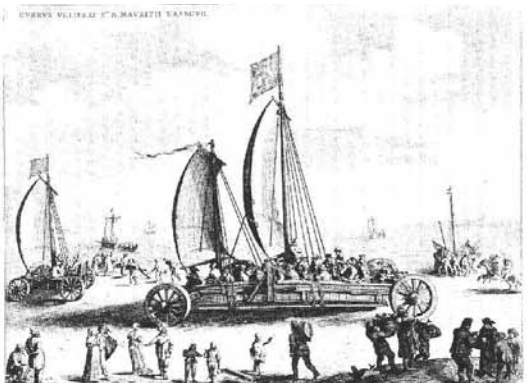
informatie uit natuurlijke taaluitingen kunnen halen als wij mensen dat gewoon zijn te doen. Er zit dus een flinke taalkundige uitdaging in dit project, dat dan ook een afnemer van toegepaste taalkunde is. Het extraheren van informatie uit teksten wordt aangepakt als een concert van afzonderlijke analysetools, waarvan de resulterende informatie bij elkaar gedirigeerd wordt tot aansprekende visualisaties.



Daarmee is dit project een gebruiker van taaltechnologie. De onderzoeksfocus van het Geleerdenbrievenproject is niet linguïstisch, maar letterkundig, (wetenschaps-) historisch en precies deze karakterisering maakte dat het project in aanmerking kwam voor steun van CLARIN.



CLARIN (Common LAnguage Resources and technology Infrastructure) is een Europese infrastructuur in opbouw met als doel om linguïstische data en tools voor alle Europese talen te kunnen aanbieden voor (her-)gebruik door met name geesteswetenschappers. Denk bij data aan woordenlijsten en corpora, en bij tools aan spellingnormaliseerders, ontleders (op woordsoortniveau en grammaticaal niveau), herkenners van eigennamen en meer. Om deze data en tools breed te kunnen inzetten voor wetenschappelijk gebruik moeten ze van goede, gestandaardiseerde metadata voorzien zijn, zodat ze vindbaar zijn, en moeten ze zo gebouwd en georganiseerd zijn, dat ze op elkaar toepasbaar zijn. En het is ook wenselijk dat processen herhaalbaar zijn, dus data en tools moeten 'persistent identifiers' krijgen, en zo gearchiveerd worden dat ze tot



Zeilwagen van Simon Stevin

in lengte van jaren via het web ontsloten zijn. Het Geleerdenbrievenproject is een dankbare afnemer van CLARINs taaltechnologie. Het profiteert van zowel diensten in natura als van financiering door CLARIN-NL van een demonstrator-traject waarmee de keten van bron tot visualisatie versneld en vereenvoudigd wordt afgelegd. Omgekeerd profiteert CLARIN van het Geleerdenbrievenproject op de manier zoals elke leverancier van services profiteert van zijn klanten: het ontwikkelen van use-cases die er in de praktijk toe doen, zodat de diensten ontwikkeld kunnen worden naar een voluit productief niveau.

Een belangrijke bevinding uit dit project zal zijn welke taaltechnologische tools nu werkelijk cruciaal zullen blijken te zijn bij het bereiken van inzichtgevende visualisaties. Om een voorbeeld te noemen: een tussenstap is het identificeren van sleutelwoorden in de tekst. Daar zijn statistische algoritmen voor. Nu zijn de spellingen van Nederlandse, Franse en Latijnse woorden in de 17e eeuw allemaal afwijkend van die van nu, en bovendien minder standaard. Laat je de sleutelwoordanalyse los op de ruwe tekst, dan krijg je mogelijk slechtere resultaten dan wanneer je eerst een spellingsnormalisatie slag doet. Dat zouden we graag eens constateren, en ook kwantificeren. Dit project biedt een uitstekende gelegenheid voor dergelijke analyses als bijproduct.

Tenslotte, als het project afgelopen is, gaat de stekker er niet uit. De bronnen blijven toegankelijk, de tools operationeel. Nieuwe (teams van) wetenschappers kunnen de resultaten gaan checken, de processen met andere parameterwaarden gaan overdoen, nieuwe inzichten als aantekeningen toevoegen, en duurzaam verwijzen naar de resultaten in hun publicaties.

Daarvoor staan DANS en CLARIN garant.

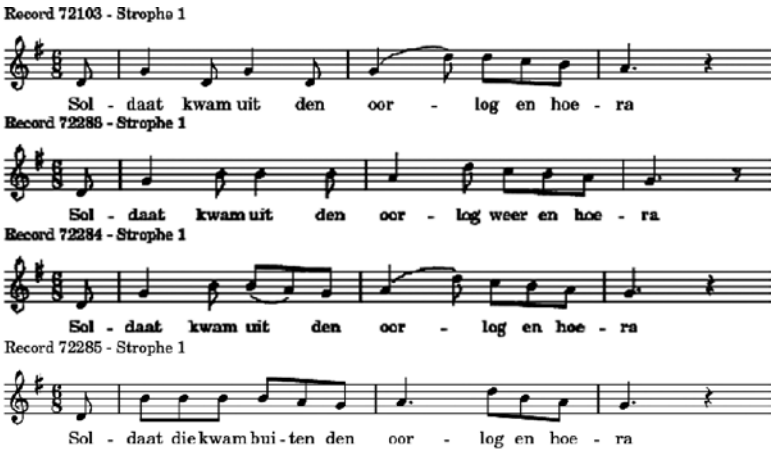
[1] *www.nwo.nl/nwohome.nsf/pages/NWOP_52XC3B*

Een melodieënzoekmachine voor de Nederlandse liederenbank

In de tweede helft van de twintigste eeuw reisden de veldwerkers Will Scheepers en Ate Doornbosch met hun bandrecorders stad en land af om Nederlandse volksliedeeren vast te leggen die verloren dreigden te gaan. Het radioprogramma Onder de groene linde van Ate Doornbosch speelde daarbij een belangrijke rol. Reacties van luisteraars leidden tot talloze nieuwe opnames. De in totaal ongeveer 7.000 opnames zijn intussen gedigitaliseerd en worden bewaard in het Meertens Insitituut in Amsterdam. Hiermee herbergt het Meertens Instituut een waardevolle collectie cultureel erfgoed. Sinds juni 2007 zijn de opnames publiek toegankelijk via de website van de Nederlandse liederenbank (www.liederenbank.nl). Om een bepaald liedje te vinden kan in de liederenbank in de meta-data (zoals titel, zanger, opnamedatum, etc.) worden gezocht.

Peter van Kranenburg
Meertens Instituut
Frans Wiering
Universiteit Utrecht

Een lang gekoesterde wens van het Meertens Instituut was om ook in staat te zijn op melodische inhoud te zoeken. In samenwerking met de Universiteit Utrecht is van 2006 tot 2010 aan dit probleem gewerkt in het WITCHCRAFT-project1. Dit artikel geeft een kort overzicht van de werkwijze en de resultaten.



Figuur 1 Voorbeeld van vier gerelateerde melodieën. Enkel de incipits ofwel de eerste woorden of openingslijn, zijn afgebeeld.

Een belangrijke voorwaarde voor het slagen van een dergelijk project is een interdisciplinaire aanpak. In Music Information Retrieval worden methoden ontwikkeld om op inhoud gebaseerd naar muziek te zoeken in grote verzamelingen, zoals op internet, of in digitale archieven. In de Muziekwetenschap, met name in het Volksliedonderzoek, wordt onderzoek gedaan naar verwantschappen tussen volksliedmelodieën. Beide disciplines zijn relevant binnen het WITCHCRAFT project.

Één van de eerste stappen was het begrijpen van de verwantschap tussen de volksliedmelodieën. Omdat deze mondeling worden overgedragen wil er nogal eens iets veranderen en soms kan de opeenvolging van veranderingen er zelfs tot leiden dat de oorspronkelijke melodie onherkenbaar is. Een belangrijk concept dat in het volksliedonderzoek wordt gebruikt om melodische verwantschap te beschrijven is *tune family* (melodie-familie). Hiermee wordt een groep melodieën aangeduid die als variant van elkaar kunnen worden beschouwd. Een voorbeeld is afgebeeld in figuur 1.

De domeinexperts aan het Meertens Instituut kennen de collectie *Onder de groene linde* door en door. Zij hebben – veelal impliciete – kennis over het soort relaties tussen de melodieën, die zij gebruiken om de melodieën in te delen in de juiste melodie-familie. Om deze kennis expliciet te maken is een annotatiemethode ontwikkeld waarmee de domeinexperts de gelijkenisrelaties tussen een groot aantal melodieën hebben geannoteerd. Hieruit bleek dat aspecten als ritme en contour een rol spelen bij het herkennen van een melodie, maar dat de aanwezigheid van karakteristieke motiefjes in een melodie het belangrijkst is.

In het WITCHCRAFT project zijn verschillende benaderingen onderzocht. De methode die uiteindelijk het meest succesvol is gebleken, gebruikt toonhoogte, ritme en structuur. De methode is gebaseerd op uitlijning

van reeksen. Als voorbeeld een uitlijning van twee reeksen woorden:

De ------ kat zit op het grijze paaltje
De kleine hond zit op het ------ paaltje

scores: 1 -1 0,5 1 1 1 -1 1
totaal: 3,5

De twee zinnen zijn uitgelijnd door op geschikte plaatsen gaten in te voegen en te accepteren dat twee gerelateerde woorden, zoals kat en hond, met elkaar uitgelijnd worden (substitutie).

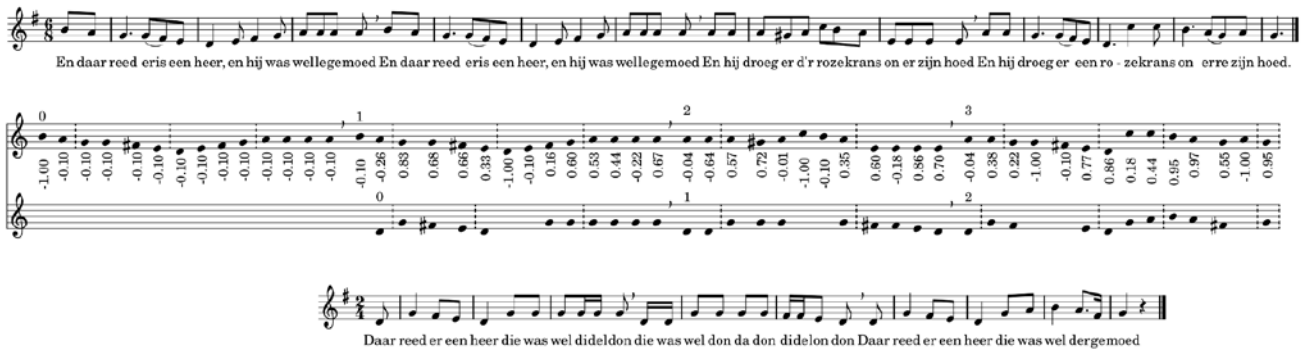
In de computerwetenschap bestaat een algoritme dat voor twee reeksen symbolen de optimale uitlijning kan uitreken, dat wil zeggen, waar gaten ingevoegd moeten worden om de uitlijning met de hoogst mogelijke score te verkrijgen. De score wordt berekend door de scores van de individuele associaties bij elkaar op te tellen. In het voorbeeld hierboven geven we een exacte match score 1, een niet-exacte match 0,5 en een gat -1. De totaalscore voor de uitlijning is dus 3,5.

In het geval van de melodieën zijn de noten de symbolen die we uitlijnen. De berekening van de scores stelt ons in staat muzikale kennis te verwerken in het algoritme. Dat is een zeer belangrijke eigenschap. Hier vinden we het aanknopingspunt voor een interdisciplinaire benadering. In feite beantwoordt een individuele score de vraag: hoe graag willen we dat deze twee symbolen (noten) met elkaar uitgelijnd worden? Het antwoord is muzikaal bepaald en hangt af van een aantal factoren.
De volgende voorwaarden bleken in de meeste gevallen voldoende: de **toonhoogte** dient ongeveer overeen te komen, het **metrisch gewicht** (de 'belangrijkheid' van de noot) dient overeen te komen, en de **noten** dienen zich op ongeveer dezelfde plaats binnen de zin te bevinden (bijvoorbeeld beide aan het begin van de zin, beide aan het eind, of beide in het midden). Een voorbeeld van zo'n uitlijning, inclusief scores, is afgebeeld in figuur 2.

Wanneer we een onbekende melodie hebben (de zoekvraag) en we willen gerelateerde melodieën in de collectie vinden, kunnen we als volgt te werk gaan. We berekenen de uitlijning (en daarmee de score) van de onbekende melodie met elk van de melodieën in de collectie. Vervolgens sorteren we deze volgens de scores van de uitlijningen. Dat levert een resultaatlijst op waarin de meest gelijkende melodie bovenaan staat. Bij een test met 360 zoekvragen en een collectie van bijna 5.000 melodieën blijkt dat in 90% van de gevallen een relevant zoekresultaat bovenaan de resultaatlijst staat en dat zich in 99% van de gevallen een relevant zoekresultaat bij de eerste 10 melodieën bevindt.

Met deze methode is een prototype van een melodieënzoekmachine voor de Nederlandse liederenbank gemaakt. Momenteel gewerkt aan een robuuste implementatie die publiek beschikbaar kan worden gesteld.

Hoewel het ontwikkelde model voldoende krachtig is om een zoekmachine voor de liederenbank te maken, nodigt het resultaat van het annotatie-experiment, namelijk dat motieven het belangrijkst zijn voor de herkenning van een melodie, uit tot nader onderzoek naar manieren om motieven te gebruiken voor het zoeken naar melodieën. Dit leidt tevens tot nieuwe onderzoeksvragen waarin relaties gelegd kunnen worden met de wijze waarop mensen melodieën onthouden of herkennen. In ieder geval zijn de huidige resultaten een goed voorbeeld van de wijze waarop muzikale kennis en computationele methoden gecombineerd kunnen worden in een succesvolle toepassing.



Figuur 2. Voorbeeld van een uitlijning van twee melodieën inclusief substitutiescores.

De netwerken van 2015
De Nederlandse letteren verbonden met beeld, geluid, krantenpapier en gps

De Digitale Bibliotheek voor de Nederlandse Letteren is aan het einde van 2010 een steeds sneller groeiende website met informatie over meer dan 20.000 auteurs, en met de volledige en zo goed als foutloze teksten van meer dan 5.000 boeken en van vele honderden volledige tijdschriften en naslagwerken. Ze vormen samen een hoogwaardige (op XML gebaseerde) bibliotheek. Sinds 2009 scant DBNL bovendien in een hoog tempo boeken en tijdschriften. De website bestrijkt daardoor een steeds groter deel van de Nederlandse cultuurgeschiedenis. In 2010 verwerkt DBNL meer dan 2,5 miljoen boekpagina's. DBNL is daarmee een van de grootste digitaliseringprojecten van Nederland en het zal de komende jaren alleen maar harder gaan.

René van Stipriaan
DBNL

Netwerken

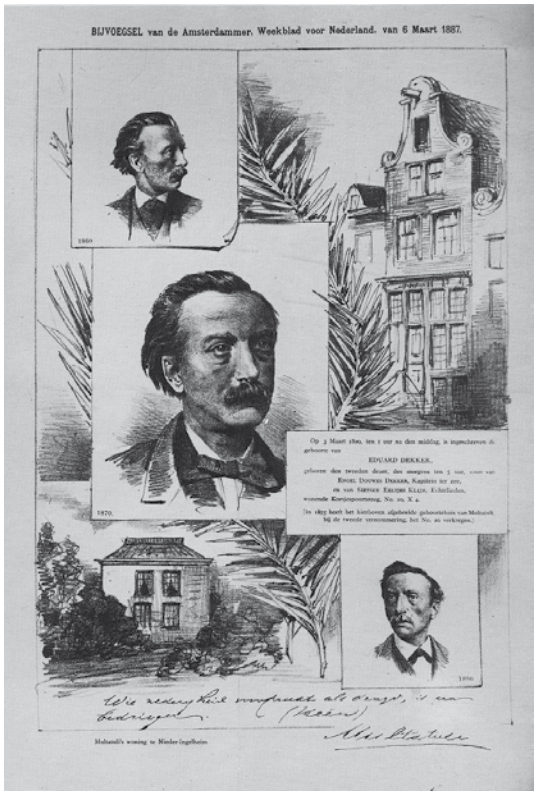
Alle gescande teksten worden daarna naar goed gestructureerde en gemetadateerde XML-documenten omgezet. Zo worden ze geschikt gemaakt voor alle mogelijke doelgroepen: niet alleen voor onderzoekers en studenten, maar ook voor scholieren, bibliotheekgebruikers en anderen die aan deze teksten specifieke informatie willen ontlenen. De DBNL-teksten worden op dit ogenblik op grote schaal verbonden aan andere netwerken. De zo lang eigenstandig opererende website www.dbnl.org raakt daardoor steeds nauwer verbonden aan de webomgevingen van de Vlaamse en Nederlandse openbare bibliotheken, en aan de onderzoeksomgeving van het Meertens Instituut en het Huygens Instituut. Er zullen ook lijnen gaan lopen naar websites van dagbladen, uitgevers, literaire tijdschriften en zelfs het Centraal Boekhuis. Iemand met interesse in de Nederlandse literatuur, taal en cultuur moet vanuit www.dbnl.org altijd verder geholpen kunnen worden naar andere betrouwbare kennisbronnen en diensten.

Multatuli

Vanaf het online gaan van de DBNL in 2000 heeft ze toegewerkt naar opname in hoogwaardige netwerken. In 2000 was de tijd nog niet rijp, in 2005 ook niet, maar de laatste twee jaar groeit in de Nederlandse culturele sector het besef dat twee verbonden websites aanzienlijk meer kracht kunnen ontwikkelen dan twee afzonderlijke websites. Het slim stroomlijnen van de datamodellen vraagt de komende jaren veel aandacht, maar de opbrengst zal enorm zijn. Wie op de DBNL-website nu Multatuli aanklikt vindt de volledige tekst van ten minste 8 biografieën, zo goed als alle primaire werken in verschillende edities, ca. 200 artikelen en studies over Multatuli en een enkel audiobestand naast een schamel aantal links naar eventuele nuttige websites. Hier wordt de komende jaren een veelheid aan opties aan toegevoegd, liefst vergaand geïntegreerd in de DBNL-omgeving. Het Multatulimuseum biedt er uitdrukkelijk zijn waren en diensten aan. Er zijn e-books te koop of voor korte tijd te huur. Wie een specifiek boek van Multatuli op papier wil hebben kan het met één beweging als printing on demand bestellen of opsporen bij een antiquariaat. Wie het gedrukte boek wil lenen krijgt, eventueel met behulp van GPS, de weg naar de dichtst bijzijnde bibliotheek gewezen, waar het boek al klaar ligt als hij of zij er arriveert. Even gemakkelijk zijn recensies op te vragen die in de loop der jaren over specifieke boeken in kranten zijn verschenen, en ook radioprogramma's en tv-documentaires die ooit aan Multatuli zijn gewijd. In een Google Earth-achtige atlas zijn alle plaatsten te vinden die verband houden met Multatui's leven en werken. De *Max Havelaar*-verfilming van Fons Rademakers is te bekijken, eventueel met een koppeling van elke scene aan de betreffende pagina in het boek. En als we toch bezig zijn dan kunnen ook meteen de Twitter-, Facebook- en Hyves-pagina's integreren die aan Multatuli gewijd zijn, om het laatste nieuws over de grootste schrijver uit de Nederlandse letteren te kunnen vernemen.

Zoeken en vinden

Er is niet veel dat ons belet om onze informatie op deze manier te verknopen met



informatie van anderen. Er wordt al met overgave aan gewerkt, maar de praktijk is weerbarstig. Lang hebben velen in de erfgoedsector in de illusie geleefd dat digitaliseren een kwestie was van scannen, indexen draaien en zoekmachines voor de neus van gebruikers neerzetten. Het ging zo goed als vanzelf. Het vindbaar maken van data had geen prioriteit. De oude trouwe documentalisten werden vervangen door handige ICT'ers, die beweerden voor alle problemen van de erfgoedsector oplossingen te hebben. Die oplossingen bleven veelal uit. Er zijn de afgelopen jaren vele miljoenen verspild aan slecht uitgedachte en uitgevoerde projecten[1]. Het wordt tijd om de documentalisten, maar dan in een nieuwe gedaante, hun rentree te laten maken. Ze zullen de matrix van de Nederlandse cultuurgeschiedenis, en uiteindelijk ook de cultuurgeschiedenis van Europa en de wereld gestalte moeten geven en onderhouden. Want kennis van zaken is in het verknopen van informatie beslissend. Wanneer informatie goed gestructureerd is, dan ligt het rendement voor een veelheid aan internetdienst voor het grijpen. Techniek is daarin dienend, en niet leidend, zoals lang is gedacht.

Er zal veel aan de kwaliteit van de data en aan de uitwisselbaarheid moeten worden gesleuteld, maar elke inspanning in het fijn slijpen van datamodellen levert in beginsel een blijvend resultaat op. En verder moet er nog enorm veel gedigitaliseerd en ontsloten worden. Er is zoveel ervaring opgedaan met hoogwaardige digitalisering van teksten, dat tegen betrekkelijk lage kosten in korte tijd een hoge opbrengst bereikt kan worden. Als we ervan uitgaan dat de Nederlandse cultuurgeschiedenis, met zijn zwaartepunt de literatuur en de taal- en literatuurwetenschap, bij elkaar ongeveer 50 miljoen pagina's beloopt, dan moet het mogelijk zijn om die vóór 2020 geheel en ook tegen de hoogste kwaliteitsstandaarden digitaal te maken. Er zijn op dit ogenblik plannen in de maak om zelfs al vóór 2015 een representatief corpus aan primaire en secundaire teksten aan te bieden. Representatief betekent in dit geval dat het heel moeilijk is voor de gemiddelde gebruiker om teksten te bedenken die niet in het corpus voorkomen. De overheid zal daar stevig in moeten investeren, maar deze investering zal een zodanig rijke, hoogwaardige en comfortabele webomgeving opleveren dat de consument voor het onderhoud en de verdere uitbreiding zal willen betalen.

Het is 2010 en nu het gaat het pas echt beginnen. Het eerste decennium was aan de zoekmachines, het tweede decennium zal voor de netwerken zijn. De afgelopen tien jaar waren aan de vrije jongens, de durfals en de amateurs, het moet raar lopen willen de echte professionals het de komende tien jaar niet overnemen. Het internet is een te mooi medium om in chaos ten onder te laten gaan.

[1] Zie artikel 'De nieuwe bibliotheek' in: De Gids 2010, nr. 3, p. 335-343

Verteld Verleden: gesproken getuigenissen online

In 2012 is het 50 jaar geleden dat een medewerker van het Instituut voor Geschiedenis van de Rijksuniversiteit Groningen bij de Haagse Post de opnamebanden opvroeg van een serie vraaggesprekken met prominente Nederlanders. Dit initiatief kan gezien worden als het startpunt van de aanleg van interviewcollecties in Nederland ten behoeve van de historische wetenschap: Oral History. Verteld Verleden koppelt 50 jaar Oral History verzamelen aan de mogelijkheden van nieuwe technologie voor de verbetering van de toegankelijkheid.

Roeland Ordelman
Nederlands Instituut voor Beeld en Geluid

Onder Oral History wordt verstaan het bevragen van een respondent met de bedoeling zijn specifiek persoonlijk geheugen vast te leggen zodat dit kan dienen als historische bron. Zo'n getuigenis is een primaire bron: het materiaal wordt zo opgetekend als de persoon het gezegd heeft. Daarmee onderscheidt het Oral History interview zich van een 'gewoon' interview (in tijdschrift, boek, of op radio/tv), omdat in dit laatste geval een regisseur een uitsnede uit het materiaal heeft gemaakt.

Schat aan getuigenissen

In 50 jaar is een grote schat aan gesproken getuigenissen verzameld. Een inventarisatie van de collecties van de Stichting Film en Wetenschap in 1996 leverde een lijst op van zo'n 1500 uur interviews met meer dan 1.000 personen[1]. Deze collectie is echter maar een beperkt deel van wat er aan OH materiaal in Nederland in dozen, op planken, harde schijven, en in het gunstigste geval, goed georganiseerde digitale archieven, is terug te vinden. Naast de collecties die op nationaal en regionaal niveau buiten deze inventarisatie zijn gevallen, is er sinds die tijd veel materiaal bijgekomen, mede dankzij voortschrijdende technologie (denk aan opnameapparatuur en de computer). De voortschrijdende technologie moet er nu ook voor zorgen dat dit cultureel erfgoed beter toegankelijk wordt.

Het doel van het Verteld Verleden is de mogelijkheden van moderne technologie om audiovisuele data toegankelijk te maken, onder de aandacht te brengen van Oral History collectiebeheerders in Nederland en tegelijkertijd te horen van gebruikers wat hun wensen zijn. In Verteld Verleden is er voor gekozen met een aantal collectiebeheerders een praktisch voorbeeld uit te werken en aan de hand hiervan -letterlijk- het land in te gaan om met het veld in gesprek te gaan over de mogelijkheden, onmogelijkheden en verbeterpunten bij het gebruik van technologie.

Praktisch voorbeeld

Verteld Verleden bouwt een Oral History web-portal die het mogelijk maakt om in Oral History collecties te zoeken in archieven verspreid over heel Nederland. Collectiebeheerders houden zelf het beheer over hun collecties en stellen de metadata beschikbaar via een protocol (OAI-PMH) dat andere partijen, zoals in dit geval de Verteld Verleden server, deze metadata kan binnenhalen (harvesten in jargon) en doorzoekbaar kan maken via een gebruikersinterface op de Verteld Verleden portal. Met behulp van zoektechnologie kunnen aangesloten collecties ook inhoudelijk aan elkaar gekoppeld worden. Binnen dit infrastructurele raamwerk maakt Verteld Verleden waar mogelijk gebruik van moderne technieken zoals spraakherkenning, persistent identifiers of thesauruskoppelingen.

Colofon

Verteld Verleden is een project van Nederlands Instituut voor Beeld en Geluid, DANS, Meertens Instituut en Aletta, instituut voor vrouwengeschiedenis en gefinancierd in het kader van de subsidieregeling Digitalisering met Beleid. Verteld Verleden wordt ondersteund door een klankbordgroep en een validatiegroep met vertegenwoordigers uit het erfgoedveld. Verteld Verleden ging in maart dit jaar van start en zal in ieder geval tot 2012 doorlopen. De portal is te vinden op http://verteldverleden.org.

[1] *Cirkelen om een centrum, 35 jaar interviewcollectie bij Stichting Film en Wetenschap, 1996, Mieke Lauwers*

NOTaS-Nieuws

Harry Bunt
Secretaris
NOTaS-bestuur

TST-paviljoen met NOTaS-deelnemers bij Overheid en ICT

Van 27 t/m 29 april 2010 vond in de Jaarbeurs in Utrecht het evenement Overheid & ICT plaats. Vijf NOTaS-deelnemers hadden zich op initiatief van het NOTaS-bestuur en Gridline voor deze gelegenheid verenigd en een paviljoen ingericht dat geheel gewijd was aan taal- en spraaktechnologie ten behoeve van de overheid. Gebruik van taal- en spraaktechnologie in de digitale communicatie tussen burgers en overheden is een voorbeeld daarvan; een andere is het toepassen van intelligente taaltechnologie in het zoeken in de steeds sneller groeiende hoeveelheid informatie van de overheid. In het paviljoen werden een groot aantal succesvolle taal-, spraak-, en zoektoepassingen gedemonstreerd en werd voorlichting gegeven over wat er allemaal mogelijk is. Deelnemers aan het TST-paviljoen waren: Dedicon, Gridline, de TST-Centrale, Telecats en Knowledge Concepts. De bedrijven waren enthousiast en hebben al weer geboekt voor volgend jaar. Er zijn veel leads verzameld en de algemene tevredenheid was groot.

NOTaS-microsubsidie: 150 euro

Het NOTaS-bestuur heeft besloten om in het vervolg een vergoeding van 150 euro uit te keren aan deelnemers die op een beurs of ander evenement publiciteit aan NOTaS geven, bijvoorbeeld door het verspreiden van brochures. Wel altijd tevoren afstemmen met de penningmeester.

Deelnemersbijeenkomst bij de Rabobank

In december 2009 vond een succesvolle deelnemersbijeenkomst plaats in het hoofdgebouw van de Rabobank in Utrecht, met als gastheer Ruud Smeulders, innovatiemanager bij de Rabobank. Hij verzorgde in inleiding over nieuwe ontwikkelingen die mogelijk van belang zijn voor de toepassing van taal- en spraaktechnologie.
Jan Odijk, als vertegenwoordiger van de Universiteit Utrecht, die vorig jaar deelnemer van NOTaS geworden is, vertelde over ontwikkelingen in Europese projecten waarin taal- en spraaktechnologie een rol speelt.

Europees subsidieprogramma voor taaltechnologie en MKB

Tijdens de Language Technology Days op 22-23 maart in Luxemburg heeft Roberto Cencioni, hoofd van Directoraat E, 'Digital Content and Cognitive Systems', plannen ontvouwd voor de ontwikkeling van een EU-subsidielijn voor taaltechnologie, met een focus op steun aan het MKB. Wilt u van deze ontwikkeling op de hoogte gehouden worden, of een actieve rol spelen bij het inspelen op de mogelijkheden van dit programma? Stuur dan even een berichtje naar info@notas.nl.

Te gast bij het INL

In juni 2010 waren de NOTaS-deelnemers en het bestuur te gast bij het Instituut voor Nederlandse Lexicologie in Leiden.
Jeanine Beeken, directeur van het INL, verwelkomde de deelnemers, waarna Rob Tempelaars een demonstratie gaf van de net operationeel geworden site `Meldpunt Taal' (www.meldpunttaal.nl en www.meldpunttaal.be), waar taalgebruikers van het Nederlands en onderzoekers samenwerken in het vastleggen van alle ontwikkelingen in de taal. De site lijkt meteen succesvol: observaties van taalgebruikers stromen binnen!
Linde van den Bosch, algemeen secretaris van de Nederlandse Taalunie, praatte de deelnemers bij over de ontwikkelingen rond een mogelijk vervolg van het STEVIN-programma, en peilde de belangstelling en specifieke ideeën bij de deelnemers.
Tenslotte vertelde gastspreker Kees-Jan Diepstraten (Presis BV) over de mogelijkheden om 'social media' zoals Twitter in te zetten voor zakelijk gebruik.

LinkedIn groep

Sinds een tijdje bestaat deze NOTas-groep voor LinkedIn. Hoe kunnen we er nu samen (deelnemers en geïnteresseerden) voor zorgen dat de faciliteiten van deze groep optimaal worden benut? Bijvoorbeeld de interactieve mogelijkheden zouden een prachtige aanvulling op de NOTaS-website kunnen zijn. Wij staan open voor jullie ideeën.

Stichting
NOTaS
Nederlandse Organisatie voor
Taal & Spraaktechnologie

Het CLARIN-NL project

Jan Odijk
Uil-OTS, Universiteit Utrecht, CLARIN-NL

Het probleem

Onderzoekers uit de geesteswetenschappen gebruiken grote hoeveelheden tekst als 'grondstof' voor hun onderzoek. Zo gebruiken bijvoorbeeld linguïsten allerlei grammatica's, woordenboeken en tekstuele archieven. Historici gebruiken vooral boeken, pamfletten, kranten en, in toenemende mate, transcripties van audiovisuele opnamen zoals documentaires en interviews. Literaire onderzoekers gebruiken naast commentaren en kritieken ook verschillende uitgaven van hetzelfde boek of artikel.

De hoeveelheid digitaal beschikbaar materiaal is enorm en zal de komende jaren nog eens sterk toenemen wanneer projecten zoals Libratory en Google Books eenmaal op stoom komen. Uiteindelijk zal er zoveel materiaal zijn, dat nieuwe methoden ontwikkeld moeten worden om er iets zinnigs mee te doen: traditionele manieren (het 'handmatig' doorlezen of afluisteren/bekijken) zullen dan niet meer voldoende zijn, en de standaard zoekfunctionaliteit die bijv. Google biedt zal onvoldoende precies zijn (te veel irrelevante 'hits') om goed bruikbaar te zijn.

Gelukkig is er een hele reeks taal- en spraaktechnologische tools beschikbaar om iets zinnigs met die enorme hoeveelheden data te kunnen doen. Spraak- en sprekerherkenning kunnen gebruikt worden om iets te zeggen over wie wat wanneer zei (bv in de Radio Oranje[1] demo bij het NIOD). Andere tools kunnen automatisch zinnen parsen[2] in grote corpora om vervolgens met de resultaten weer onderzoek te doen zoals nu al gebeurd in het STEVIN-onderzoeksproject Lassy[3].

Hoewel het altijd nog beter kan, kan men stellen dat veel kennis en veel bruikbare tools wel aanwezig zijn, maar dat door de afwezigheid van standaarden op dit gebied, het zeer lastig zo niet onmogelijk is om willekeurige tools op willekeurige data los te laten. De ene tool vereist een specifiek XML-dataformaat met bepaalde tags terwijl de andere tool juist platte tekst of een relationele database verwacht. De output van een parser van de universiteit van Groningen kan vervolgens niet zomaar door een analyse programma van de universiteit van Gent of Utrecht gebruikt worden, laat staan dat teksten met een verschillend formaat door elkaar heen gebruikt kunnen worden. Op dit moment wordt er voor dit soort gezamenlijk onderzoek steeds weer een transformatie van het ene dataformaat naar het andere gemaakt; iets wat op den duur niet meer realiseerbaar is. Zeker wanneer we willen dat onderzoekers uit de geesteswetenschappen geheel zelfstandig (zonder de hulp van de TST-experts) onderzoek op grote tekstcorpora gaan doen, is een vergaande vorm van standaardisering absoluut noodzakelijk.

Behalve de problemen met de verschillende data formaten en de verschillende in- en output eisen van de tools, is er nog een groot probleem. Veel data wordt speciaal voor een onderzoek verzameld. De data verzameling is (bijna) nooit het doel van het onderzoek; dat zijn de onderzoeksresultaten en de publicaties. Is een onderzoek eenmaal afgerond dan gaat men in de regel verder met een volgend onderzoek en blijven de data verweesd achter. Na verloop van tijd verdwijnen de data in persoonlijke archieven en als een onderzoeker verdwijnt (pensionering, andere baan, andere onderzoeksinteresse) dan volgen de data meest snel.

De oplossing

Het CLARIN-NL project, een nationale versie van het Europa brede 'CLARIN-Europe initiative', probeert aan de twee hierboven geschetste problemen iets te doen door een gestandaardiseerde CLARIN infrastructuur voor Geesteswetenschappen op te zetten. Ten eerste geeft het de onderzoekers de mogelijkheid om al beschikbare of nieuw gecreëerde relevante onderzoekdata op een gestandaardiseerde manier op te slaan zodat de data a) voor langere tijd bewaard kunnen worden en dus beschikbaar blijven als de onderzoeker 'verdwijnt' en b) andere onderzoekers makkelijk bij deze data kunnen komen doordat precies bekend is hoe ze zijn opgeslagen, ze vindbaar zijn via gestandaardiseerde beschrijvingen ervan(de metadata), en doordat de betekenis van elementen in de metadata is vastgelegd.

Dit laatste is zeer relevant en minder triviaal dan men dikwijls denkt. Zo schijnt er

een venussonde te zijn neergestort doordat er in een xml-bestand voor landingssnelheid '2.7' stond. De Amerikanen gingen uit van mijlen terwijl de Britten kilometers bedoelde!

Alleen maar data en metadata opslaan is dus niet voldoende: er moet precies bekend zijn wat de verschillende data voorstellen. Daarom bevat de CLARIN infrastructuur een registratie van de betekenis van elementen (datacategorieën) die voorkomen in de data en de metadata (ISOCAT[4]). Met behulp van de browsing en zoekfunctionaliteit kunnen onderzoekers op heel precieze manier zoeken in zowel de data als de metadata. Een eerste voorlopige versie van een catalogus van de data met de bijbehorende browsingfunctionaliteit is beschikbaar in het 'Virtual Language Observatory'[5]. Aan uitbreidingen van de zoekfunctionaliteit wordt momenteel hard gewerkt. De CLARIN infrastructuur zal betere zoekmogelijkheden bieden dan bv Google: de resultaten zullen preciezer zijn met mindere irrelevante hits. Dit wordt bereikt door optimaal gebruik te maken van de CLARIN-specifieke structuur van data en metadata, en de betekenis van de elementen erin.

Virtuele collecties

CLARIN zal bovendien de mogelijkheid scheppen om zelf 'virtuele' collecties te maken. Virtuele collecties zijn collecties die een onderzoeker samenstelt door delen van andere, al bestaande collecties te combineren. Dit combineren van bestaande collecties is ook mogelijk door de standaardisatie van de structuren de betekenis van de elementen die erin voorkomen.

Zowel voor browsen, zoeken en voor het maken van virtuele collecties is het essentieel dat verwijzingen naar data altijd kloppen. URLs worden daar meestal voor gebruikt, maar die zijn zeer onbetrouwbaar aangezien die regelmatig veranderen of gewoon verdwijnen. De CLARIN-infrastructuur probeert dit probleem op te lossen door gebruik te maken van zogeheten 'Persistent Identifiers' (PID's)[6].

Om dit allemaal duidelijk te maken voor zowel experts als niet-experts, verzorgt CLARIN een groot aantal trainingen en workshops waarin de mogelijkheden van de beoogde CLARIN-infrastructuur zullen worden uitgelegd.

Het door NWO betaalde CLARIN-NL programma loopt van april 2009 tot april 2015 en heeft een budget van M€ 9.01. Voor meer informatie zie de CLARIN-NL website[7].

[1] http://niod.al-m.nl/nl/thema/10/
[2] Parsen is het toekennen van grammaticale termen zoals onderwerp, gezegde, aan de verschillende woorden.
[3] www.let.rug.nl/erikt/lassy/
[4] www.isocat.org
[5] www.clarin.eu/vlo/
[6] www.ariadne.ac.uk/issue56/tonkin/
[7] www.clarin.nl

CLARIN-NL

Documentatie van Bedreigde Talen & DOBES-programma

Als u het nieuws volgt is u wellicht opgevallen dat er bericht werd van de ontdekking van een nieuwe taal in het uiterste noordoosten van India, het 'Koro'. Onderzoekers stootten 'bij toeval' hierop bij documentatiewerkzaamheden van een andere, al wel bekende taal 'Aka'. Hoewel meer van deze berichten niet zijn uit te sluiten, is het waarschijnlijker dat er bericht zal moeten worden over het uitsterven van talen. Van alle nu bekende ongeveer 7.000 talen wordt in een UNESCO rapport verwacht, dat maar de helft of wellicht zelfs maar een tiende nog zal bestaan aan het einde van deze eeuw. Oorzaken zijn tegenwoordig niet meer het fysiek uitsterven van volkeren en gemeenschappen, maar het uitsterven is nu veelal een gevolg van de druk die minderheden nu ondervinden in een anderstalige omgeving om hun taal en cultuur op te geven.

Daan Broeder en Paul Trilsbeek
Max Planck Instituut, Nijmegen

Het is van groot belang 'bedreigde talen' adequaat te documenteren en het verzamelde materiaal veilig te stellen voor de toekomst. Het gaat hierbij niet alleen om materiaal dat van belang is voor linguïsten maar ook voor antropologen en ecologen. Daarnaast is het materiaal interessant voor het grote publiek, omdat in de documentatie nu eenmaal veel informatie besloten ligt over de cultuur en de natuurlijke leefomgeving van de taalgemeenschap.

Een ander belangrijk element van de documentatie van bedreigde talen is dat het essentieel is bij revitalisatiewerk. Hierbij wordt getracht een taal weer ingang te doen vinden in de gemeenschap en het materiaal kan dan bijvoorbeeld gebruikt worden voor het vervaardigen van lesmateriaal zodat de diversiteit in stand blijft.

Data

Normaliter wordt documentatie verzameld door onderzoekers die een relatie opbouwen met informanten uit de taalgemeenschap. Vervolgens wordt dan een grote verscheidenheid aan video, foto's en geluidsopnames gemaakt waar de informanten verhalen vertellen of bezig zijn met alledaagse en minder alledaagse activiteiten zoals het bouwen van een huis. Deze opnames worden vervolgens geanalyseerd en geannoteerd en daarna worden er bijvoorbeeld lexica en grammatica's aan toegevoegd.

De eerlijkheid gebied te zeggen dat er ook zonder uitdagende expedities het nodige werk te doen is. In het verleden is al veel materiaal verzameld dat eveneens goed gearchiveerd zou moeten worden. Veelal zijn de gebruikte datadragers zoals audiotapes en zelfs Cd-roms geen stabiele dragers. Een UNESCO rapport stelt dat 80% van alle opnames over talen en cultuur zelf bedreigd zijn.

DoBeS

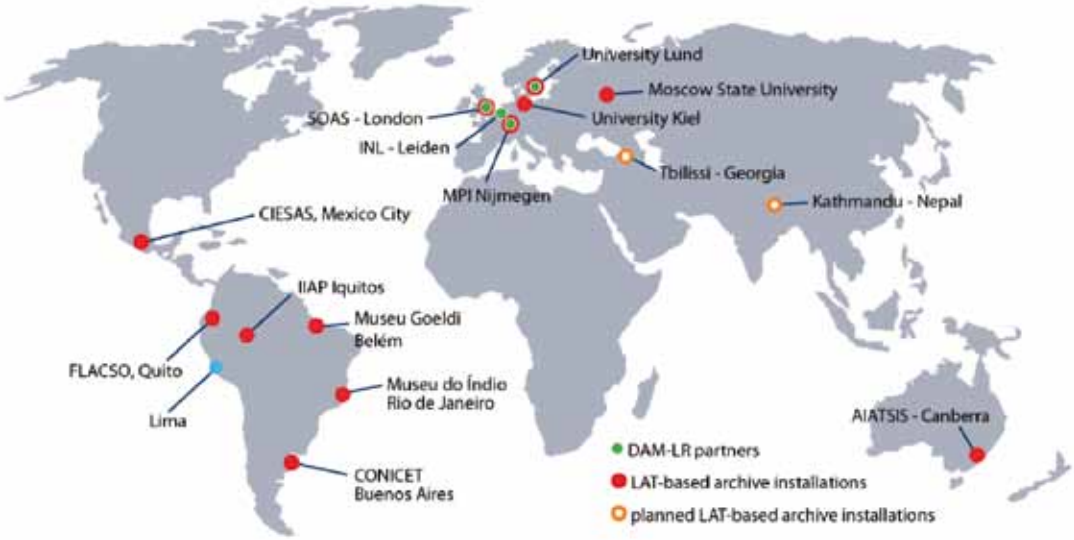
DoBeS[1] (Documentation Bedrohte Sprachen, een initiatief van de Volkswagen Stiftung)

werkt aan de documentatie van bedreigde talen. Het is in zijn soort een van de grootste programma's ter wereld. DoBeS ging van start in 2000 met zeven documentatieteams en een vertegenwoordiging van het Max-Planck Instituut voor Psycholinguïstiek (MPI) uit Nijmegen dat zorg draagt voor de technische ondersteuning. Het gearchiveerde materiaal wordt om veiligheidsredenen gekopieerd naar twee grote rekencentra en door de overkoepelende Max-Planck Geselschaft gegarandeerd voor de komende 50 jaar. Inmiddels zijn er 50 documentatieteams uitgezonden (zie figuur 1) en een groot gedeelte van het door hen verzamelde materiaal is nu opgeslagen in het archief van het MPI. De documentatieteams bestaan vaak uit onderzoekers met een grote reputatie. David Harrison, een van de onderzoekers die verantwoordelijk was voor de ontdekking van Koro waar we dit artikel mee begonnen, maakte in het verleden ook deel uit van een DoBeS documentatieteam.

De DoBeS richtlijnen stellen dat de documentatieteams materiaal verzamelen dat de taal documenteert in zijn culturele context en dat bruikbaar moet zijn voor een brede waaier van disciplines. Verder moet het begrepen kunnen worden zonder voorafgaande kennis van de gedocumenteerde taal en gebruikt kunnen voor revitalisatieprojecten zoals boven beschreven. Belangrijk is ook de eis dat al het materiaal digitaal en volgens open standaarden moet worden gearchiveerd en beschikbaar gesteld via internet. Dit laatste was niet zo vanzelfsprekend in de onderzoekswereld van bedreigde talen. Soms gaat het om opnames van personen of ceremonieën waarvan de taalgemeenschap vindt dat deze om culturele redenen niet voor iedereen toegankelijk moeten zijn of worden er dingen besproken die de informant kunnen benadelen. Het MPI heeft daarom speciale software ontwikkeld waarbij de oorspronkelijke dataverzamelaars op ieder moment kunnen specificeren welke personen toegang hebben tot de data.
De MPI LAT archief software[2] stelt de documentatieteams en de door hen aangewezen collega's in staat om zelf via het web data te kunnen up- en downloaden, maar toch gebeurt, omwille van de efficiëntie, nog een groot gedeelte van het preparatie werk (zoals digitalisering) door het MPI archiveringsteam.

Regional Archives Initiative

De DoBeS documentatieteams onderhouden meestal goede contacten met wetenschappelijke instituten of universiteiten in de landen waar zij hun documentatiewerk verrichten. Hierdoor werd het MPI archiveringsteam zich bewust van het ontbreken van professionele archiveringsmogelijkheden voor talige data in deze landen. Besloten werd tot het Regional Archives Initiative[3], waar kopieën van de door het MPI/DoBeS ontwikkelde archiveringssoftware werden geïnstalleerd bij



geselecteerde partners over de hele wereld. Dit bevordert de ontwikkeling van een lokale infrastructuur voor archivering en helpt lokale onderzoekers bij het verkrijgen van steun voor hun werk.

Deze 'regional archives' hebben kopieën van het materiaal dat voor het DoBeS-programma in hun land is verzameld, maar ook van het materiaal dat door lokale projecten is aangedragen. Op dit moment zijn er acht archiefinstallaties binnen het Regional Archives Initiative opgebouwd en vier binnen andere samenwerkingsprojecten. Het is de bedoeling dat er een netwerk voor data-uitwisseling tussen de archieven gebouwd zal worden. Mogelijkheden om hiervoor aan te sluiten bij de binnen het EU CLARIN project[4] ontwikkelde e-infrastructuur worden onderzocht.

[1] DoBeS, www.mpi.nl/DOBES/
[2] LAT, www.lat-mpi.eu/
[3] Regional Archives Initiative, www.mpi.nl/DOBES/regional_archives/
[4] CLARIN, www.clarin.eu/

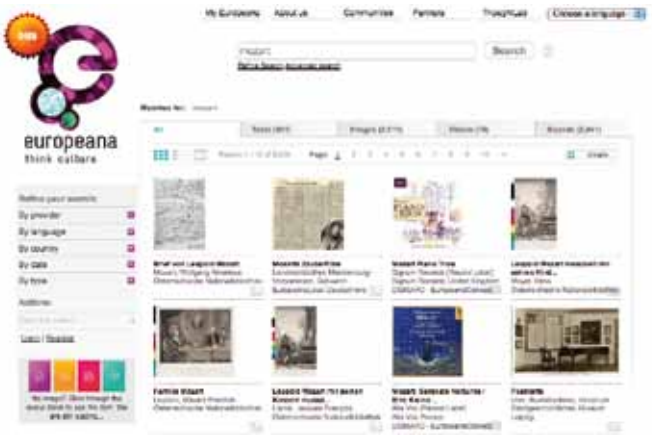
Europeana
Meertalig zoeken in metadata

Europeana is Europa's toegangspoort voor gedigitaliseerd cultureel erfgoed. Het combineert informatie over gedigitaliseerde objecten uit 27 landen: de lidstaten van de Europese Unie plus Noorwegen, Zwitserland en IJsland. Europeana ontsluit nu (oktober 2010) ruim 13 Miljoen objecten uit diverse domeinen: bibliotheken, archieven, musea en audiovisuele collecties. Europeana bevat niet de objecten zelf, maar de door deze instellingen aangeleverde metadata en links naar het volledige digitale object en de lokale pagina waar het object beschreven staat.

Jan Molendijk
Europeana
Foundation

Uitdagingen

Bij het goed meertalig doorzoekbaar maken van de gegevens in Europeana stuiten we onder meer op de volgende moeilijkheden.



Ten eerste de omvang. En dan niet een teveel, maar een tekort aan omvang: een typisch metadata record zoals wij dat in Europeana aangeleverd krijgen bevat ca. 50 woorden en identifiers. Dat is natuurlijk onvoldoende voor statistische zoekmethoden en het bepalen van relevantie van een bepaald zoekwoord binnen die 'tekst'. Daar staat tegenover dat de metadata records wel gestructureerd zijn – bepaalde velden zijn aangeduid als titel, beschrijving, maker etc.

Een tweede belemmering van taalkundige analyse is dat er niet bij verteld wordt in welke taal een record, of zelfs een veld binnen een record opgesteld is. Een bibliotheekrecord uit Nederland kan heel goed een Zweeds boek beschrijven, met een Zweedse titel.

Ten derde: gebruikers verwachten misschien dat Europeana de volledige tekst van al die boeken en archiefstukken doorzoekbaar maakt. Europeana is echter opgezet als een metadata verzameling: de volledige objecten blijven berusten bij de aanleverende instellingen. Van veel objecten is ook geen volledige tekst beschikbaar (oudere tekstuele werken, handschriftelijk materiaal, foto's, schilderijen, audiovisueel materiaal). Dat betekent dat als we op dit moment een 'volledige tekst zoekoptie' zouden toevoegen, modern gedrukt tekstueel materiaal onevenredig veel aandacht zou krijgen.

Wegen naar een oplossing

Een van de manieren om de metadata te verbeteren is het aanmoedigen van de instellingen om ons de beste data die ze hebben te sturen. Tot nu toe deden we dat door een eigen formaat te specificeren en de aanleverende instellingen te vragen hun metadata in dat formaat aan te leveren. Vanwege de veelheid van dataformaten leidde dat in veel gevallen tot het platslaan van de originele metadata om het in het Europeana formaat te persen. ESE (Europeana Semantic Elements) is gebaseerd op Dublin Core met daaraan toegevoegd een aantal velden die specifiek zijn voor Europeana. Ondanks de ontegenzeggelijke kwaliteiten van Dublin Core gaat bij deze benadering veel van de oorspronkelijke rijkdom van metadata beschrijvingen verloren. Bovendien dreigde het Europeana dataformaat ESE 'yet another standard' te worden. Wij denken dat er al genoeg standaarden zijn. De oplossing ligt niet in het toevoegen van nog een standaard, maar in het vinden van manieren om bestaande standaarden met elkaar te verbinden.

Europeana heeft nu samen met de domeinen en de wetenschappelijke wereld een nieuw en veel flexibeler datamodel ontwikkeld, EDM (Europeana Data Model)[1], waarmee we de oorspronkelijke metadata in haar volle rijkdom kunnen opnemen. EDM is nadrukkelijk bedoeld als een intern dataformaat, niet als de nieuwe standaard. De instellingen leveren ons hun bestaande metadata in een door henzelf gekozen XML formaat aan. Uiteraard moeten we behalve deze oorspronkelijke data

een mapping file krijgen, om te vertellen welk veld uit de originele data naar welk EDM-veld afgebeeld wordt. Ook kan aangegeven worden welke velden in welke taal aangeleverd worden. Voor de grotere standaarden zullen we in samenspraak met de domeinen standaard mapping files beschikbaar stellen, zodat de hoeveelheid werk in het maken van de afbeelding geminimaliseerd wordt: alleen de eventuele lokale afwijkingen van en aanvullingen op de standaard moeten aangegeven worden. Waar mogelijk gebruikt EDM concepten uit bestaande datamodellen/ontologieën zoals SKOS, CIDOC CRM en Dublin Core.
EDM is gebaseerd op de principes van Semantic Data en geformuleerd in termen van eigenschappen, objecten en relaties. Dat stelt ons in staat om als we benoemde objecten (zogenaamde 'entities') zoals plaats- of persoonsnamen in de data herkennen, een verwijzing aan te brengen naar het concept in plaats van de tekstuele tekenreeks. Dat maakt het mogelijk om iemand die zoekt op Paris, resultaten uit Parijs te laten zien. Dan heb je natuurlijk altijd nog de mogelijkheid, dat eigenlijk gezocht wordt op de film 'Paris, Texas', op Paris de prins van Troje, of op Paris Hilton, mediafenomeen.

In de werkstroom van het inlezen van de metadata records gaat veel aandacht naar het normaliseren van gegevens. Bijvoorbeeld jaartallen en datums worden op heel veel verschillende manieren aangeleverd. Het is onwaarschijnlijk dat de duizenden instellingen die leveren aan Europeana ooit tot een enkel datumformaat zullen besluiten. Wat we wel kunnen doen is de binnenkomende informatie normaliseren, en de genormaliseerde versie naast de oorspronkelijke opslaan en gebruiken in zoekopdrachten.

Meertaligheid gezien vanuit de eindgebruiker

Voor de eindgebruiker van Europeana bestaat 'toegang in de eigen taal' uit drie elementen.

Ten eerste: het interface van Europena.eu en andere kanalen die we ontwikkelen moet in alle Europese talen beschikbaar zijn, zodat iedereen zijn of haar eigen taal kan gebruiken. Hiertoe hebben we een beperkte 'crowdsourcing' benadering gekozen: we hebben een speciaal CMS gebouwd dat mensen in het Europeana netwerk in staat stelt om bij te dragen aan die vertalingen, of ze te corrigeren.

Ten tweede wil de gebruiker zoeken in de eigen taal. Hierbij helpt het dat we vanuit de logbestanden kunnen zien dat gebruikers in ca. 70% van de gevallen zoeken op persoonsnamen, in ca. 20% zoeken op plaatsnamen. De hiervoor beschreven benaderingen van het identificeren van personen en plaatsen in de metadata en het linken naar externe bronnen om die persoon of plaats te beschrijven helpen daar al enorm bij. Voor de resterende zoekvragen proberen we de zoektaal te raden en de zoekvraag uit te breiden met vertalingen ervan.

Ten derde willen gebruikers de zoekresultaten in hun eigen taal kunnen lezen. Voor het vertalen van zoekresultaten zullen we bestaande webservices integreren zoals Yahoo BabelFish en Google Translate. De keus om te vertalen laten we bij de gebruiker, evenals de keuze welk van de beschikbare webservices hij of zij wil aanroepen.

Hoe ver we zijn

De productieversie van Europeana bestaat en kan op www.europeana.eu doorzocht worden. De zoekresultaten zijn al goed en kunnen nog veel beter worden met de hierboven geschetste technieken. Een aantal prototypes zijn al toegankelijk via het Europeana ThoughtLab[2].
In de komende maanden implementeren we de backend processen om EDM te kunnen inlezen, en experimenteren we met enkele instituten met het aanleveren van EDM. Uiteraard converteren we zelf de reeds aangeleverde gegevens naar EDM. Daartoe proberen we zoveel mogelijk relevante betrouwbare bronnen met persoons- en plaatsnamen en andere concepten te identificeren, zodat we die kunnen gebruiken om naartoe te linken. Het invoeren van EDM zal niet al onze uitdagingen rondom meertaligheid in een klap oplossen. Het zal ons wel de mogelijkheid geven aan die oplossingen te werken.

[1] *version1. europeana.eu/web/europeana-project/technicaldocuments/*
[2] *www.europeana.eu/portal/thoughtlab.html*

IMPACT
Centre of Competence voor massa-digitalisering

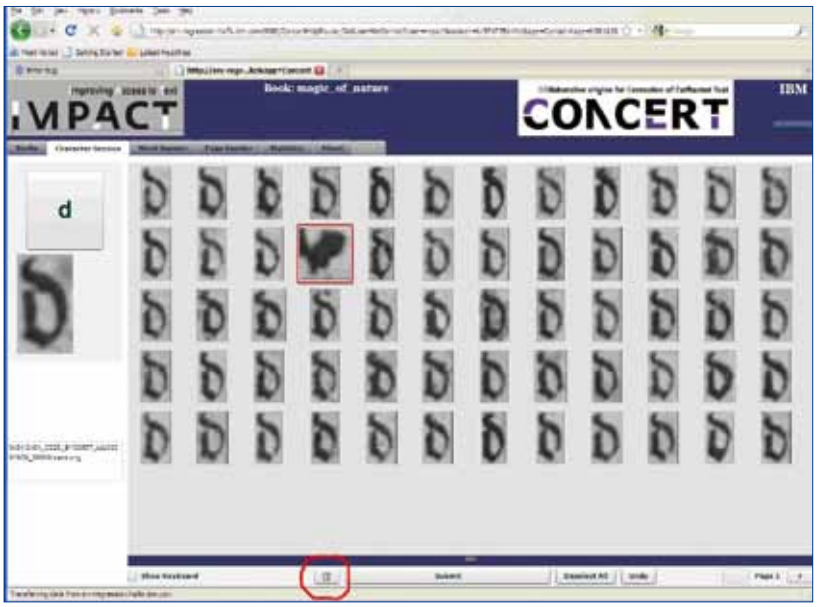
Steeds meer bibliotheken en archieven digitaliseren hun collecties. Zo heeft de Koninklijke Bibliotheek zich bijvoorbeeld ten doel gesteld om in 2013 10% van alle Nederlandse gedrukte publicaties digitaal aan te bieden. Om dit mogelijk te maken moeten miljoenen pagina's uit boeken, kranten en tijdschriften worden gescand en vervolgens omgezet naar tekst door middel van optische tekenherkenning (OCR). Op deze manier wordt het mogelijk de tekst te doorzoeken en te bewerken, net als bij teksten die van oorsprong digitaal zijn.

Hildelies Balk en Lieke Ploeger
Koninklijke Bibliotheek

Voor de digitalisering van historisch materiaal is de bestaande commerciële OCR-software echter nog niet geschikt: door de oude lettertypes, druktechnieken, beschadigingen en soms ingewikkelde lay-out zijn de resultaten vaak nauwelijks bruikbaar. Correctie achteraf is duur, tijdrovend en niet geschikt voor digitaliseren op grote schaal. Daarnaast is er nog de 'historische taalbarrière': het feit dat historische teksten door spelling- en andere variatie moeilijker doorzoekbaar zijn. Tot slot is de kennis over digitalisering ongelijk verspreid over instellingen, zowel binnen Nederland als in Europa, waardoor het wiel steeds opnieuw wordt uitgevonden.

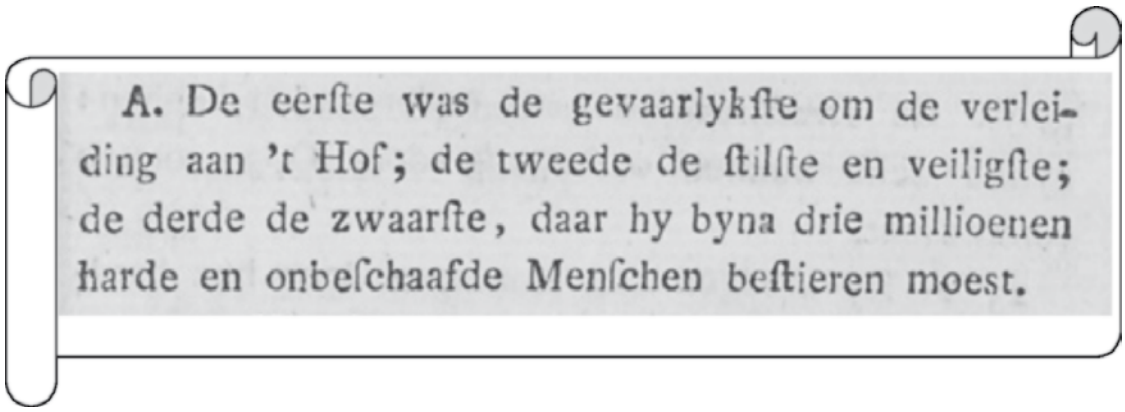
Vandaar dat in 2008 IMPACT (Improving Access to Text) van start ging, een grootschalig Europees samenwerkingsproject gefinancierd binnen het Zevende Kaderprogramma van de EC, dat dit soort problemen aanpakt. IMPACT wordt gecoördineerd door de Koninklijke Bibliotheek en uitgevoerd door een consortium van zesentwintig instellingen uit Europa, Israel en Rusland (nationale en regionale bibliotheken, universiteiten, onderzoeksinstellingen en bedrijven). Het doel van IMPACT is om de toegang tot gedigitaliseerde historische tekst aanmerkelijk te verbeteren door innovatie van OCR- en taaltechnologie. Ook wordt begin 2011 het 'IMPACT Centre of Competence' gelanceerd als centraal aanspreekpunt voor alle bibliotheken, archieven en musea die betrokken zijn bij de digitalisering van tekstmateriaal.

Allereerst wordt er binnen IMPACT gewerkt aan verschillende innovatieve methodes om de huidige OCR-software te verbeteren. De toonaangevende bedrijven IBM en ABBYY zijn bezig een 'zelflerende' OCR-machine op te zetten die zich automatisch aanpast aan elk nieuw boek dat gedigitaliseerd wordt. Onderdeel hiervan is een online Cooperative Correction applicatie, genaamd CONCERT, waarmee vrijwilligers uit heel Europa online OCR-resultaten kunnen verbeteren. Een veelvoorkomende OCR-fout is bijvoorbeeld dat de lettercombinatie 'r' en 'n' ('rn') als een 'm' wordt gezien. De CONCERT tool groepeert alle gevallen waarin een 'm' herkend werd op één scherm, waardoor gebruikers gemakkelijk de onjuiste letters kunnen markeren. Ook zal er een nieuwe en verbeterde versie van de ABBYY Finereader (versie 10) uitgebracht worden, waarin enkele nieuwe methodes, ontwikkeld door IMPACT, zijn opgenomen.



Het tweede probleem dat IMPACT aanpakt is de historische taalbarrière: diverse onderzoeksinstituten, waaronder bijvoorbeeld het Instituut voor Nederlandse Lexicologie (INL), ontwikkelen

Een voorbeeld uit 'Kort begrip der waereld-historie voor de jeugd' van J.F Martinet uit 1789.


A. De eerſte was de gevaarlykſte om de verlei-
ding aan 't Hof; de tweede de ſtilſte en veiligſte;
de derde de zwaarſte, daar hy byna drie millioenen
harde en onbeſchaafde Menſchen beſtieren moest.

Resultaat van de optische tekenherkenning aan het begin van het project:

A. De eerde was de gevaarlykflti om de verleiding aan 't Hof; de tweede de ftillie en veiligde; de derde de zwaarde, daar hy byna drie millioenen harde en onbefchaafde Menfchen beftieren moest.

Resultaat begin 2010:

A. De eerste was de gevaarlykste om de verleiding aan 't Hos; de tweede de stilste en veiligste; de derde de zwaarste, daar hy byna drie millioenen harde en onbeschaafde Menschen bestieren moest.

bronnen voor de ontsluiting van gedigitaliseerd materiaal, zoals computerwoordenboeken voor oud Nederlands. Ook worden er lexica gemaakt voor Named Entities: dit zijn specifieke namen van personen en plaatsen. Met speciaal door IMPACT ontwikkelde technieken wordt ervoor gezorgd dat de historische lexica zo efficiënt mogelijk gemaakt worden en dat een OCR-programma ze zo goed mogelijk kan gebruiken om betere OCR-resultaten te verkrijgen. Zo kan IMPACT er ook voor zorgen dat woorden in historische spelling (waereld) in een gedigitaliseerde tekst net zo makkelijk te vinden zijn voor gebruikers als moderne woorden (wereld).

Voor het Nederlands is IMPACT al een heel eind opgeschoten. Ook voor bijvoorbeeld 16e-eeuws Duits zijn al aardige resultaten behaald. Een van de grootste uitdagingen voor de laatste twee jaar van het project is om dit voor niet minder dan 7 Europese talen (Engels, Frans, Spaans, Pools, Tsjechisch, Bulgaars en Sloveens) ook voor elkaar te krijgen. Bij de technieken voor het bouwen van lexica is taalonafhankelijkheid een belangrijke vereiste: het doel is om tot een aanpak te komen voor toegankelijkheid en verbetering van gedigitaliseerde tekst die de specifieke talen overstijgt.
IMPACT zal tot slot het proces van massadigitalisering in Europa sterk verbeteren door de verspreiding van techniek, kennis en ervaring. Hiervoor worden verschillende strategische tools ontwikkeld zoals een website, helpdesk, decision support tools, een trainingsprogramma en een 'Centre of Competence', waarin de resultaten van IMPACT na afloop van de subsidie (2008-2011) beschikbaar blijven en verder ontwikkeld kunnen worden voor de gebruikers. Dit IMPACT Centre of Competence zal begin 2011 gelanceerd worden en zich ervoor inspannen dat het Europese culturele erfgoed sneller, beter en goedkoper online beschikbaar komt.

Tweetalige zoekservice voor de Erfgoedsector

In de erfgoedsector ontstaan steeds meer algemene zoekdiensten, die door verschillende instellingen te gebruiken zijn. Eén daarvan is de tweetalige zoekservice voor Fries Erfgoed. Deze service is zo opgezet dat alle Friese erfgoedinstellingen de dienst kunnen gebruiken voor het tweetalig zoeken in hun collectie - in het Fries en in het Nederlands.

Tigran Spaan
GridLine BV

De zoekservice is opgezet in opdracht van de Bibliotheekservice Fryslân (BSF). Projectleider Bertus Douwes hierover: 'De grote idee is een optimale ontsluiting van de collectie Fryslân via een meertalig trefwoordenthesaurus. Onder noemer 'Collectie Fryslân' worden via de provinciale Aquabrowser de gezamenlijke collecties ontsloten van Tresoar (Frysk Histoarysk en Letterkundich Sintrum), de basisbibliotheken van het Fries Bibliotheken Netwerk, Historisch Centrum Leeuwarden, Historisch informatiecentrum Noordoost Fryslân en de landelijke databanken.'

Het project vervult een belangrijke rol, niet alleen binnen de erfgoedsector maar ook als nieuwe stap in de Roadmap Fryske Taaltechnologie (Arjen Versloot - Fryske Akademy en Tigran Spaan, GridLine, augustus 2009). In dit kader werd al eerder een tweetalige zoekservice opgezet voor overheidsinformatie, met als klant de Provincie Fryslân. De Roadmap voorziet in verdere toepassing van de services, en ook in automatische classificatie en uiteindelijk een vertaalservice Fries-Nederlands.

Tweetalige Quabrowser voor Fries Erfgoed.

De software is niet alleen geschikt voor het talenpaar Fries-Nederlands, maar ook voor andere tweetalige situaties met een streektaal en het standaard Nederlands. En zelfs binnen een taal kan de technologie ingezet worden, bijvoorbeeld wanneer bezoekers met lekentermen zoeken die moeten worden vertaald in de terminologie van de erfgoedinstelling zelf. Bertus Douwes: 'De geautomatiseerde meertalige ontsluiting is landelijk inzetbaar. Uitgangspunt is dat beide talen beschikken over gedigitaliseerde woordenboeken. Op deze manier is het mogelijk om zowel landelijk als regionaal (taal) doelgroepen mogelijk te bedienen zoals Engels, Arabisch, enz.'

Daarnaast past het project in een serie initiatieven om door middel van webservices thesauri beschikbaar te maken voor de erfgoedsector. Voor het Rijksbureau voor Kunsthistorische Documentatie (RKD) voerde GridLine in 2006 een eerste project uit waarbij de Nederlandstalige Arts and Architecture Thesaurus beschikbaar werd gemaakt als webservice gebaseerd op SKOS (Simple Knowledge Organisation System), het standaard formaat voor de codering van thesauri. Dit project heeft recentelijk navolging gekregen bij Beeld en Geluid, zie elders in deze DIXIT.

Een vierde en laatste peiler van het project is de wens van de landelijke bibliotheekorganisaties om verder te vernieuwen; om deze reden werd het project ondersteund door bibliotheken.nl. Zoekservices met taalondersteuning maken deel uit van deze vernieuwing. Belangrijk voor de bibliotheken is dat deze services onafhankelijk zijn van de gebruikte zoekmachine. Daarom is gekozen voor een service die een tweetalige thesaurus levert. Deze kan worden gebruikt door iedere zoekmachine voor verbetering van de resultaten.

GridLine heeft hiertoe eerst een tweetalige thesaurus afgeleid. Basis hiervoor waren verschillende thesauri, waaronder de Gemeenschappelijke Onderwerpsontsluiting (GOO) van de Nederlandse wetenschappelijke bibliotheken, de woordenboeken van de Fryske Akademy, verschillende catalogi en de trefwoordtoekenningen hierin, en GridLine's taalmodules voor het Nederlands uit de GridLine TaalServer. Via de thesauruseditor GridWalker T konden tweetalige

domeinexperts het resultaat controleren en bijwerken. Deze thesaurus heeft Gridline vervolgens als proof of concept gekoppeld aan de AquaBrowser, in combinatie met een disambigueringsmodule, die bij zowel Nederlandse als Friese homoniemen de juiste equivalente vertaling geeft.

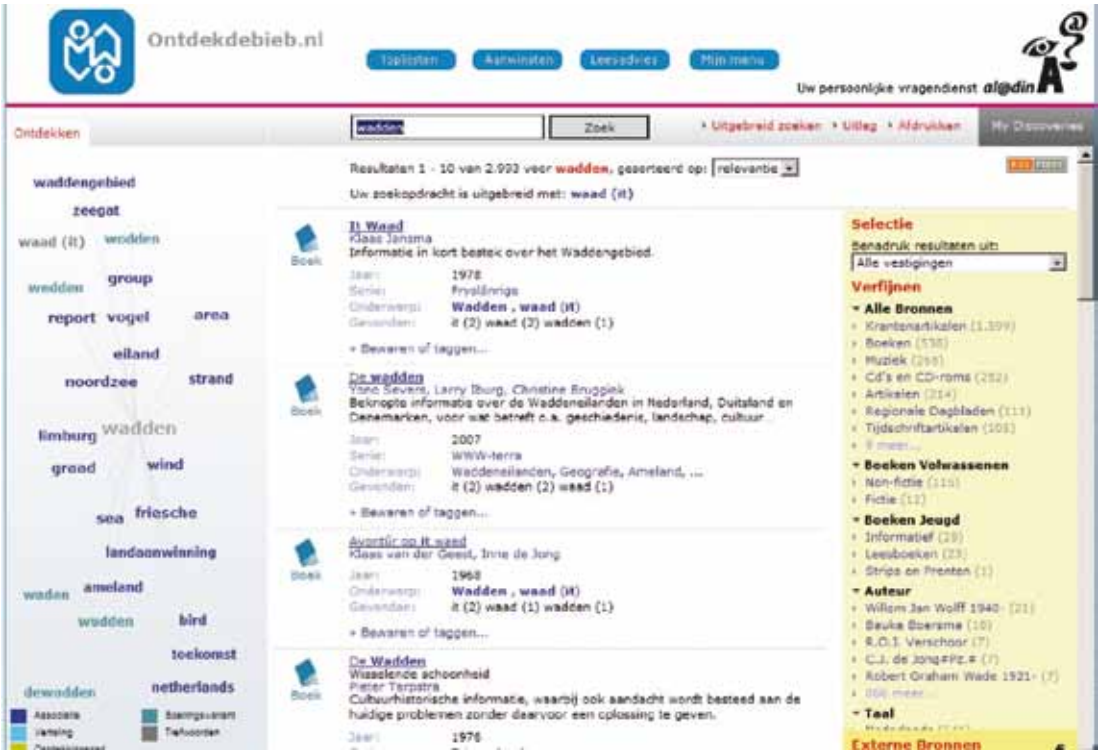
'De meertalige ontsluiting voor het Fries Erfgoed maakt alle titels zichtbaar die in collectie Fryslân betrekking hebben op het opgegeven zoekwoord', aldus Bertus Douwes. 'Dit ongeacht de taal waarin het zoekwoord is gesteld. Zo geeft het zoekresultaat van het zoekwoord 'kerk' niet alleen de Nederlandstalige titels over dit onderwerp, maar ook de Friestalige titels met het trefwoord 'Tsjerke'. Omgekeerd resulteert het zoekwoord 'Tsjerke' naast Friestalige titels ook in Nederlandstalige titels met het trefwoord 'Kerk'.'

'Voor de klant betekent dit dat het zoekresultaat alle materialen bevat die beschikbaar zijn via de provinciale zoekmachine, ongeacht de taal waarin de zoekvraag is gesteld. De klant wordt breder en beter geïnformeerd. Als we ervan uitgaan dat 50% van de gebruikers zowel Fries als Nederlands kunnen lezen, dan gaat het om de helft van de directe gebruikers van de voorzieningen, de leden, ruwweg 80.000 gebruikers. Maar de zoekmachine van Collectie Fryslân is ook voor niet-leden toegankelijk, dus in potentie gaat het om 50% van het aantal inwoners van de provincie Fryslân, zo'n 320.000 mensen.'

Voor de vertaalservice is binnen de provincie Fryslân belangstelling vanuit erfgoedinstellingen zoals het Fries Museum en de Fryske Akademy. Ook vanuit het project Foarwurd is belangstelling voor de vertaalservice. 'Foarwurd' is een initiatief gericht op het activeren van een website voor ouderen en kinderen die moeite hebben met lezen.

Een onderdeel van dit project is het schrijven van een boek in de streektaal, het Fries, met behulp van een 'vertaalkompjotter'. Voor dit onderdeel bestaat belangstelling in de vertaalservice. Gridline verwacht dat de service zal worden uitgebreid en dat meer Friese instellingen zich zullen aansluiten bij het initiatief.

Eén van de thesaurus-tools uit het GridWalker T-pakket.



Het belang van spraakweergave van online content

De meeste content op het internet is tekst en dat kan voor een grote groep mensen problemen opleveren. Veel apparaten waarmee je het internet op kunt (bijvoorbeeld smartphones), zijn vanwege het kleine scherm niet erg geschikt voor het weergeven van grote stukken tekst. Bovendien creëert tekst op internet problemen voor die 20% van de Westerse bevolking die moeite heeft met lezen: mensen met een visuele handicap, een forse taalachterstand of dyslexie. Een voorleesfunctie voor online tekst kan hier uitkomst bieden. Door de teksten voor te lezen wordt het de gebruikers makkelijker gemaakt de informatie te begrijpen en kunnen mensen die moeite hebben met lezen beter functioneren in onze informatiemaatschappij.

Jolet Jung
ReadSpeaker

Spraakweergave

De ReadSpeaker voorleesfunctie biedt spraakweergave van online content in meer dan 30 talen. Maar ook kan de voorleesfunctie beschikbaar worden gemaakt in een dialect. Zo heeft de gemeente Den Haag haar website toegankelijk gemaakt door een voorleesfunctie aan te bieden in zowel de Nederlandse taal als het Haagse dialect. Hiermee wordt de betrokkenheid van de burgers vergroot en verlaagt het tegelijkertijd de drempel, die sommige mensen kunnen ondervinden, om online informatie te zoeken.

De spraakweergave is direct beschikbaar, zonder dat de gebruiker een programma hoeft te downloaden, en hierdoor zeer toegankelijk. Niet alleen (mobiele) websites kunnen worden voorgelezen met de voorleesfunctie, maar ook online documenten en webformulieren. Zo kan de voorleesfunctie mensen helpen bij het invullen van formulieren en het lezen van documenten.

Spraakweergave in de Culturele Erfgoedsector

De 'Culturele Erfgoedwereld' maakt veel gebruik van tekst. Belangrijke documenten, verklaringen van gebeurtenissen en onderschriften bij allerlei soorten afbeeldingen vormen een essentieel onderdeel van veel collecties en zijn zonder spraakondersteuning voor mensen met een leesmoeilijkheid niet of nauwelijks toegankelijk. Daarom wordt de voorleesfunctie reeds gebruikt op talrijke websites, zoals die van het Rijksmuseum Amsterdam, The British Postal Museum & Archive, Exploring Surrey's past en het Goethe Instituut.

Embedded Highlighting

Ook de World Digital Library website heeft een stem gekregen dankzij de voorleesfunctie. De spraakweergave wordt hier ondersteund door een voorleescursor (Embedded Highlighting). De voorleescursor is een zeer nuttige optie voor tekstrijke websites waar toegankelijkheid van groot belang is, bijvoorbeeld voor mensen met leesproblemen. Onderzoek heeft uitgewezen dat het gebruik van een voorleescursor het 'begrijpend lezen' bevordert. De voorleescursor laat de tekst die voorgelezen wordt tegelijkertijd oplichten, op woordniveau en/of zinsniveau, afhankelijk van wat de bezoeker heeft ingesteld. Op de website van World Digital Library kan nu bijvoorbeeld worden geluisterd (of worden meegelezen) naar een korte beschrijving van de 'American Declaration of Independence'[1].

[1] *www.wdl.org/en/*

NOTaS

Stichting NOTaS
Secretariaat: Postbus 31070
6503 CB NIJMEGEN
T: 024-352 88 88
W: www.notas.nl
E: notas@malta-online.nl

NOTaS behartigt de belangen van bedrijven en kennisinstellingen die actief zijn op het terrein van Taal- en Spraaktechnologie (TST). Dit doet zij o.a. door middel van bijeenkomsten, lobbyactiviteiten en het tijdschrift DIXIT.

Deelnemers Stichting NOTaS

CLST
Postbus 9103, 6500 HD NIJMEGEN
Bezoekadres: Erasmusplein 1, 6525 HT NIJMEGEN
T: 024-361 16 86
W: www.ru.nl/clst
E: clst@let.ru.nl

Het Centre for Language and Speech Technology (CLST) onderzoekt en adviseert op het gebied van taal- en spraaktechnologie. Belangrijke speerpunten zijn robuuste ASR en de inzet van TST ten behoeve van het onderwijs en voor mensen met communicatieve beperkingen. CLST is tevens betrokken bij verscheidene projecten waarin audiomateriaal en teksten van cultureel erfgoed voor geesteswetenschappelijk onderzoek ontsloten worden.

Data Archiving and Networked Services (DANS)
Postbus 93067, 2509 AB Den Haag
T: 070-349 44 50
W: www.dans.knaw.nl
E: info@dans.knaw.nl

DANS is een instituut van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), dat mede wordt ondersteund door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). DANS zorgt voor de opslag en blijvende toegankelijkheid van onderzoeksgegevens in alle disciplines. Daartoe ontwikkelt DANS zelf duurzame archiveringsdiensten, bevordert het dat anderen dat doen, en werkt samen met databeheerders om ervoor te zorgen dat zo veel mogelijk data vrij beschikbaar komen voor gebruik in het wetenschappelijk onderzoek.

Dedicon
Postbus 24, 5360 AA GRAVE
Bezoekadres: Traverse 175
5361 TD GRAVE
T: 0486-48 64 86
W: www.dedicon.nl
E: info@dedicon.nl

Dedicon is in Nederland dé organisatie die lectuur en informatie toegankelijk maakt. Zo kan iedereen lezen wat hij wil in de vorm die hem past. Binnen de diensten van Dedicon speelt de moderne taal- en spraaktechnologie een belangrijke rol.

Dialogs Unlimites BV
Adres: Takkebijsters 17 sub 11, 4817 BL BREDA
T: 076-572 47 77
W: www.dialogsunlimited.com
E: info@dialogsunlimited.com
Dialogs Unlimited is de ontwikkelaar van de unieke ImmediateDialogs™ software suite, een complete product familie voor een geïntegreerde Self Service Delivery voor websites en spraakapplicaties, waarvan o.a. ImmediateVoice™ en de ivWebcommunicator™ deel uit maken. Dankzij de online ImmediateDialogs™ ontwikkelomgeving (SaaS) kunnen, zonder kostbaar specialisme in te huren of op te bouwen, in eigen beheer applicaties gebouwd en onderhouden worden.

Dutchear
Postbus 5050, 2600 GB DELFT
Bezoekadres: Brassersplein 2
2612 CT DELFT
T: 015-219 11 11
W: www.dutchear.nl
E: info@dutchear.nl

Dutchear is expert op het gebied van spraaktechnologie. De oplossingen van Dutchear maken processen en interfaces efficiënter, veiliger en gebruikersvriendelijker. Dutchear ontwikkelt en levert applicaties aan onder andere Justitie, Politie, KPN, Wehkamp, TNT en lokale overheden. Dutchear is een dochterbedrijf van TNO.

GridLine B.V.
Adres: Keizersgracht 520,
1017 EK AMSTERDAM
T: 020-616 20 50
W: www.gridline.nl
E: info@gridline.nl

GridLine is een succesvol IT-bedrijf dat gespecialiseerd is in taaltechnologie voor het Nederlands. Een aantal van onze producten en diensten: Klinkende Taal - plugin die ambtenaren helpt begrijpelijke brieven te schrijven; FAST; SharePoint; Lucene; NedLex - informatieplatform voor advocaten; de TaalServer - voor alle taalopgaven in uw bedrijf; opinion mining, information retrieval, kennismanagement. GridLine, voor professionals die het Nederlands gebruiken.

Knowledge Concepts
Adres: De Handboog 9,
5283 WR BOXTEL
T: 0411- 61 08 02
W: www.knowledge-concepts.com
E: sales@knowledge-concepts.com

Gedreven door passie en ervaring in dit vakgebied ontwikkeld en realiseert Knowledge Concepts vooruitstrevende systemen ten behoeve van informatieverwerking. Onze oplossingen onderscheiden zich in aspecten als: verzamelen, samenvoegen, uitlijnen, opslaan, ontsluiten en verspreiden van informatie en kennis. Dit realiseren wij met name door innovatief gebruik van taaltechnologie in combinatie met o.a. zoekmachines, document management systemen en portals.

OSTT/Sint Maartenskliniek
Adres: Hengstdal 3,
6522 JV NIJMEGEN
T: 024-365 97 18
W: www.ostt.eu
www.maartenskliniek.nl
E: ostt@maartenskliniek.nl

Het 'Ontwikkelcentrum voor Spraak- en Taaltechnologie ten behoeve van spraak- en Taalpathologie en revalidatie algemeen' is erkend als expertisecentrum door het Ministerie van het VWS aan de Sint Maartenskliniek. In het kader van het OSTT werkt de Sint Maartenskliniek samen met de afdeling Taalwetenschap van de Radboud Universiteit en met het UMC St. Radboud in Nijmegen.

Q-go.com B.V.
Adres: 'Diemercircle' Eekholt 40, 1112 XH DIEMEN
T: 020-531 38 00
W: www.q-go.nl
E: info@q-go.com

Q-go is een internationale aanbieder van Natural Language Search technologie. De natuurlijke-taaltechnologie wordt ingezet in applicaties waarmee onder andere banken, verzekeraars, telecombedrijven en logistieke dienstverleners hun online klantenservice verbeteren, hun kosten verlagen, en hun inzicht in de klant vergroten. Q-go wendt 40% van haar opbrengsten aan voor R&D.

ReadSpeaker
Adres: Dolderseweg 2A,
3712 BP HUIS TER HEIDE
T: 030 - 692 44 90
W: www.readspeaker.com
E: infoNL@readspeaker.com

ReadSpeaker is een internationale organisatie opgericht in Zweden, met inmiddels kantoren in Nederland, Duitsland, Engeland, Frankrijk, Spanje, de VS en Zweden. ReadSpeaker is marktleider op het gebied van het omzetten van tekst op het internet naar spraak via TTS, en richt zich met haar diensten op toegankelijkheid en gemak. Wij leveren innovatieve producten voor het voorlezen van websites (de eerste ooit), het toepassen van spraak voor andere web-based applicaties en RSS feeds en applicaties voor het gebruik op mobiele apparaten.

Telecats
Postbus 92, 7500 AB ENSCHEDE
Bezoekadres: Colosseum 42, 7521 PT ENSCHEDE
T: 053 488 99 00
W: www.telecats.nl
E: info@telecats.nl

Telecats ontwikkelt producten en diensten voor spraakinteractie en -analyse met een focus op telefonie. Met onze oplos¬singen voor Interactieve Voice Response en (open vraag) spraakherkenning worden telefoongesprekken (deels) geautoma-tiseerd. Met spraakanalyse kunnen inhoud en toon van live gesprekken en audioarchieven worden ontsloten. De taal- en spraaktechnologieën zijn als webservices beschikbaar voor derden die deze willen integreren in hun eigen propositie.

TST- Centrale, p/a Instituut voor Nederlandse Lexicologie
Postbus 9515, 2300 RA LEIDEN
Bezoekadres: Matthias de Vrieshof 2-3 2311 BZ LEIDEN
T: 071-527 24 95
W: www.inl.nl/tst-centrale
E: servicedesk@inl.nl

Bent u op zoek naar Nederlandstalige digitale taalmaterialen? Bij ons bent u aan het juiste adres! De TST-Centrale (Centrale voor Taal- en Spraaktechnologie) is de Nederlands-Vlaamse centrale voor beheer, onderhoud en distributie van Nederlandstalige taalmaterialen zoals tekst- en spraakcorpora, lexica en taalkundige tools. De taalmaterialen zijn veelal met overheidsgeld gefinancierd en worden door de TST-Centrale onderhouden en beschikbaar gesteld voor onderwijs, onderzoek en ontwikkeling.

Universiteit van Tilburg – DCI/Tilburg Centre for Cognition and Communication (TiCC)
Postbus 90153, 5000 LE TILBURG
Bezoekadres: Warandelaan 2, 5037 AB TILBURG
T: 013-466 91 11
W: www.uvt.nl
E: uvt@uvt.nl

In onderwijs en onderzoek van het Departement Communicatie- en Informatiewetenschappen (CIW) ligt het accent op aspecten van menselijke communicatie, zowel talige als niet-talige, inclusief mens-machine interactie, kennisrepresentatie, text mining, machine learning, kunstmatige intelligentie, computationele semantiek, semantische annotatie, en dialoogmodellen en -systemen. Het department omvat sinds September 2008 het Tilburg Centre for Cognition and Communication (TiCC). Het departement verzorgt onder andere de bloeiende opleiding Bedrijfscommuncatie en Digitale Media, de masteropleiding Human Aspects of Information Technology, en samen met de Radboud Universiteit Nijmegen de research masteropleiding Language and Communication.

Universiteit Twente - HMI
Postbus 217, 7500 AE ENSCHEDE
Bezoekadres: Drienerlolaan 5, 7522 NB ENSCHEDE
T: 053-489 91 11
W: www.utwente.nl
E: info@utwente.nl

De HMI-groep van de Universiteit Twente (UT) is een ondernemende onderzoeksgroep die zich op het gebied van mens-machine interactie zowel met fundamenteel als met meer toepassingsgericht onderzoek bezig houdt. Rondom de campus van de Universiteit Twente bevinden zich een groot aantal spin-off bedrijven waarvan enkele ook op enigerlei wijze bezig zijn met Taal & Spraaktechnologie en Mens-Machine Interactie.

Universiteit Utrecht – Utrecht institute of Linguistics OTS
Adres: Janskerkhof 13,
3512 BL UTRECHT
T: 030-253 60 06
W: www-uilots.let.uu.nl
E: uil-ots@let.uu.nl

Het UiL OTS is een onderzoeksinstituut binnen de Faculteit Geesteswetenschappen van de Universiteit Utrecht en stelt zich ten doel onderzoek te doen naar menselijke taal. De volgende drie dimensies bepalen het onderzoek: (i) de architectuur van het menselijke taalsysteem, (ii) de cognitieve systemen die ten grondslag liggen aan de verwerving en de verwerking van taal, en (iii) de wijze waarop taal wordt gebruikt in communicatie tussen mensen en tussen mens en machine.

Sponsoren/Samenwerking

Stichting DEN
Postadres: Postbus 90407, 2509 LK Den Haag
Bezoekadres: KB-complex, Prins Willem Alexanderhof 5, Den Haag
T: 070 – 314 03 43
F: 070 – 314 01 00
E: den@den.nl

Stichting DEN is het nationale kenniscentrum voor digitaal erfgoed. DEN bevordert en bewaakt de kwaliteit van digitalisering en digitale dienstverlening door de erfgoedsector. DEN investeert in kennisontwikkeling op het gebied van open ICT-standaarden en andere landelijke kwaliteitsprincipes voor duurzame digitalisering. Hiermee versterkt DEN de toekomstvastheid en de publieksgerichtheid van een nationale infrastructuur voor digitaal erfgoed.

CLARIN-NL
Postadres: UIL-OTS, Janskerkhof 13, 3512 BL Utrecht
Bezoekadres: UIL-OTS, Achter De Dom 24, k. 1.05, 3512 JP Utrecht
T: 030 - 253 6730
F: 030 - 253 60 00
E: clarinnl@uu.nl
W: www.clarin.nl

Nederlands Instituut voor Beeld en Geluid
Postvak BG55, Postbus 1060, 1200 BB Hilversum
Bezoekadres: Media Park, Sumatralaan 45, Hilversum
T: 035 - 677 16 52
F: 035 - 677 33 07
W: www.beeldengeluid.nl

Het Nederlands Instituut voor Beeld en Geluid is een cultuurhistorische instelling die het audiovisueel erfgoed dat uit historisch of cultuurhistorisch oogpunt van nationaal belang is verzamelt, conserveert en toegankelijk maakt voor zoveel mogelijk gebruikers.

Ondersteuning CumLingua Taal & Communicatie
Adres: Bronkhorstweg 48, 5363 TZ Velp (N.B.)
T: 0486 – 471 554
W: www.cumlingua.com
E: info@cumlingua.com
- Copywriting, Redactie & Vertalingen
- Taallessen
- Taalstijl- en communicatieadvies

Leonard Nijmegen bv
Adres: Groenestraat 294, 6531 JC Nijmegen
T: 024 - 378 26 43
W: www.leonard.nl
E: nijmegen@leonard.nl

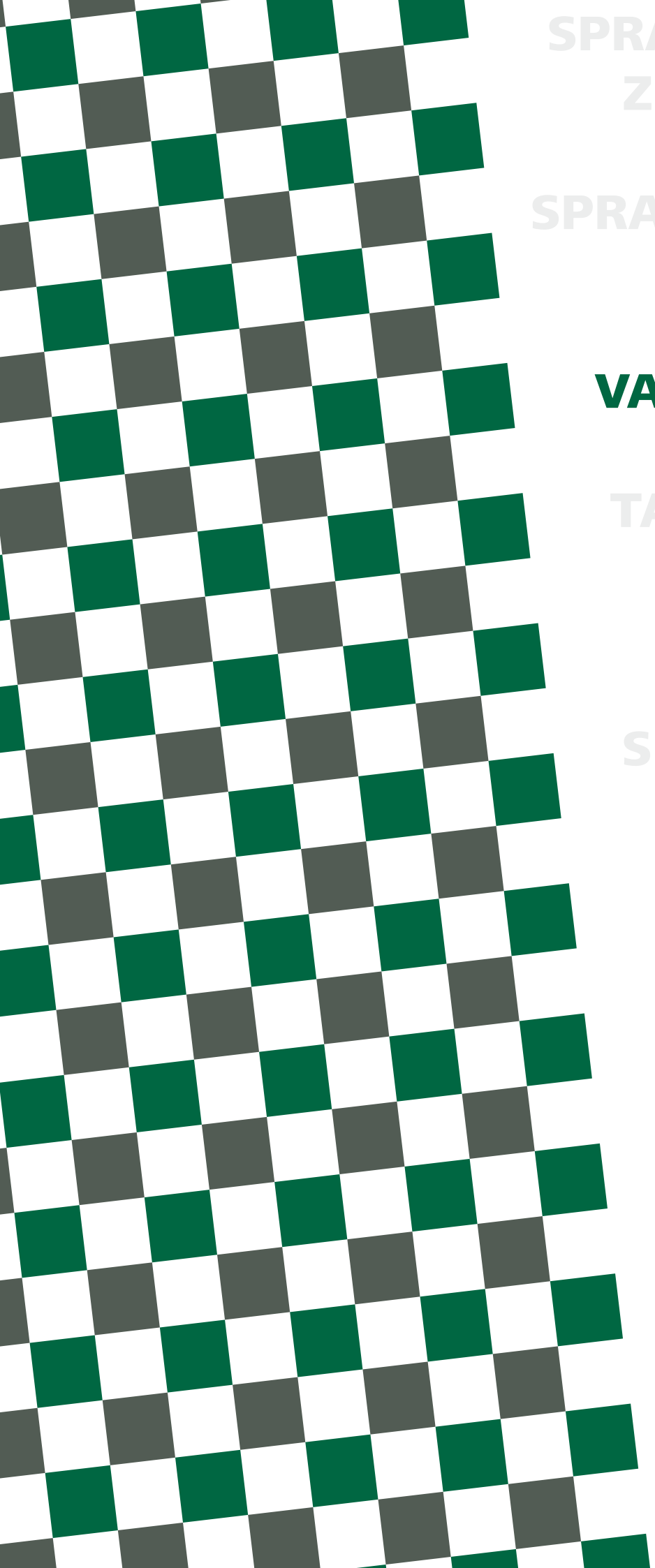
Bedenken, Maken, Doen
Leonard = communicatie+vormgeving+internet+drukwerk

Deykerhoff/Accountants
Postadres: Postbus 2325, 5202 CH, 's-Hertogenbosch
Bezoekadres: Rompertsebaan 70, 5231 GT, 's-Hertogenbosch
T: 073 - 623 24 50
W: www.deykerhoff.nl
E: denbosch@deykerhoff.nl

Deykerhoff/Accountants is een dynamisch accountantskantoor dat eigenzinnigheid paart aan betrouwbaarheid. Onze pro-actieve instelling is gericht op een grote betrokkenheid en een op maat gesneden persoonlijk contact met de klanten.

Malta & de Keyzer
Adres: Toernooiveld 300, 6525 EC Nijmegen
T: 024-351 21 08
W: www.malta-online.nl
E: welkom@malta-online.nl

Malta & de Keyzer, specialisten in office-management en secretariaat.



SPRAAKHERKENNING
ZEGT U HET MAAR
DIGITALISEREN
SPRAAKTECHNOLOGIE
DIALOOG
REPRESENTEER

VANZELFSPREKEND

VOICE RESPONSE
TAALTECHNOLOGIE
SPRAAKANALYSE
OPEN VRAAG
VERSTAAN
SPRAAKSYNTHESE
AUDIOMINING
BEGRIJPEN
EMOTIEDETECTIE
NASPREKEN
OPLIJNEN
LUISTEREN
HOREN
IVR
HERKEN
VOIP

TELECATS