Googelen door archieven Een revolutie in archief-toegang of sciencefiction?

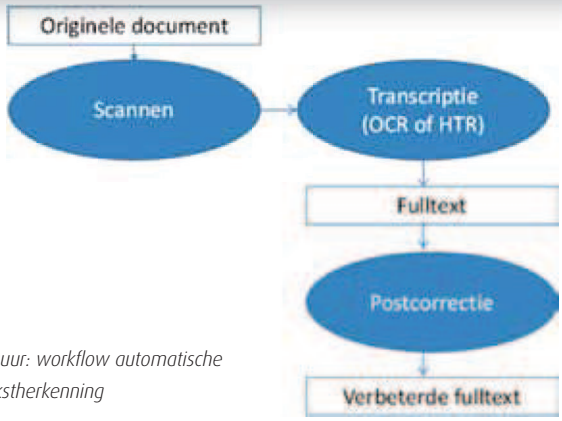
Edwin Klijn ■

Het is de droom van menig archiefgebruiker: vanachter je computer met een simpele muisklik zoeken door grote hoeveelheden historische documenten. Nog mooier: vanuit deze documenten worden doorverwezen naar andere gerelateerde stukken en collecties, achtergrondinformatie op Wikipedia, (historische) kaarten en tijdsbalken met gebeurtenissen. Ook veel onderzoekers likkebaarden bij het idee: handgeschreven kerkregisters uit de achttiende eeuw, omgewerkt naar spreadsheets waar je allerlei kwantitatieve analyses op kunt loslaten, reisverslagen uit de Gouden Eeuw die je op idioom kunt analyseren en een rijk scala aan nieuwe mogelijkheden.

Terug vanuit de wolken op de grond: we zien dat anno 2016 de digitale revolutie onmiskenbaar zijn stempel heeft gedrukt op de archiefsector. Toch zijn we nog maar net begonnen. Nog altijd is slechts een fractie van de archiefcollecties in Nederland gedigitaliseerd (8 procent ENUMERATE 2014).[1] Weliswaar zijn er steeds meer archiefbeschrijvingen online te vinden, maar zelden kan een gebruiker tot op documentniveau door archiefcollecties heen zoeken en zelden ook zijn er koppelingen tussen collecties gemaakt. Archiefinstellingen beschikken over een rijkdom aan "data" die nog nauwelijks gevonden kan worden in het digitale domein. Maar op het grensvlak van de Digital Humanities en de digitale collectie-ontsluiting zijn er ontwikkelingen gaande die daarin weleens verandering kunnen brengen. Technologie rondom automatische tekstherkenning van historische documenten lijkt de ontsluiting van archiefcollecties naar een hoger plan te tillen. Leverden de resultaten van de tot voor kort nog gebrekkige software teleurstellende resultaten op, in 2016 lijkt de automatische tekstherkenning voorzichtig de sprong van experimenteel naar operationeel te maken. Hoe werkt deze technologie, welke resultaten worden er gehaald en ten slotte: welke mogelijke gevolgen kan dit alles hebben voor het toegankelijk maken van archieven?

Automatische tekstherkenning

De workflow van automatische tekstherkenning is in grofweg drie onderdelen te verdelen: scannen, transcriptie en postcorrectie. Een scan maken van een tekstdocument maakt nog niet dat je door de tekst kunt zoeken. Om van een plaatje een machineleesbare tekst te maken (transcriptie), kun je gebruik maken van Optical Character Recognition (OCR)-software zoals bijvoorbeeld Abbyy Finereader (marktleider) of Tesseract (open source). Deze software analyseert de vorm van een letterteken of woord en vertaalt dit naar een machineleesbare variant. Vergelijkbaar met bijvoorbeeld Google Translate of Siri, maakt de software hierbij fouten. Deze foutmarge wordt doorgaans aangeduid als de Word Error Rate (WER), ofwel het percentage woorden dat incorrect wordt omgezet. Hoe hoger de WER hoe slechter de kwaliteit van de machineleesbare tekst. Software als Abbyy Finereader registreert overigens niet alleen de woorden, maar ook de coördinaten van deze woorden op de pagina en specifieke lay-outkenmerken (bijvoorbeeld de aanwezigheid van een afbeelding). Dit is handig om bijvoorbeeld de leesvolgorde te reconstrueren of later zoektermen op webpagina's te markeren. De OCR-kwaliteit die kan worden gehaald is sterk afhankelijk van de conditie van het originele materiaal. Automatische metingen van de Koninklijke Bibliotheek voor kranten uit het decennium 1990-2000 wijzen op een "word error rate" van circa 12 procent.[2] >>



Figuur: workflow automatische tekstherkenning

Pagina uit een set aantekeningen. Op de pagina staan beschrijvingen en afbeeldingen over de muis. (Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0)

>> Gepubliceerd tekstmateriaal is doorgaans gedrukt op stevig papier, met een duidelijk font en voorspelbaar formaat, rechte regels en een overzichtelijke lay-out. Anders is het voor archiefmateriaal. Recent onderzoek in het VOLAUTO-project (zie kader), met een steekproef met getypte documenten uit het Centraal Archief Bijzondere Rechtspleging (Nationaal Archief), kwam uit op een WER van 19 procent. Voor handgeschreven bronnen zijn de scores vaak aanzienlijk minder goed; de Handwritten Text Recognition (HTR)-software in het Huygens ING-project rondom de Resoluties van de Staten Generaal kwam voor een testset uit op een WER van maar liefst 68 procent.[3] Juist voor handgeschreven documenten valt men vaak terug op zelflerende software. De resultaten van het handmatig overtypen van een aantal documenten wordt hier geëxtrapoleerd naar de rest van een collectie. Zowel Transkriptorium in het READ-project als MONK (zie kader) maken gebruik van dergelijke technologie.
Wat kun je met machineleesbare tekst waarvan 1 van de 5 woorden incorrect is omgezet? In het geval van het Centraal Archief Bijzondere Rechtspleging bijzonder veel. Een fragment van een pagina met deze score:[4]

> Naam en voornamen: #ACHTERNAAM#,#VOORNAAM# Geboortep!aats en-datum: Zaandam, 29 October 1897 Echtgenoote van / Beroep: voorheen agent van Politie, t Laatste woonptaats en adres: Kanaalstraat 25 II Amsterdam Persoonsbewijs-no.:z 2 01239 afgegeven te Leeuwarden Nationaiiteit (evt. vroeger) Nederlander die ervan verdacht wordt: joodsche personen in macht van den vijand te hebben gebracht,terwijl hij in dienst was van de S.D. Terzake gehoord, verkiaarde verdachte mij het votgende: dat hij in dienst was getreden van Lippmann en Rosenthal voor de inventarisatie van joodsche goederen,vervolgens overgegaan naar de S.D.,afdeeling joodsche zaken te Amsterdam Verdachte heb ik, optastvan den Chef Opsporingsdienst D.P.M. op 28 Mei 1945 bewaring , togesteid, in het Huis van Bewaring I te Amsterdam P.O.D. Amsterdam. Mode! A

Met behulp van slimme software kun je hieruit namen, geografische locaties, datums, organisaties en domeinspecifieke termen (vijand, joodsche goederen, bewaring) halen en standaardiseren, hier zoektoegangen op bouwen en deze data koppelen aan andere collecties (linked data).
Wanneer er een machineleesbare tekst is, volgt er vaak nog een laatste verbeterslag: de zogenaamde post-correctie. Hier wordt de automatisch gegenereerde machineleesbare tekst door allerlei softwarematige bewerkingen verbeterd. Een van de bekendste open source software op dit gebied is TICCL (zie kader).

OCR is relatief goedkoop

OCR'en met standaard software is doorgaans niet heel prijzig; afhankelijk van de leverancier betaal je 0,5 tot 3 cent per scan. Vergeleken met de paginaprijs voor het scannen van archiefdocumenten (ongeveer 20-30 cent per scan) is dit een relatief bescheiden investering. Steeds meer leveranciers kunnen de machineleesbare tekst wegschrijven in ALTO (Analyzed Layout and Text Object)-xml-bestanden, waarin de positie van de woorden en de lay-out van de documenten wordt vastgelegd. Hiermee kunnen bijvoorbeeld treffers in de scan worden "ge-highlight". Postcorrectie en andere nabewerking is nu vooral nog maatwerk. De code is vaak vrijelijk beschikbaar, maar het vergt wel een handige programmeur om het te integreren in een digitaliseringsworkflow.

Zoekt en gij zult vinden?

Wanneer een archiefcollectie gescand is en machineleesbaar is gemaakt, is alles nog niet doorzoekbaar gemaakt voor de doorsnee webgebruiker. Vaak worden er zoeksystemen omheen gebouwd, die de grote hoeveelheden data indexeren en via een webinterface toegankelijk maken. Er bestaan veel verschillende zoeksystemen, die met een veelvoud van verschillende algoritmes de webgebruiker de weg wijzen. De argeloze webzoeker is zich hier niet van bewust, maar een zoeksysteem is heel bepalend voor wat men uiteindelijk vindt. Recent onderzoek van het Centrum Wiskunde en Informatica (Myriam Traub et al, 2016) naar de "retrievability bias" in het krantencorpus van Delpher illustreert dit uitstekend; slechts 2,7 van de 102 miljoen artikelen zijn 1 of meer keren bekeken door gebruikers.[5] Een beperkt aantal artikelen is heel veel bekeken, het overgrote deel van de 2,7 miljoen eenmalig. Een belangrijke

Tekening van een Burro. Gemaakt in Buitenzorg, Java in 1827 door Pieter van Oort. (Naturalis Biodiversity Center, Archief van de Natuurkundige Commissie voor Nederlands-Indië. Copyright: Public Domain Mark 1.0)

Project Volautomatische Archiefontsluiting

Dit pilotproject, mogelijk gemaakt door Archief2020 en BRAIN, is uitgevoerd in 2015-2016 door het Netwerk Oorlogsbronnen, het Nationaal Archief, het IMPACT Centre of Competence en het Centre for Language and Speech Technology (Radboud Universiteit Nijmegen). Aan de hand van een kleine testset van getypte en hybride documenten uit het Centraal Archief Bijzondere Rechtspleging (CABR) is onderzoek gedaan naar de kwaliteit van machineleesbare tekst (gegenereerd door Abbyy 11 software) en post-correctie (met TICCL en FROG). Op basis van een kleine test set van 89 documenten is er een word error rate gemeten van 19 procent (volgorde onafhankelijk). Het eindrapport bevat aanbevelingen voor configuratie en metingen van de resultaten. Zie eindrapport en documentatie: www.oorlogsbronnen.nl/volauto

MONK, Naturalis en uitgever Brill: Making Sense of Illustrated Written Archives

MONK, ontwikkeld aan de Rijksuniversiteit Groningen, is een geavanceerd systeem voor handschrift- en beeldherkenning. Met dit systeem en aanvullende contextuele informatie over diersoorten, plaatsnamen en habitats, zal worden geprobeerd om enkele negentiende-eeuwse dagboeken van ontdekkingsreizigers uit de collectie van Naturalis nader te ontsluiten. Het doel is een systeem te ontwikkelen dat de stukken uit het archief van de Natuurkundige Commissie doorzoekbaar maakt en contextualiseert met gerelateerde informatie. Zie: bit.ly/DagboekenNWOBrill

READ (Recognition and Enrichment of Archival Documents)

READ is een nieuw Europees Horizon2020-project (budget 8,2 miljoen euro), dat zichzelf ten doel heeft gesteld in de komende drie jaar tools te ontwikkelen die kunnen worden ingezet om vol- en semiautomatisch archiefdocumenten fulltext doorzoekbaar te maken en te transcriberen.
Wat het READ project uiteindelijk gaat opleveren is: publicaties, onderzoeksdata met “ground truth”-sets (volledig correcte datasets) en een “virtual research environment”, waar “software as a service” wordt aangeboden en testdata wordt gedeeld. Dit platform bouwt voort op Transkribus dat al is ontwikkeld in het Transcriptorium-project. Voor meer informatie over Transkribus: https://transkribus.eu/Transkribus/ en het READ-project: http://read.transkribus.eu/

TICCL (Text Induced Corpus Clean-up)

TICCL, ontwikkeld door Martin Reynaert vanuit het CLARIN-programma, is software die kan worden ingezet om machineleesbare tekst te verbeteren, onder meer door spellingcorrectie vanuit een historisch lexicon, een lijst van Bekende Historische Letterteken Verwarringen, ranking-features afgeleid uit het corpus en andere hulpmiddelen. Zie code (open source): https://github.com/martinreynaert/TICCL

verklaring hiervoor vormt de ranking (volgorde van presentatie) zoals Delpher deze hanteert; artikelen die door gebruikers zijn aangeklikt komen hoger in de ranking. En als ze hoger staan, worden ze ook weer sneller aangeklikt. Ook zoeksystemen als Google werken met dergelijke mechanismen, die niet of nauwelijks gedocumenteerd zijn maar wel heel bepalend zijn voor de vindbaarheid van gedigitaliseerd materiaal.

In ALTO (Analyzed Layout and Text Object)-xml-bestanden wordt de positie van de woorden en de lay-out van de documenten vastgelegd.

Kansen

Automatische tekstherkenning zet de deur open naar allerlei nieuwe manieren om archiefcollecties te ontsluiten. Zo zou je op basis van vorm- en/of lay-outkarakteristieken grote hoeveelheden documenten automatisch kunnen groeperen. Voor het CABR – een archief met een diversiteit van verschillende documenten variërend van foto’s, juridische documenten, vragenlijsten, lidmaatschapskaarten – heeft dit veel potentie. Ook interessant is het Famous Hands-experiment dat binnen het READ-project zal worden uitgevoerd: lukt het om op basis van de digitale registratie van de kenmerken een specifiek handschrift (Automatic Writer Identification) op te sporen in andere archieven?
Een interessante ontwikkeling is ook de integratie van automatische tekstherkenning in zoeksystemen. Hiermee kan zonder transcripties grote hoeveelheden scans doorzoekbaar worden gemaakt. Verbeteringen in de technologie vereisen dan niet dat er telkens nieuwe transcripties worden aangemaakt, maar kunnen meteen worden toegepast. Een systeem zoals MONK werkt nu feitelijk al op deze manier, in combinatie met zelflerende handmatige invoer om MONK slimmer te maken bij het herkennen van specifieke zoektermen. ■

Woordenlijst

- Word error rate (WER): percentage woorden dat incorrect is omgezet door de software.
- Ground truth-document: een volledig correct machineleesbaar document, gebruikt om metingen mee te verrichten.
- Named entity recognition (NER): het softwarematig herkennen van namen (personen, plaatsen, dingen, zaken) in digitale tekstcorpora.
- Optical Character Recognition (OCR): automatische tekstherkenning, waarbij woorden en leestekens softwarematig worden omgezet naar een machineleesbaar formaat.
- Handwritten Text Recognition (HTR): het (semi-)automatisch herkennen van handgeschreven tekstbronnen zoals dagboeken, registers, brieven, et cetera.

Dit artikel is mogelijk gemaakt door Archief2020, BRAIN, het ministerie van VWS, het VSBfonds en het vfonds. Met speciale dank aan het Nationaal Archief.

Noten

1 ■ http://www.den.nl/art/uploads/files/Enumerate-core-survey-NL2013-2014.pdf
2 ■ OCR-confidence level per decade, KB intern (ongepubliceerd).
3 ■ https://www.historici.nl/nieuws/geautomatiseerde-handschriftherkenning-tools-en-archieven-een-verslag.
4 ■ Namen, locaties en datums zijn aangepast in verband met privacy van betrokkenen.
5 ■ oai.cwi.nl/oai/asset/24651/24651B.pdf

Edwin Klijn ■ programmamanager Netwerk Oorlogsbronnen (www.oorlogsbronnen.nl).