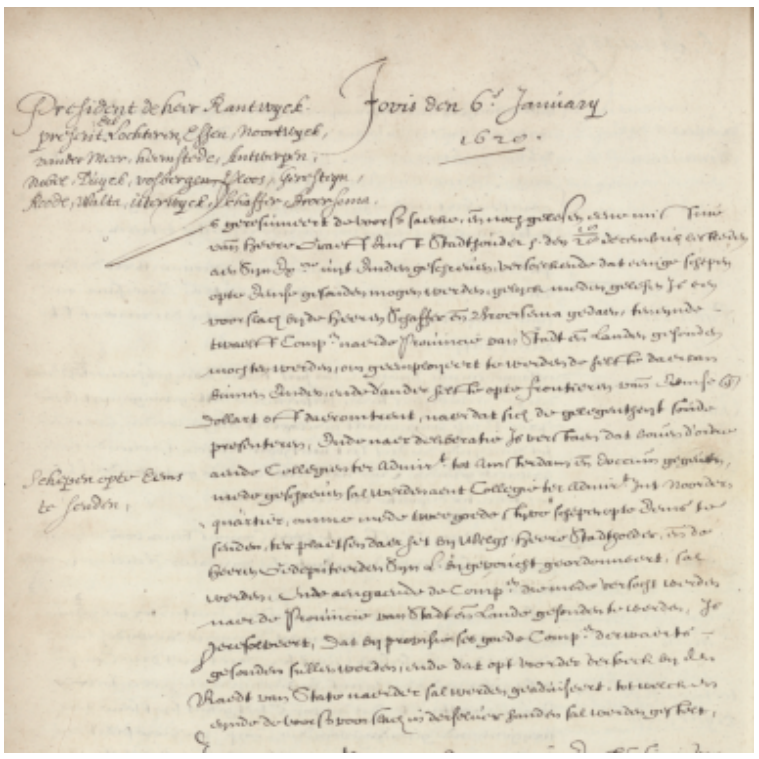
EU-project Transcriptorium ontwikkelt HTR-tool

Automatisch herkennen van handschriften weer dichterbij

Om historische tekstbronnen geschikt te maken voor digitaal gebruik, moeten ze machineleesbaar zijn. Het EU-project Transcriptorium experimenteerde met Handwritten Text Recognition. ***Edwin Klijn***



Een pagina uit de Resolutie van de Staten-Generaal *bron Huygens ING*

Voor handgeschreven teksten volstaat OCR-technologie niet. In het EU-project Transcriptorium (met als partners onder andere het Huygens ING, de Universiteit van Innsbruck en het Instituut voor Nederlandse Lexicologie) is drie jaar geëxperimenteerd met Handwritten Text Recognition (HTR). Deze technologie probeert interactief en voorspellend handgeschreven gedigitaliseerde teksten in machineleesbare teksten om te zetten. Op een workshop, gehouden op 27 november in Den Haag, werden de eindresultaten besproken én uitgeprobeerd, in een hands-on sessie rondom de transcriptietool Transkribus.

60% goed

Véronica Romero (Universitat Politècnica de València) introduceerde het Transcriptorium-project, waarbij HTR-technologie als test is ingezet op handgeschreven materiaal van de Engelse filosoof Jeremy Bentham en schrijfster Jane Austen. De resultaten laten zien dat al veel kan worden bereikt met HTR-technologie door het inzetten *prior knowledge methods*, zoals layout analyse, tekstregeldetectie en -extractie en lexical and language modelling.
Binnen het Transcriptorium-project is ook een pilot met HTR-technologie uitgevoerd op het handgeschreven deel van de Resoluties van de Staten-Generaal. Jesse de Does, computationeel taalkundige bij het Instituut voor Nederlandse Lexicografie (INL) legt uit: “De eerste resultaten waren schrikbarend: 68% van de woorden was incorrect! Na finetuning van de software slaagde men erin de Word Error Rate (WER) terug te brengen tot 40,4%. Het is een goed teken dat experts tijdens een workshops van het project zich vooral bekommeren om de WER. Want 40% fout kun je net zo goed zien als 60% goed.”

Meerdere toepassingen

Walter Ravenek (Huygens ING) vertelde hoe door toepassing van tools van onder meer de Stanford Natural Language Processing Group gedigitaliseerde corpora beter toegankelijk kunnen worden gemaakt op onder meer datum, geografische locatie en personen. Günther Mühlberger (Universiteit van Innsbruck) introduceerde de opvolger van het Transcriptorium-project: READ (Recognition and Enrichment of Archival Documents). READ richt zich sterk op de toepassing van HTR-technologie bij het digitaal toegankelijk maken van archiefcollecties. Het project wil nadrukkelijk oplossingen bieden die toepasbaar zijn op grote hoeveelheden documenten. READ gaat op basis van de Transkribus-tool verder bouwen aan een cloud-service waarin diensten worden aangeboden op het gebied van HTR, lay-outanalyse, document understanding en language modelling. Ook gaat er geëxperimenteerd worden met automatische handschriftherkenning (Famous Hands).

Enorme sprong mogelijk

De workshop Automated Handwritten Text Recognition liet zien dat een van de uitdagingen is om oplossingen te ontwikkelen die relatief goedkoop zijn en kunnen worden geïntegreerd in het productieproces van massadigitaliseringsstraten. Als men de beperkingen van de huidige technologie accepteert, is het mogelijk om met relatief kleine investeringen een enorme sprong te maken in het toegankelijk maken van archieven.

Edwin Klijn werkt bij het Nederlands Instituut voor Oorlogsdocumentatie (NIOD)
huygens.knaw.nl

GELEZEN

Beyond Open Access to Open Publication and Open Scholarship, John W. Maxwell (Simon Fraser University, Canada)
Dirk Roorda

Dit artikel maakt duidelijk dat het bij Open Access niet alleen om gaat dat het lezen van artikelen gratis wordt, maar dat het digitale paradigma een revolutie aan het bewerkstelligen is in de wetenschappelijke communicatie. Het artikel bevat een aantal catch-phrases die aangeven wat er aan de hand is.

Hier zijn er alvast twee:
1. Lenige wetenschap (Agile scholarship). Vroeger betekende publiceren dat een werk afgerond was en vervolgens openbaar gemaakt werd. Nu gebruiken groepen ook elkaars tussenresultaten, die dan wel openbaar moeten zijn. Zo is het afronden losgekoppeld van het openbaar maken.
2. Publiceren is publiek verzamelen (gathering an audience). In de digitale wereld is het een klein kunstje om iets openbaar te maken. De grote kunst is anderen zover te krijgen dat ze het aandacht geven. Publiceren is nu meer dissemineren geworden. Actief netwerken, met een zichtbare rol voor collega’s en het publiek. Al met al helpt dit artikel om de eigen (discipline-specifieke) activiteiten in een breder kader te zien.

http://src-online.ca/index.php/src/article/view/202

COLUMN

Zo eenvoudig is metadateren niet in de praktijk

Voor een historisch letterkundige studie die ik aan het schrijven ben, heb ik de afgelopen twee jaar een paar honderd boeken en artikelen moeten lezen. Ik las ze op papier en digitaal. Terugkijkend is het lezen van fotokopieën mij het slechtst bevallen. Fotokopieën ga ik te lijf met een potlood en markers in verschillende kleuren. Met het potlood maak ik aantekeningen in de marge, met de markers maak ik een samenvatting. Ik highlight eerst de grote lijn van het verhaal, plus passages die me om een of andere reden nuttig lijken. Vervolgens vat ik de highlights samen in een andere kleur.
Dat lijkt een redelijk efficiënt systeem, maar het komt erop neer dat je, als je iets wilt naslaan, de hele tijd in stapels kopieën aan het bladeren bent. Dan geef ik toch de voorkeur aan een boek – dat bladert makkelijker. Hoewel ik uiteindelijk vrijwel alles digitaal heb gelezen, zou ik de echte boeken alleen al om die reden niet willen missen.

Digitaal lezen doe ik op m’n iPad. Ik lees boeken het liefst in pdf-formaat, omdat je er van alles mee kunt. Je kunt een pdf bijvoorbeeld makkelijk dupliceren, zodat je een schoon exemplaar kunt bewaren naast een exemplaar om digitaal aantekeningen in te maken. Als ik maar een of twee hoofdstukken uit een boek nodig heb, verwijder ik de andere hoofdstukken uit het duplicaat.
De grootste winst van digitaal lezen zit wat mij betreft in de annotatiemogelijkheden. Ik gebruik daar een buitengewoon handige app voor – iAnnotate – waarmee je diverse soorten aantekeningen aan een pdf kunt toevoegen: beeld, geluid, teksten, hightlights in alle kleuren van de regenboog, uitroeptekens, vraagtekens, stempels – je kunt het zo gek niet bedenken.

Toen mij eenmaal duidelijk was welke onderwerpen ik in mijn studie wilde opnemen, ben ik mijn digitale bronnen gaan verrijken met metadata. Zo zette ik bij alle theologische verhandelingen bijvoorbeeld het woord ‘theotag’. Vervolgens kun je alle bronnen op dat woord doorzoeken – al dan niet via een index – wat veel tijd scheelt.
De theorie achter metadateren is relatief simpel, maar de afgelopen twee jaar heb ik ondervonden hoe weerbarstig de praktijk kan zijn. Om goede metadata te kunnen maken, moet je eerst patronen in je bronnen herkennen. Maar om heldere patronen in je bronnen te herkennen, moet je er eerst veel hebben gelezen. En moet je ze grondig hebben gelezen.

foto Leo van Velzen

Ik zou hier graag vertellen hoe schoon en helder gestructureerd mijn digitale bronnenverzameling eruitziet, maar dan zou ik liegen. Ik zweer nog altijd bij digitaal lezen en metadateren is echt buitengewoon handig, maar ik kom pdf’s tegen met highlights in vier kleuren, met inconsistente metadata en met aantekeningen die me ooit helder waren, maar die ik nu niet meer begrijp.
Echt heldere metadatering van letterkundige bronnen vraagt niet alleen zeer veel discipline, maar ook een flinke dosis helderziendheid. Pas als je van tevoren weet wat je in die bronnen gaat aantreffen, kun je je onderzoek beginnen met een consistente, heldere set metadata, verankerd in een gedegen theoretisch kader.
Wie weet hoe je dat aanpakt met historisch letterkundige bronnen waar tot nu toe nauwelijks onderzoek naar is gedaan, moet het mij een keer uitleggen.

Ewoud Sanders
Taalhistoricus en journalist. Sanders is vaste medewerker van onder meer NRC Handelsblad en Onze Taal.