Googelen door archieven, droom of werkelijkheid?

“Googelen door archieven is een droom en wens van gebruikers.” Met die zin opent Irene Gerrits (directeur Collectie & Publiek Nationaal Archief) de studiedag ‘Googelen door archieven’ bij het Nationaal Archief op 13 oktober 2016. Bronnenbeheerders krijgen steeds meer te maken met de vraag ‘digitaliseren, en dan?’. Beschikbaarheid van digitale bronnen is niet genoeg, toegankelijkheid is gewenst. Er zijn veel vragen en nog weinig antwoorden. Tijdens de studiedag deelt de sector kennis over projecten die een oplossing kunnen bieden. “Veel inspiratie gewenst, met de droom om te kunnen googelen door archieven in het vizier”, aldus Gerrits.

Dagvoorzitter Charles van den Heuvel (Huygens ING) start met een dankwoord aan organisatoren, een overzicht van zijn project Golden Agents en het programma. Vragen die centraal staan zijn: Hoe kunnen we handmatige en automatische technieken combineren? Hoe gaan we om met de verschuiving van metadata op (sub)collectieniveau naar fulltext doorzoeken van individuele documenten? En wat betekent dat voor de praktijk, voor het werk van onderzoekers en collectiebeheerders? Hoe gaan we het materiaal beschikbaar stellen? De gecombineerde Muller- en Fruinzaal zit vol en het project volautomatische archiefontsluiting is als eerste aan de beurt.

Ook Optical Character Recognition software heeft voorkeuren

Anne Gorter (Nationaal Archief) en Edwin Klijn (Netwerk Oorlogsbronnen) presenteren hoe zij de afgelopen maanden met standaard Optical Character Recognition (OCR)-software gescande documenten volautomatisch toegankelijk hebben gemaakt. De resultaten van de pilot, met een testset uit het Centraal Archief Bijzondere Rechtspleging (Nationaal Archief), zijn veelbelovend. Gorter laat een dossier zien uit dit archief dat zo’n vier kilometer beslaat. Het is een verzameling papier van verschillende kwaliteit, dikte, kleur, formaat, met uitgelopen inkt, typefouten in het document en verschillende opmaken.

Anne Gorter (foto: Anne Reitsma)

Een aantal bevindingen: Documenten scannen tegen een zwarte achtergrond geeft het beste resultaat. Vervolgens zijn de scans voorbewerkt door ze bijvoorbeeld recht te zetten en te ruime kaders weg te snijden. Daarna gaat de OCR-software aan het werk. Wat blijkt, maar liefst 81% van alle 30.000 woorden zijn correct door de computer omgezet.

Nu Gorter uit de doeken heeft gedaan waar OCR niet van houdt, is de beurt aan Klijn om te vertellen waar OCR wél van houdt. Een eenvoudige lay-out bijvoorbeeld of mooie rechte regels. “Door ground truth documenten – handmatig overgetypte documenten die dus het 100% gewenste resultaat geven – naast de ge-OCR’de documenten te leggen zie je de verschillen. En als vier van de vijf woorden dan goed zijn gescand, wat kun je er dan mee? Heel veel. In geval van het informatierijke CABR-archief kun je al redelijk nauwkeurig personen, organisaties, geografische locaties en datums eruit halen”, aldus Klijn.

Met tools als Frog en Ticcl zijn Klijn en Gorter vervolgens de OCR gaan verbeteren, ofwel het toepassen van 'post-correctie' door er bestaande lijsten tegenaan te houden, zoals GeoNames voor geografische locaties. Maar de post-correctie zit nog in een experimentele fase. Meet-methodieken zijn nog erg in ontwikkeling en worden nu nog veel beïnvloed door menselijke interpretatie. Klijn: "Sommige tools worden gelukkig doorontwikkeld. Het zou mooi zijn om ze domein-specifiek het te trainen. Door bijvoorbeeld lijsten toe te voegen zoals de WO2-thesaurus die in het Netwerk Oorlogsbronnen wordt ontwikkeld". En dan een voorbeeld uit de praktijk. De software haalt meer locaties uit het machine-leesbare bestand dan in het ground truth bestand staan. Aan de andere kant kan de software de datum er helemaal niet uithalen, terwijl die wel in ground truth staat. De zaal lacht om het voorbeeld 'Den Helder': "Doet de software het goed en selecteert het de plaats in Noord-Holland, of de persoon Glenn Helder, een voetballer?".

Edwin Klijn (foto: Anne Reitsma)

Terug naar de hoofdvraag 'is het zinvol om te OCR-en en nader toegankelijk te maken met Named-entity recognition (NER)?' Klijn: "Ja het is zinvol. Als je volautomatisch vier van de vijf woorden correct hebt, kun je heel veel...! Dan kun je kilometers aan archief op documentniveau toegankelijk maken. Archiefdocumenten waarvan inhoudelijk nog niet veel bekend is kun je op basis van vorm- en beeldherkenning groeperen. Dan is onderzoek op een heel andere manier mogelijk: vergelijkingen, statistische analyses, nieuwe onderzoeksvragen en koppeling met andere informatiebronnen. Er gaat een hele nieuwe wereld open met deze technieken en met het verbinden van collecties. Genoeg kansen!"

Terecht komt de vraag uit de zaal om welke kosten het gaat bij de toepassing van deze technieken. Het OCR-en kost ongeveer 2 cent per pagina en de kosten van Named-entity recognition zijn nog niet helder omdat de techniek nog volop in ontwikkeling is. Klijn sluit af: "Deze technieken zijn goed voor het beter doorzoekbaar maken van getypte of hybride documenten die voldoen aan bepaalde kenmerken. Voor handgeschreven materiaal is dit alles minder geschikt. Dat is een wezenlijk verschil".

Maandelijkse records in machine learning met MONK

Prof. dr. Lambert Schomaker (RUG Groningen, ALICE Instituut) en Andreas Weber (Universiteit van Twente) presenteren de casus van automatische handschriftherkenning in 19e-eeuwse dagboeken in het MONK-systeem. Dit systeem biedt een ontsluitingsmethode voor archieven – waaronder handschriften – uit verschillende culturen en talen en doet 24/7 aan 'machine learning'. Het bevat een omgeving voor de opslag en annotatie van gescande documenten én herkennings- en zoekalgoritmen. Momenteel bevat het MONK-systeem 400 documenten en 70.000 scans.

Lambert Schomaker (foto: Anne Reitsma)

"Ons doel is vrij ambitieus, een Europese Google for handwriting", aldus Schomaker

MONK focust op het 'wat'. Het OCR-en van documenten gaat over het herkennen van woorden (keyword spotting). Handschriften geven dan grote problemen. Want je hebt te maken met gesuggereerde letters, contractie, etc. Kortom, een handschrift! Hoe moet OCR daar karakters in herkennen? Schomaker laat een voorbeeld zien van een handschrift waarbij het herkennen van karakters lastig is: "Kijk eens, er hangt een fraaie doch storende krul van een d uit de toekomst boven de m!". OCR is dus niet werkbaar voor hem. MONK belooft dan ook geen transcriptie maar is er vooral op gericht de doorzoekbaarheid van handgeschreven materiaal te verbeteren.

Wat erg helpt is het labelen van woorden door mensen. Moeilijk materiaal wordt beter vindbaar als het gelabeld wordt. Denk daarbij aan het handmatig aangeven dat een 'datum' een 'datum' is. Zodat de computer bij het volgende document de datumaanduiding zelf als zodanig eruit kan halen. "Elke maand worden er records gebroken in het machine learning", zegt Schomaker met enige trots. En visuele woordmodellen worden hergebruikt. "Het fascinerende is dat met redelijk weinig input zo redelijk snel een index wordt opgebouwd waarmee je een basis creëert voor transcriptie", aldus Schomaker. Klijn stipte het in de ochtend ook al aan. Dat verschil, tussen het verbeteren van de doorzoekbaarheid en het antomatisch generen van een transcriptie, is belangrijk om aan te geven.

Weber neemt ons mee in de praktijk met een project waarbij documenten van de 'Natuurkundige commissie van Nederlands Indië' uitdagend materiaal vormen voor ontsluiting met behulp van het MONK-systeem. Doel van dit project is de ontwikkeling van een infrastructuur en het open toegankelijk maken van het archief. De uitdagingen liggen in het maken van links tussen materialen, de mix van talen, schrijvers en stijlen (soms op één pagina) en het gebruik van tekst en beeld door elkaar heen.

Andreas Weber (foto: Anne Reitsma)

Hoe is het om met MONK te werken? We krijgen een inkijkje in het label-proces. Opvallend is dat een pagina automatisch wordt gesplitst in lijnen maar labels vervolgens aan individuele woorden worden toegekend. Het verschil met tekstherkenning door een mens, is dat MONK hier volgens Weber aanzienlijk beter in is. Maar een expert die bekend is met het domein kan tot nog betere resultaten komen. En tenslotte weer een vraag over de kosten. Schomaker: "Er

is interesse vanuit het bedrijfsleven. De infrastructuur heeft ook wel langdurige bekostiging nodig. Het leveren van continuïteit en een goed businessmodel best lastig".

De Wolpertinger Transkribus

Een handgeschreven archiefdocument fotograferen en het automatisch getranscribeerd terugvinden in een app. Het lijkt de toekomst maar komt steeds dichterbij! "The Netherlands and England are pioneering when it comes to the usage of these techniques. And a lot of work has been done for me", zegt Günter Mühlberger. Hij stelt vast dat machine learning, of het werken met neural networks, in verschillende domeinen begint te overheersen. Hij is dan ook blij dat steeds meer archieven hun collecties digitaliseren. Dit zijn collecties die meestal heel weinig of nog niet bekeken zijn; archieven zijn 'booming'. Tot slot zijn Digital Humanities (big) data driven, dus zij hebben materiaal nodig om mee te werken.

Transkribus is een multi-inzetbaar platform (een 'virtual research environment') en een transcriptie tool. Het is een infrastructuur voor archieven, studenten uit de geesteswetenschappen, leveranciers van technologieën en het publiek (vrijwilligers). De geavanceerde transcriptie-tool wordt door Mühlberger vergeleken met een Wolpertinger; een Duits fabeldier waarvan het lichaam een samenstelling is van verschillende dieren. "Transkribus is also a creature that can do anything!", zegt hij.

Mühlberger over vrijwilligers: "Why do we call it 'crowdsourcing'? Who wants to be part of a crowd? Let's talk about people who are interested in the content and become experts while working with it. The point is to let people see how valuable their efforts could be, so they'll get involved in these projects. I believe in expert sourcing!".

Günter Mühlberger (l) en Charles van den Heuvel (r) (foto: Anne Reitsma)

Wat kan een archiefonderzoeker, of een archief z'n onderzoeker bieden met Transkribus? In een private omgeving werken aan de transcriptie van (zelf) geüploade documenten. Dat kan indien gewenst met duizenden tegelijk met een File Transfer Protocol (FTP). Zo kan een onderzoeker 'shoppen' naar documenten bij verschillende archiefinstellingen en er uiteindelijk mee werken in één omgeving. De transcriptie vindt vervolgens op regelniveau plaats (niet woorden). En door de scan en de getranscribeerde tekst te linken kan de automatische handschriftherkenning ingezet worden.

Dat kan al na handmatige transcriptie van zo'n vijftig pagina's. Er is dan doorgaans genoeg training data voor de computer. "Een beperking van de Text Recognition tool is bijvoorbeeld de behoefte aan woordenboeken. Hoe meer woorden er gedefinieerd zijn in de tool hoe beter hij kan herkennen". Als een tekst automatisch getranscribeerd is worden de resultaten gemeten aan de hand van de 'word' (WER) en 'character error rate' (CER). Dat geeft het percentage foutief omgezette woorden of karakters aan. De WER ligt volgens Mühlberger het

dichts bij de realiteit. Na een correctieslag zou een document full text doorzoekbaar zijn, en deelbaar met anderen.

Tot slot geeft Mühlberger ons een kijkje in de toekomst. Wat kunnen we nog verwachten? Onder meer een tabel editor waarmee voorgedrukte modellen in administratie bijvoorbeeld opgenomen kunnen worden. Of een E-learning interface, waarmee mensen kunnen leren transcriberen. En een scan-app, waarmee gefotografeerde documenten worden geüpload naar Transkribus en automatisch getranscribeerd. De Transkribus-tool wordt momenteel verder uitgebouwd in het Europese Horizon2020 READ-project (Recognition and Enrichment of Archival Documents).

Pitch je archief

Na een goede lunch staan er vier collega's klaar om hun collectie te pitchen voor de heren Mühlberger en Schomaker. Ieder krijgt vijf minuten om de eigenschappen van zijn archief toe te lichten en waarom juist die collectie geschikt (of ongeschikt) zou zijn voor het automatisch ontsluiten via het Transkribus of MONK-systeem. Het geeft een praktische inkijk op de beide systemen. Weber trapt af met veldnotities uit eind 18e / begin 19e-eeuw, onderdeel van de collectie van Naturalis. "Het is een prachtige collectie die aan historici laat zien hoe biologen de natuur destijds documenteerden en aan biologen een unieke kijk op Zuidoost-Azië in die periode geeft", vertelt hij.

Maar het is een rommelig geheel voor een computer. "Sloppy" geschreven, met regelmaat op doorschijnend papier, verschillende auteurs op één pagina aan het werk, veel gebruikte dierennamen zijn veranderd en last but not least is op veel pagina's tussen de tekst getekend. Mühlberger ziet geen match voor Transkribus: "De mix van talen zou de tool aankunnen, mits door dezelfde auteur geschreven. De leesvolgorde is wel een probleem, net als de later ingevoegde woorden in zinnen. Transkribus kan daar op dit moment niet mee omgaan".

Publiek (foto: Anne Reitsma)

Schomaker ziet het positiever: "De tabel is het belangrijkste voor de bioloog, maar het moeilijkst te lezen voor de machine. Belangrijk is dat MONK de pixels eerst leest en dan de taal als een saus erover ziet. Voor dit systeem zijn de verschillende talen en schrijvers dus ook geen probleem. Daar kun je modellen voor maken. En het probleem met het doorschijnende papier kan opgelost worden door de achterkant ook te scannen en de doorgedrukte inkt te wissen". Hij drukt Weber nog wel op het hart dat de term 'sloppy' niet toepasselijk is, omdat het voor de computer niet uitmaakt of iets slordig of netjes geschreven is, zolang de stijl maar eenduidig is.

De computer geeft de oplossing bij grote hoeveelheden

"Ik had ambitieuzer kunnen zijn nu ik de mogelijkheden van MONK en Transkribus gezien heb en er meer tabellen in kunnen gooien", start Erik Odegard (Universiteit Leiden) zijn

levendige pitch van 2,5 kilometer 17e-eeuws VOC-archief. Schomaker en Mühlberger zijn het erover eens dat het prachtig materiaal betreft en dat met zo'n hoeveelheid bijna de enige mogelijkheid voor een goede transcripties de computer is.

Klijn pitcht een set van honderden dagboeken, beschreven tijdens de Tweede Wereldoorlog. Deze staan online maar zijn niet doorzoekbaar en hij vraagt zich af wat hij kan doen om de collectie machine leesbaar te maken. Het dagboek van Toby Vos is bijvoorbeeld prachtig geïllustreerd en ze schrijft in een hele regelmatige en voorspelbare stijl. Etty Hillesum schreef ook heel keurig. Het manuscript van Hitzerus Mees is daarentegen rommelig. "Daar zou ik niet aan beginnen, dan kunnen we onze tijd beter besteden", antwoord Schomaker dan ook op de laatste. Maar transcriptie van het dagboek van Hillesum ziet hij wel zitten, net als zijn collega Mühlberger. Hij concludeert op deze pitch dat "handschriften die voor mensen moeilijk zijn om te lezen, dat ook voor computers zijn".

Nico Vriend (foto: Anne Reitsma)

Nico Vriend (Noord-Hollands Archief) ziet uitkomst in de toegankelijkheid van indexen. Door het automatisch ontsluiten van vier meter aan indexen zou honderdveertig meter aan archiefmateriaal van het Ministerie van Koloniën (1910-1919) beschikbaar komen. Een efficiënte redenatie. 'Can Handwritten Text Recognition reveal the unseen?' Deze indexen zijn opgesteld in een voorgedrukte tabel. Dat kan voor- en nadelen opleveren.

Schomaker ziet geen heil in het handmatig ontleden van de tabelstructuur omdat het 't zoeken door de inhoud niet beter maakt. Mühlberger is het daarmee eens. "De woorden kun je goed ontsluiten dus ik zou me niet richten op de betekenis van de tabelstructuur. Waar ik wel nog vraagtekens bij heb is het koppelen van de index aan de documenten waar ze naar verwijzen. Met de manier waarop de inventarissen er nu uit zien zie ik dat niet voor me. Maar die boomstructuur kan ook losgelaten worden met de mogelijkheid full tekst te doorzoeken".

Amerikaans recht en de Nederlandse digital born archieven?

Mette van Essen (Nationaal Archief) neemt ons in de loop van de middag mee in de wereld van E-Discovery. Dit is een methodiek waarbij machine learning technieken worden ingezet bij de automatische classificatie van ongestructureerde data. Het komt oorspronkelijk voort uit het Amerikaanse recht, waarbij rijke partijen in rechtszaken elkaar overladen met terabytes waar een vorm van bewijs in gevonden kan worden. "Kennen jullie die Amerikaanse series of films waarbij advocaten zichzelf dagen opsluiten tussen tientallen dozen met papier om die ene speld in de hooiberg te vinden waarmee ze kunnen winnen? Dat is dus volledig achterhaald en doen ze alleen nog zo omdat het er aantrekkelijker uitziet op beeld", aldus Van Essen.

Uitgangspunt bij E-Discovery is het weggooien van voor jou onbruikbare informatie. Hoewel steeds meer vergelijkbare technieken gebruikt worden, verschilt dat van de omgang met 'big data'. Daarbij wordt naar inzicht in een grote hoeveelheid data gezocht. Binnen het informatiemanagement zoeken we naar bruikbare onderdelen van het E-Discovery proces. Met name in de waardering van informatie voor hergebruik. De groeiende digitale werkprocessen in aantal en hoeveelheid geven ruis en is voor het Nationaal Archief

aanleiding om dit onderzoek te doen. Van Essen laat wat berekeningen zien: “Als ik handmatig al mijn mails moet keuren voor langdurige bewaring, ben ik zestien jaar bezig”.

Waardering van de correspondentie door medewerkers zelf is een traditionele en onwerkbare oplossing. Technologie aan het begin van het proces kan de helpende hand bieden. Door automatische selectie blijft dan nog maar veertig of zelfs tien procent van het totaalaantal mails over om handmatig te keuren. “Het moeilijkste is echter om data te krijgen om mee te testen. E-mails zijn natuurlijk privacygevoelig”, aldus Van Essen. Daarom wordt een testset van het Nationaal Archief als use-case gebruikt. Schomaker ziet de noodzaak in selectie niet: “Opslag wordt alleen maar goedkoper, bedrijven zullen niet begrijpen wat jullie probleem is”. Waarop Van Essen antwoord: “Je kunt wel alles bewaren, maar er is zoveel ruis dat je informatie niet terugvindt. En in de toekomst zal vast alles beter gaan, maar we hebben nu pijn en zoeken daarom naar een oplossing”.

Live demo van kunstmatige intelligentie in het archief

Met de presentatie ‘Digitaliseren, en dan? Artificial intelligence in het archief: het Digital Humanities Lab’ weet José de Kruif de aandacht van de toehoorders te vangen. “Ik ga proberen een live demo te geven, een hachelijke situatie natuurlijk, maar ik ga het proberen”, leidt ze haar presentatie in. Het DH-Lab van de Universiteit Utrecht houdt zich bezig met de ontwikkeling van digitale methoden en technologieën in (en buiten) de geesteswetenschappen. Denk daarbij praktisch. “Onderzoekers kunnen bij ons komen met een specifieke vraag op behoefte. Waarna wij op zoek gaan naar geschikte software. Soms heeft die een kleine aanpassing nodig en soms is het nodig een tool te bouwen die nog helemaal niet bestaat”, aldus De Kruif.

Meeschrijven (foto: Anne Reitsma)

Ze demonstreert ons onder meer ‘Texcavator’, een tool waarmee het krantenarchief van de Koninklijke Bibliotheek (zes miljoen pagina’s) full text doorzocht kan worden. Daarbij kunnen visualisaties, tijdlijnen en heat maps ingezet worden. Onderzoekers hebben zelfs de mogelijkheid een set kranten op hun eigen computer zetten. Kunstmatige Intelligentie maakt het mogelijk relaties tussen woorden te leggen. Bijvoorbeeld dat de relatie tussen Parijs en Frankrijk hetzelfde is als die tussen Amsterdam en Nederland). Maar net als de mens is geen techniek zonder fouten. Als De Kruif live zoekt op het woord ‘computer’ rolt als eerstgenoemde een bericht uit 1894 eruit. Een fout van de OCR die een ander woord voor deze moderne term aanzag. Ook leuk: ‘archivaris’ komt in de krant vaak voor in relatie met de woorden 'directeur en leidinggevende’.

Tot slot nog wat tips van De Kruif: “Maak korte trajecten, daarmee voorkom je dat iets uit de hand loopt. Bewaak je tijdpad en geef bij meerwerk op tijd aan dat het langer gaat duren. Werk met veel ontwikkelaars en weinig aansturing. Bewaak consistentie, houd je aan wat je je voorgenomen had. En test uitgebreid!”.

Moderne technieken in de archiefpratijk

Panelleden Ellen Fleurbaay (Stadarchief Amsterdam), Ceciel Huitema (Nationaal Archief), Thomas van Maaren (Picturae) en Ida Nijenhuis (Huygens ING) nemen aan het eind van de

studiedag standpunt in een aantal stellingen. De eerste is dat de behandelde 'Google-technieken' een studiezaal overbodig maken. "Dienstverlening moet verschuiven naar een digitale onderzoekomgeving", aldus Huitema. Fleurbaay ziet vooral een verschuiving in de manier van archiefonderzoek: "Er staat dan wel veel materiaal online, maar aan de hoeveelheid mensen in de studiezaal is niet zoveel veranderd. Want door online veel efficiënter te werken is een studiezaalbezoek goed voorbereid. Zo'n bezoek kan zelfs worden geïnspireerd door online aanwezigheid van het archief". Schomaker wijst vanuit de zaal nog wel op een verbeterpunt in de digitale beschikbaarstelling van archieven: snel door grote hoeveelheden documenten te 'bladeren' en overzicht houden.

De vraag of onderzoekers in de toekomst vaker bij zullen dragen aan de ontsluiting van erfgoedcollecties, in plaats van er vooral over te publiceren, wordt niet goed opgepikt. De panelleden delen de wens wel, maar krijgen geen helder beeld van de praktijk. Op de laatste stelling, 'het werk van crowdsourcingsprojecten kan veel efficiënter benut worden', wordt een duidelijk gezamenlijk standpunt ingenomen. Namelijk dat vrijwillige invoerders of 'the crowd' eigenlijk een groep experts zijn die collecties soms beter kennen dan medewerkers van een archief. Uit de zaal komt de opmerking: "Er zijn vrijwilligers die middeleeuwse schepenboeken beter kunnen lezen dan onze archivarissen". Fleurbaay is het daarmee eens en sluit de paneldiscussie af: "Onderschat de kwaliteit van de crowd niet!".

Programma:	bit.ly/DefProg13okt
Sfeerimpressie:	bit.ly/Video13okt
Presentaties:	bit.ly/SlideshareNOB
Foto's:	Anne Reitsma Fotografie
Tekst:	Tessa Free (Netwerk Oorlogsbronnen)

NETWERK
KENNISCENTRUM GEFACILITEERD DOOR HET NIOD
OORLOGS
BRONNEN

Mogelijk door steun van:
VSBfonds
vfonds VRIJHEID