De experts aan het woord: wat kan de archiefsector met nieuwe digitale technologie?

Edwin Klijn ■

Archiefinstellingen beschikken over een rijkdom aan 'data' die nog nauwelijks gevonden kunnen worden in het digitale domein. Maar op het grensvlak van *Digital Humanities* en digitale collectieontsluiting 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. Leverde de tot voor kort nog gebrekkige software teleurstellende resultaten op, anno 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 de archiefsector? In het novembernummer van het *Archievenblad* las u over twee projecten, deze maand zijn de experts aan het woord: Günter Mühlberger, Martin Reynaert, Lambert Schomaker en Jacco van Ossenbruggen.

Wat kunnen archieven met nieuwe technologie en artificiële intelligentie?

Mühlberger: 'De nieuwe technologie betekent een revolutie voor archieftoegang. We zullen in staat zijn op nieuwe manieren, met een hogere nauwkeurigheid, door archiefstukken te zoeken. Er zijn drie belangrijke terreinen met grote potentie voor archieven: patroonherkenning, *machine learning* en taaltechnologie. De combinatie van deze technologieën zal ons in staat stellen om automatisch kwalitatief hoogwaardige transcripties te creëren, fijnmaziger te zoeken op woorden in de tekst (ook als ze niet in de transcriptie staan, zoals bijvoorbeeld historische spellingsvarianten) en onderzoekers in staat stellen zelf de mate van accuratesse van hun zoekactie in te stellen.'

Schomaker: 'Ik zie grote kansen bij het inschakelen van publiek om collecties beter toegankelijk te maken. Als men bijvoorbeeld aangeeft dat een specifieke vorm het woord 'koning' is, kan de computer dit extrapoleren en alle 'koningen' in de tekst herkennen. In veel crowdsourcingprojecten wordt helaas de handmatige invoer niet aangewend om de software zelf te verbeteren. Als je het slim aanpakt, leer je de computer over beeldpatronen, bladspiegels, woordvormen en semantische relaties, zodat die kennis ook weer elders gebruikt kan worden.'

Reynaert: 'De technologie om analoge tekstcollecties machine-leesbaar te maken, valt stilletjes aan wel productiewaardig te noemen. Het is heel jammer dat het nog zo weinig wordt toegepast op archiefcollecties. Het biedt alle mogelijkheden om zicht te krijgen op wat er zich zoal in die kilometers aan archiefdozen bevindt. Tekstherkenningssoftware heeft grote behoefte aan bestaande nadere toegangen. Archiefinstellingen hebben van oudsher nadere toegangen op de collecties aangelegd, denk aan papieren namenindexen, kaartenbakken et cetera. Deze kunnen we nu prima gebruiken om de digitale toegang te verbeteren.'

Van Ossenbruggen: 'Mijn onderzoek richt zich vooral op wat je ermee kunt zodra collecties – hetzij door middel van de metadata en/of fulltext – digitaal beschikbaar zijn gemaakt. Wat mij daarbij opvalt, is dat erfgoedinstellingen op heel verschillende manieren de collectiedata online beschikbaar stellen. Je ziet dat er veel verschillende zoeksystemen worden gebruikt, waar men doorgaans maar weinig inzicht heeft in hoe deze software werkt. Welke hits komen bovenaan als je op bepaalde termen zoekt (ranking), hoe maakt het zoeksysteem gebruik van de registratie van het klikgedrag (zelflerende patronen) van gebruikers, hoe gaat het zoeksysteem om met historische spellingsvarianten? Het zoeksysteem is vaak ook voor de erfgoedinstelling een *black box*.
Als onderzoeker zou je op zijn minst hierover geïnformeerd moeten worden door de archiefinstelling die het zoeksysteem aanbiedt. Als ik autorijd, hoef ik niet elk schroefje te kennen, maar is het wel fijn als ik weet dat ik moet schakelen als ik bergop rijd. Ik pleit dus voor meer kennis en training van archiefmedewerkers in hoe digitale zoektechnologie werkt. Want deze is voor een groot deel bepalend hoe je gebruikers uiteindelijk je collecties raadplegen.'

Günter Mühlberger (Anne Reitsma Fotografie).

Günter Mühlberger
Studeerde Duitse taal en literatuur, filosofie, psychologie en pedagogie aan de Universiteit van Innsbruck. Hoofd van de onderzoeksgroep Digitalisering en Digitale Duurzaamheid aan de Universiteit van Innsbruck. Nu programmamanager van het READ-project, voorheen betrokken bij verschillende Europese projecten zoals IMPACT (Improved Access to Text), Europeana Newspapers en METAe (waar de basis werd gelegd voor het ALTO-formaat). Zijn onderzoeksveld: digitale humaniora, Handwritten Text Recognition (HTR), machine-learning, taaltechnologie.

Hoe zie je de verhouding tussen Digital Humanities (DH) en erfgoed? Gescheiden werelden of groeien ze naar elkaar toe?

Mühlberger: 'Beide hebben een lange traditie en moeten zichzelf opnieuw uitvinden. Digitale geesteswetenschappers dienen de context en structuur van data te begrijpen. Erfgoedinstellingen moeten weten hoe onderzoekers werken en wat ze nodig hebben. Archiefinstellingen zouden zich moeten focussen op hun kernactiviteiten: het duurzaam bewaren van de originelen (analoog en digitaal) en het bieden van betrouwbare diensten aan hun verschillende doelgroepen. Een algemeen publiek kan misschien het beste worden bediend met een webinterface, maar onderzoekers willen gewoon direct bij de data. Een technische infrastructuur die beide groepen bedient, is van groot belang. Archiefinstellingen zouden zich beter moeten voorbereiden op een van de meest veelbelovende gebruikersgroepen voor de toekomst: machines!'

Schomaker: 'Ik zie deze werelden naar elkaar toegroeien. Het is wel spijtig om te constateren dat er na het CATCH-programma[1] geen geld door de overheid is vrijgemaakt voor soortgelijke initiatieven. De continuïteit in de samenwerking tussen beide is hiermee natuurlijk niet gediend. Het is ook jammer om te zien dat veel erfgoedinstellingen na de afsluiting van zo'n programma hun focus weer richten op hun website en de particuliere gebruiker. Nederland liep voorop in deze samenwerking, maar de grote investeringen worden nu vooral in het buitenland gedaan, met name in Duitsland maar ook in andere Europese landen.'

Reynaert: 'DH-onderzoekers en erfgoedprofessionals werken helaas maar zelden samen. Er bestaat een enorme kloof tussen beiden. En ook binnen de DH-wereld is er nog een wereld van verschil. Op DH-conferenties zie je mensen die fantastische tools ontwikkelen, maar ook mensen die fijn met Excel-bestanden aan het knutselen zijn. Er ligt prachtig materiaal in de archieven. Wel is het vaak moeilijk te vinden en is er weinig gestandaardiseerd. Ik denk dat er grote kansen liggen in een meer intensieve samenwerking tussen beide groepen. De onderzoekswereld is nauwelijks bekend met de problemen waar archieven mee worstelen. Hoe meer de problemen van de archiefinstellingen bij ons bekend zijn, hoe beter wij kunnen meedenken over oplossingen.'

Van Ossenbruggen: 'Beide werelden moeten goed van elkaar weten wat ze doen. Denk aan mijn verhaal over de auto.'

Wat kun jij leren van een archiefmedewerker en andersom?

Mühlberger: 'Vooropgesteld, archiefdocumenten – nog meer dan bijvoorbeeld bibliotheekcollecties – zijn unieke bouwstenen voor de reconstructie van de geschiedenis van de mensheid. Ze zijn vaak uniek, persoonlijk en geven veel informatie 'achter de schermen'. De kunst is om de informatie uit de documenten te contextualiseren met koppelingen naar referentiedata buiten het archief.
Archiefmedewerkers beschikken over heel specifieke kennis hoe een archief in elkaar zit. Zonder deze kennis is het voor de computerwetenschapper vaak onmogelijk de documenten correct te interpreteren. Andersom kunnen archiefmedewerkers van computerwetenschappers leren dat experimenteren en uitproberen uiteindelijk vaak meer oplevert dan theoretiseren en problematiseren. Je moet vertrouwen hebben in technologie én kritisch blijven. Zelfs een mislukt project brengt je vaak verder. Probeer het nog eens en laat niet meteen het grotere idee los.

>>

Martin Reynaert (foto Harold Miesen, Tilburg University).

Lambert Schomaker (foto Rijksuniversiteit Groningen).

Martin Reynaert
Promoveerde in 2005 aan de universiteit te Tilburg op het onderwerp 'Text-Induced Spelling Correction'. Is nu onderzoeker bij Tilburg University en het Centre for Language and Speech Technology aan de Radboud Universiteit te Nijmegen en werkt onder meer in het project Nederlab. Was en is betrokken bij verschillende CLARIN-NL en CLARIAH-projecten (TICCLOPS, OpenSonar, PICCL). Ontwikkelde in dit verband de Text-Induced Corpus Clean-up (TICCL)-software. Zijn onderzoeksveld: natural language processing (NLP), computerlinguïstiek, taaltechnologie, kunstmatige intelligentie.

Lambert Schomaker
Promoveerde in 1991 aan de Rijksuniversiteit Nijmegen op een proefschrift over 'Simulation and Recognition of Handwriting Movements'. Op 1 januari 2001 is hij benoemd tot professor in Kunstmatige Intelligentie aan de Rijksuniversiteit Groningen. Is nu hoogleraar aan de Rijksuniversiteit Groningen en wetenschappelijk directeur van ALICE (Artificial Intelligence and Cognitive Engineering). Op dit moment betrokken bij de doorontwikkeling van het MONK-systeem. Zijn onderzoeksveld: kunstmatige intelligentie, patroonherkenning, machineleren, cybernetica.

>> Computerwetenschappers zijn ook gewend aan imperfecte data. Waar het hen vooral om gaat, is de beschikbaarheid van zo veel mogelijk data, om daarmee bijvoorbeeld imperfecte data te verbeteren. Archiefinstellingen handhaven vaak hoge standaarden en zijn niet altijd bereid imperfectie te accepteren. Maar als je werkt met grote hoeveelheden data, moet je soms genoegen nemen met een zekere foutmarge. Imperfectie kan heel bruikbare data opleveren waarmee de toegang aanzienlijk wordt verbeterd. In andere onderzoeksdisciplines, zoals bijvoorbeeld astronomie, is het overigens heel gewoon dat je werkt met onvolledige data.'

Schomaker: 'Ik heb heel goede ervaringen met archiefmedewerkers: ze zijn vaak enthousiast over hun collecties, zijn zeer betrokken en beschikken over unieke kennis. Ze weten dingen over het materiaal en de systematiek die je nergens in de metadata zult aantreffen en waar de 'crowd' je ook niet aan kan helpen. Die kennis is onontbeerlijk voor ons werk. Ik maak mij dan ook wel enige zorgen in hoeverre dergelijke informatie wordt doorgegeven in organisaties.
Wat archivarissen van computerwetenschappers kunnen leren, is dat het soms verstandig is om de eis van perfectie los te laten en je te focussen op hoe je het optimale uit imperfectie kunt halen. Archivarissen hoeven niet op iedere vraag een antwoord te weten. Het belangrijkste is dat ze mensen de goede richting wijzen.'

Reynaert: 'Ik ben als buitenstaander niet of nauwelijks op de hoogte van de problemen waar archiefmedewerkers zoal mee worstelen. Mijn tools zouden in sommige gevallen heel handig kunnen zijn, bijvoorbeeld bij het ordenen van archieven of het (semi)automatisch genereren van namenindexen op basis van tekstherkenning.'

Van Ossenbruggen: 'Laat ik eens beginnen met de collectiespecialist. Die heeft kwaliteit vaak hoog in het vaandel en beschikt over veel kennis over de collectie. Als deze roept dat haar of zijn collectie heel bijzonder is, klopt dit vaak ook in de praktijk. Het laatste wat we moeten willen, is het unieke uit alle collecties verwijderen en een grote eenheidsworst maken. Respect hebben voor diversiteit en tegelijkertijd standaardiseren waar het kan, daar gaat het om.
Er zou wat mij betreft wel wat meer mogen worden geëxperimenteerd in de archievenwereld. Je leert heel veel van uitproberen, juist wanneer het niet goed gaat. Als ik nu terugkijk op alle projecten waarin ik heb gewerkt, zijn ze bijna allemaal mislukt, in die zin dat nooit alle oorspronkelijke doelen zijn gehaald. Maar vaak zijn er juist wel andere doelen behaald, en heb ik het idee dat ons begrip is toegenomen op een manier die zonder al die 'mislukte' projecten onmogelijk zou zijn geweest. Durf op je bek te gaan, zou ik zeggen.'

Hoe ziet volgens jou een archiefinstelling eruit in 2040? Hoe zou die eruit moeten zien? Is de archivaris nog nodig?

Mühlberger: 'Ik hoop dat archiefinstellingen zich vooral blijven focussen op de rol die ze inmiddels al zo'n 200 jaar vervullen; ze moeten zich vooral toeleggen op het bewaren van de originelen. Archiefinstellingen moeten ernaar streven alles wat ze hebben te digitaliseren. Hiervoor is een paradigmashift noodzakelijk. Digitalisering zou niet – zoals je vaak ziet – afhankelijk moeten zijn van projectsubsidies. Middelen in de reguliere begroting moeten worden vrijgemaakt om een stevige, permanente basis te vormen voor massadigitalisering van collecties. Dit geldt ook voor middelgrote en kleinere instellingen. Een klein team met beperkte middelen kan soms al heel wat doen. Dan lukt het misschien niet in drie of vijf jaar, maar in dertig tot vijftig jaar tijd is ook goed.'

Schomaker: 'Dit is bijna een politieke vraag: archiefinstellingen zullen blijven bestaan zolang de samenleving het bewaren van haar erfgoed belangrijk vindt. Het is lastig te voorspellen welke politieke stromingen het de komende decennia voor het zeggen gaan krijgen. Zoals ik al eerder betoogde, is de kennis van de archiefmedewerker elementair voor de verdere digitale ontsluiting van de archieven. Onderzoekers van het Japanse bedrijf Hitachi hebben laten zien dat je 360 terabyte kunt opslaan op een plaatje glas van twee bij twee centimeter, heel robuust. De 'bits' in de kwartsstructuur blijven honderden miljoenen jaren intact. Het idee is al geopperd om cultureel erfgoed via zo'n medium naar Mars te transporteren. Gewoon, *just in case*...
Ik denk dat het belangrijk is dat archieven zich focussen op wat ze al jaren doen: het duurzaam beheren en bewaren van de oorspronkelijke documenten. Om papieren stukken te bekijken heb je gelukkig alleen maar fotonen nodig. Voor digitaal ben je afhankelijk van snel verouderende systemen en bestandsformaten. Daar ligt een grote uitdaging voor de archiefsector. Ik heb archiefmedewerkers nodig die mij vertellen dat een rare krul in de linkermarge 'solvit' ('betaald!') betekent. Dit gaat de 'crowd' mij niet vertellen, ben ik bang.'

Reynaert: 'Ik hoop natuurlijk dat er veel meer gedigitaliseerd gaat worden door de archiefinstellingen. Hoe meer hoe beter, zou ik zeggen. Het openlijk beschikbaar stellen van digitale collecties heeft als bijkomstig voordeel dat er beter onderzoek kan worden gedaan naar verbetering van de digitale ontsluiting. Hoe meer data, hoe beter wij ons werk kunnen doen en onze tools kunnen aanscherpen.
Archiefinstellingen zullen er zeker nog zijn in 2040, tenzij er iets heel ergs gebeurt met onze samenleving. Naast digitalisering moeten archiefinstellingen zich wat mij betreft vooral bezighouden met het bewaren van de originelen. Het is nog maar de vraag of verregaande digitalisering ook leidt tot een afname van bezoek aan de instelling. Neem het Rijksmuseum als voorbeeld; daar is het toch niet rustiger geworden toen ze hun gehele collectie in hoge resolutie vrijgaven? Als het om topstukken gaat, zullen mensen toch altijd ook de originelen willen zien. Originele stukken hebben zich overigens soms meer bewezen dan digitale objecten; de meest duurzame drager voor tekstbestanden is nog altijd het kleitablet.'

Van Ossenbruggen: 'Zoals ik al eerder betoogde: archieven maken allerlei keuzes die voor de onderzoeker en het publiek in het algemeen grote consequenties kunnen hebben. Overigens niet alleen wat betreft zoeksystemen, maar ook op een flink aantal andere gebieden. Ik noem hier maar even het archiveren van 'born digital'-materiaal. Als ik het terugbreng tot mijn eigen ervaring: toen laatst mijn oude smartphone de geest gaf, was ik in één keer een flink stuk van mijn fotografisch geregistreerde leven kwijt. Ik maak mij eigenlijk meer zorgen over het duurzaam bewaren van digitaal erfgoed dan van analoog materiaal; papier is niet zomaar weg.
We hebben het vooral gehad over de doorzoekbaarheid van archief. Wat daarnaast altijd belangrijk blijft, is dat je de context van een collectie goed begrijpt: wat zit er wel en niet in, hoe is deze collectie gevormd en ook: welke relaties hebben deze documenten tot andere collecties? Zoals ik al eerder betoogde, is deze context van groot belang voor onderzoekers en andere gebruikers van de collecties. Hier ligt een belangrijke taak voor de archiefinstellingen weggelegd, juist ook binnen het digitale domein.
Kortom: het is belangrijk dat archieven zichzelf de kennis en vaardigheden toe-eigenen om zich ook in de toekomst te ontfermen over het bewaren en het beschikbaar stellen van ons erfgoed. Daarnaast zouden de keuzes die hierbij worden gemaakt, veel explicieter naar buiten moeten worden gebracht in het publieke debat. Het werk van de archivaris blijft in de toekomst, misschien wel meer dan ooit, een onmisbare schakel in onze informatiemaatschappij.' ■

Jacco van Ossenbruggen (foto CWI).

Jacco van Ossenbruggen
Computerwetenschapper. Hoofd van de Information Access-groep bij het Centrum Wiskunde & Informatica (CWI) en universitair hoofddocent bij de Web & Media-onderzoeksgroep aan de Vrije Universiteit Amsterdam. Zijn onderzoeksveld: semantisch web, ontologieën, linked data, informatica, onderzoeksmethodieken big data in digitale humaniora.

Noot

1 ■ http://www.nwo.nl/en/research-and-results/programmes/Continuous+Access+To+Cultural+Heritage+%28CATCH%29

Edwin Klijn ■ programmamanager van Netwerk Oorlogsbronnen (www.oorlogsbronnen.nl). Dit artikel is mogelijk gemaakt door Archief 2020, BRAIN, het ministerie van VWS, het VSBfonds en het Vfonds. Met speciale dank aan het Nationaal Archief.