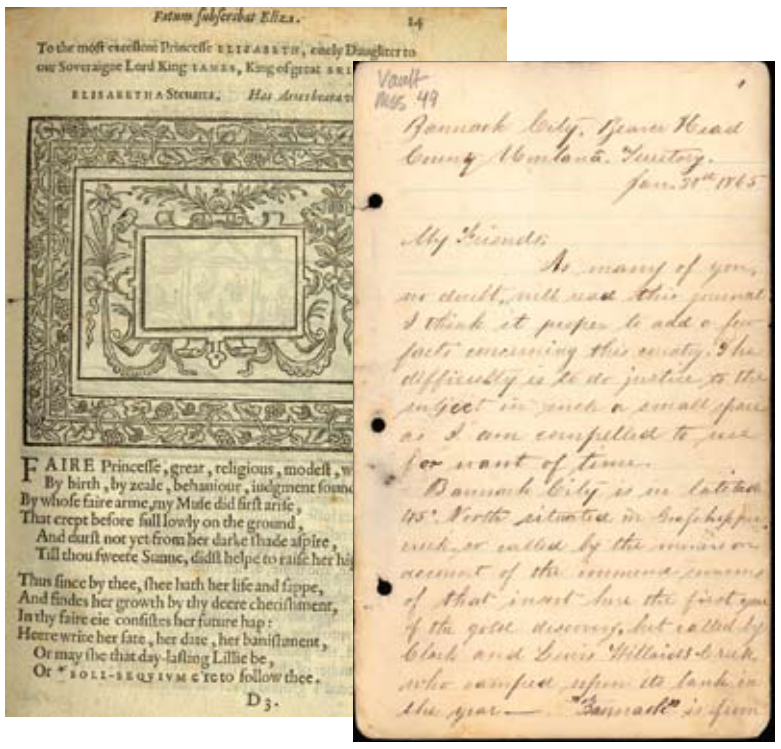
IMPACT: historische tekst echt digitaal maken

Een consortium van vijftien partners uit Europa en Israel werkt onder leiding van de Koninklijke Bibliotheek aan technieken om optische tekenherkenning geschikt te maken voor oude teksten.

HILDELIES BALK



Historische teksten worden wereldwijd op grote schaal gedigitaliseerd. Per jaar gaan miljoenen pagina's uit boeken, kranten en tijdschriften door de scanner. Vervolgens worden ze omgezet naar een tekstformaat door middel van optische tekenherkenning (OCR). Het doel is steeds om de gebruiker ter wille te zijn. De gedigitaliseerde teksten moeten op dezelfde manier doorzocht en bewerkt kunnen worden als documenten die digitaal zijn ontstaan.

Dat lukt aardig bij moderne teksten met een eenvoudige lay-out, gedrukt in de twintigste eeuw. Maar het doel wordt vaak niet bereikt als het gaat om drukwerk van vóór 1850. Oude lettertypes en druktechnieken, ingewikkelde lay-out, beschadigingen in het materiaal en spellingsvarianten; allemaal redenen waarom het moeilijk en duur is om een echte digitale (doorzoekbare, te bewerken) versie te maken van een tekst van voor 1850. De bestaande software voor automatische OCR levert bedroevend lage resultaten op. Postcorrectie - achteraf bijwerken - is duur en tijdrovend en niet geschikt voor digitaliseren op grote schaal. Dit is een grote hindernis bij de overgang van de relatief kleine digitaliseringsprojecten van de afgelopen jaren naar de werkelijke massadigitalisering van de toekomst.

Een consortium van vijftien nationale bibliotheken, universiteiten en bedrijven uit Europa, Israel en Rusland, geleid door de Koninklijke Bibliotheek, heeft op zich genomen om te werken aan een oplossing. In het voorjaar van 2007 is een gezamenlijk voorstel voor subsidie binnen het zevende kaderprogramma van de Europese Commissie ingediend, onder de naam IMPACT (IMProving ACces to Text). In dit project zal worden gewerkt aan verschillende innovatieve technieken voor de verbetering van automatische OCR, uiteenlopend van een zelflerende OCR machine tot de bouw van verfijnde computerlexica. De diverse tools en applicaties zullen zo worden ontworpen dat ze *standalone* kunnen functioneren maar ook geïntegreerd kunnen worden in een gebruiksvriendelijk platform. Naast de verbetering van de techniek zal in het project ook veel aandacht worden besteed aan het delen van kennis en *'best-practices'* op het gebied van massadigitalisering. De startdatum van het project, dat vier jaar gaat duren, is 1 januari 2008.

www.kb.nl

Archiveren in de academische psychologie

Tijd om data te gaan koesteren

In het psychologisch onderzoek op de Nederlandse universiteiten bestaat nog geen traditie op het gebied van archiveren en hergebruiken van gegevens. Data genoeg, maar de animo om ze duurzaam te bewaren is wisselend. In sommige onderzoeksrichtingen bestaat wel belangstelling, maar voordat het zo ver is zijn er nog genoeg haken en ogen, concludeert Caroline Voorbrood. Zij deed bij DANS een verkennend onderzoek naar het datalandschap in de Psychologie.

CAROLINE VOORBROOD

Academische psychologen in Nederland voelen als zo velen de druk om te publiceren. Zonder onderzoeksdata echter geen publicatie. Wat gebeurt er eigenlijk met deze onderzoeksdata nadat erover is gepubliceerd? In hoeverre worden ze gearchiveerd en hergebruikt? Hoe ziet de data-infrastructuur van de psychologie in Nederland er uit? Het zijn enkele vragen waarop het project Inventarisatie Psychologie van DANS (Data Archiving and Networked Services) een antwoord moest vinden. Doel: inzicht krijgen in de bestaande data-infrastructuur en in de behoeften van academisch werkzame psychologen op het gebied van systematisch archiveren, hergebruiken, koppelen, verrijken en delen van onderzoeksdata. Onder archiveren wordt hier verstaan het toegankelijk maken en houden van databestanden: opslag, maar ook het beschrijven van de data door metadata toe te kennen met technische en meer contextuele informatie.

ELLEN BOUMA

De wensen en gebruiken op dit gebied blijken onder psychologen sterk uiteen te lopen, al kan op een enkele uitzondering na gesteld worden dat psychologen hun onderzoeksdata niet systematisch archiveren. Er worden op ruime schaal back-ups gemaakt, waarbij het toegankelijk maken en houden van de databestanden uit het oog wordt verloren. Daartoe ontbreekt het de psychologen aan tijd, geld en soms inzicht in het nut van archiveren. De onderzoeksdata zijn vaak nog wel ergens, maar ze raken zoek of onbruikbaar.

De behoefte om in deze situatie verandering te brengen blijkt wisselend. Een positieve houding is er bijvoorbeeld bij degenen die onderzoek doen naar een bijzondere populatie; psychologen met een zeer grote output aan onderzoeksdata; psychologen die werken met longitudinaal onderzoek en psychologen die behoefte hebben aan het waarborgen van het behoud van hun gegevens. Voor degenen die vooral experimenteel onderzoek doen is systematisch en eventueel centraal archiveren echter veelal niet interessant. Een experiment opnieuw uitvoeren is sneller dan het archiveren van de onderzoeksdata, zeggen ze. Daarnaast zouden de data door hun specifieke karakter niet interessant zijn voor anderen.

Veel psychologen signaleren ook een gevaar in het systematisch archiveren van psychologische data: er zou een grote 'datavuilnishoop' kunnen ontstaan. Wanneer de vakgenoten dan tenslotte door de bomen het bos niet meer kunnen zien zou ongebruikt blijven van de gearchiveerde data het gevolg kunnen zijn. Dit gevaar is te voorkomen door een vergaande specificering van de data in disciplines, liefst per onderzoekthema. Andere zorgen en discussiepunten zijn het waarborgen van de kwaliteit van de data, intellectueel eigendom, het ontstaan van 'werkpaardjes' versus 'luxepaardjes' en methodologische bezwaren. Technisch-inhoudelijke wensen of inzichten met betrekking tot een databank zijn er weinig: 'Het is goed zoals DANS het doet'.

Psychologen denken bij de vraag naar hergebruik van data vooral aan van tevoren vastgestelde samenwerkingsverbanden. Maar hergebruik vindt daarnaast ad hoc plaats op individueel verzoek. Psychologen kunnen op deze manier hun data nader toelichten en kunnen enige controle houden over de (kwaliteit van de) nieuwe onderzoeksvraag. Hieruit blijkt dat er behoefte bestaat aan psychologische onderzoeksdata, maar tegelijk geven de academische psychologen aan zelf genoeg onderzoeksdata op de plank te hebben liggen om over te publiceren.

Binnen een deel van de academische psychologie bestaat dus wel de behoefte en bereidheid om onderzoeksdata systematisch centraal te gaan archiveren. Psychologen kunnen verschillende haken en ogen benoemen van het goed archiveren van hun data. Bij het vormgeven van een archief is de inbreng van psychologen daarom onontbeerlijk. Het startpunt zal liggen bij enkele gedreven individuen in de academische psychologie. Waardevolle psychologische onderzoeksdata zullen nieuwe data aantrekken. De publicaties worden al op waarde geschat. Het is tijd om de psychologische onderzoeksdata te gaan koesteren.