Edwin Klijn*

Van 'oud' geheugen naar digitaal brein

Massadigitalisering in de praktijk

In 1941 fantaseerde de Argentijnse schrijver en voormalig bibliothecaris Jorge Luis Borges in het inmiddels befaamde korte verhaal 'La Biblioteca de Babel' over een mythische bibliotheek, die alle boeken in alle talen uit alle tijden omvatte. Deze moeder van alle bibliotheken herbergde alle denkbare informatie, van totaal betekenisloze verzamelingen met willekeurige tekens tot de alomvattende catalogus van de bibliotheek, inclusief alle duizenden vervalsingen ervan. Toen wereldkundig werd gemaakt dat de Bibliotheek werkelijk alle boeken in haar collectie had, overheerste er in eerste instantie een gevoel van buitensporig geluk; 'Alle mensen voelden zich de baas van een ongeschonden, geheime schat'.[1] Eindelijk kon antwoord worden gegeven op de basismysteries van de mensheid, te weten de oorsprong van de Bibliotheek en van de tijd. De hoge verwachtingen liepen spoedig uit op een deceptie: de antwoorden waren er ongetwijfeld, maar niemand kon ze vinden. Borges' Bibliotheek omvatte zo veel informatie dat de zoektocht naar kennis even oneindig was als de boekenrijen op de planken in de talloze zeshoekige galerijen. De officiële 'zoekers' – de inquisiteurs – kwamen altijd uitgeput terug van hun expedities, verhaalden over een gebroken trap die hen bijna fataal was geworden en praatten met de bibliothecaris over galerijen en trappen. Niemand verwachtte ooit wat te vinden, hetgeen uiteindelijk een zeer diepe depressie tot gevolg had.

Borges' imaginaire bibliotheek wordt vaak opgevoerd als metafoor voor het internet; ook het web is een immense vergaarbak van zeer diverse informatie. In 2010 werd de omvang van *the global village* geschat op 5 miljoen terabyte. Ter vergelijking: het menselijk brein kan 1 tot 10 terabyte aan informatie opslaan.[2] Elke dag nog groeit het aantal gebruikers en worden nieuwe data toegevoegd. Naast *born digital*-data belandt ook steeds meer oorspronkelijk analoog materiaal op het internet. Ons 'oude geheugen' – van oudsher berustend in archieven, bibliotheken en musea – wordt langzaam maar zeker in een nieuw digitaal brein, het internet, opgenomen.[3] Harde cijfers over de actuele stand van zaken wat betreft de digitalisering van erf-

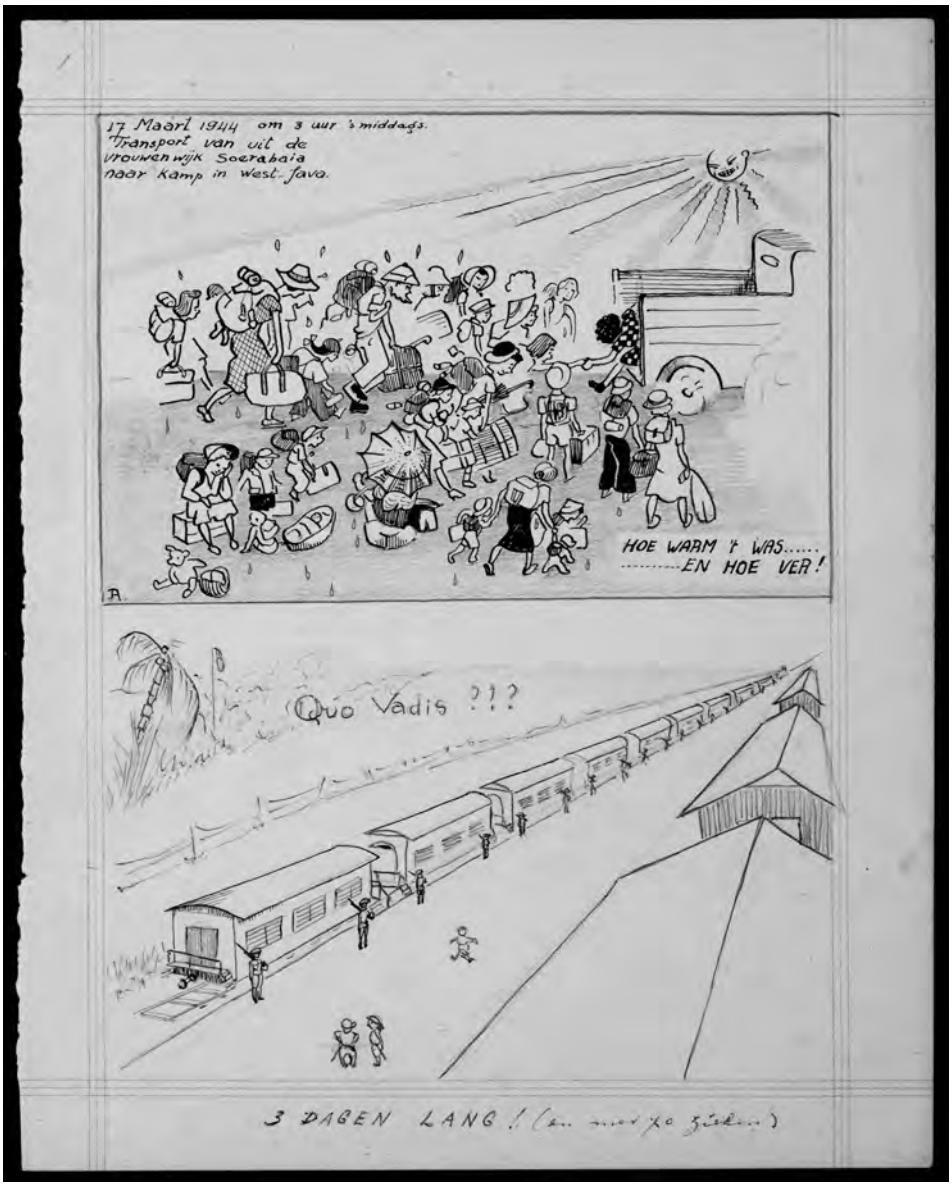
* Edwin Klijn (1970) is werkzaam bij het NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies als manager van het Netwerk Oorlogsbronnen (www.oorlogsbronnen.nl) en teamleider diensten. Van 2009 tot 2011 was hij projectmanager van het Historische Kranten-project van de Koninklijke Bibliotheek. Daarvoor was hij betrokken bij verschillende digitaliseringsprojecten. Klijn publiceerde over massadigitalisering, beeldbanken en conserving en digitalisering van foto- en audiovisuele collecties. Daarnaast schreef hij samen met Robin te Slaa *De NSB. Ontstaan en opkomst van de Nationaal-Socialistische Beweging 1931-1935* (Amsterdam 2009), genomineerd voor de Libris Geschiedenis Prijs 2010.

goedcollecties in Nederland ontbreken. Volgens een ruwe schatting uit 2008 zou pas een fractie (2 tot 3 procent) van alle Nederlandse collecties gedigitaliseerd zijn.[4]

Veel erfgoedinstellingen zijn druk bezig de stap naar het digitale domein te zetten. De Koninklijke Bibliotheek (KB) voorziet dat in 2030 alle publicaties in en over Nederland (naar schatting zo'n 730 miljoen pagina's) in digitaal formaat beschikbaar zullen zijn.[5] De archiefsector timmert flink aan de weg, maar heeft een gigantische klus te klaren: alleen al de digitalisering van de 110 kilometer archief van het Nationaal Archief vergt een inspanning die groter is dan de klus die de KB voor ogen heeft.[6] Musea, zo leert een onderzoek over 2008 van Digitaal Erfgoed Nederland, zouden van alle erfgoedinstellingen het relatief grootste deel van hun collecties hebben gedigitaliseerd.[7]

De erfgoedsector heeft inmiddels zo'n twintig jaar ervaring met digitalisering. Het waren aanvankelijk de archieven, die in de jaren negentig van de vorige eeuw databases aanlegden, met als doel vooral genealogische gegevens te ontsluiten. De grote nationale erfgoedinstellingen volgden, maar met een andere invalshoek; in de eerste jaren digitaliseerden zij op kleine schaal, experimenteel en vooral ten behoeve van virtuele tentoonstellingen en digitale schatkamers.[8] Door het tonen van de hoogtepunten uit de collectie hoopten zij webgebruikers te verleiden een bezoek aan de studiezaal te brengen. Al spoedig gingen ook de nationale instellingen digitalisering benutten om bezoekers op afstand gebruik te laten maken van bepaalde collecties. Naarmate de technologie verbeterde en expertise toenam, werd digitalisering vaker ingezet als een probaat middel om hoogwaardige digitale surrogaten van kwetsbaar analoog materiaal te maken (*preservation imaging*). De overstap van Metamorfoze, het nationaal conserveringsprogramma voor het behoud van het papieren erfgoed, van *preservation microfilming* naar *preservation imaging* in 2007 markeert deze ontwikkeling.[9]

Vanaf het begin van de 21ste eeuw ontstonden de eerste grote nationale massadigitaliseringsprojecten, zoals Staten Generaal Digitaal.[10] In deze projecten werden systematisch grote hoeveelheden originelen gescand, machine-leesbaar gemaakt en – voor zover dit nog niet voor handen was – voorzien van metadata.[11] De grote drijfveer achter de massadigitalisering was de ambitie om online toegang tot integrale delen van collecties te verschaffen. Door niet alleen op kwaliteit te focussen, maar ook op doorvoersnelheid, werd de kostprijs per pagina aanzienlijk gereduceerd. De laatste jaren heeft de trend van massadigitalisering zich voortgezet. De meest aansprekende voorbeelden hiervan zijn het programma Beelden voor de Toekomst[12] en het Historische Krantenproject.[13] Dat steeds vaker massadigitalisering niet meer op projectbasis plaatsvindt, maar integraal deel uitmaakt van de werkzaamheden is een recente ontwikkeling; het Stadsarchief Amsterdam, de Koninklijke Bibliotheek en het Nationaal Instituut voor Beeld en Geluid zijn voorbeelden van organisaties waar dit op van toepassing is. Een dergelijke overgang van project naar lijnorganisatie illustreert het 'tot wasdom komen' van digitalisering.



Pagina uit een plakboek bestaande uit 36 tekeningen, twee briefkaarten, vier krantenknipsels, een voor- en achterblad en twee tussenvellen. Het plakboek betreft de kampen Vrouwenwijk Soerabaja, Kamp Tangerang en Kamp Adek Batavia en is gemaakt en samengesteld door Mevrouw Tineke Robson-Augustijn, gedurende haar internering in deze kampen. Dit beeld is de elfde tekening en bevat de tekst; '17 maart 1944 om 3 uur 's middags. Transport van uit de vrouwenwijk Soerabaia naar kamp in West-Java. Hoe warm het was... en hoe ver! Quo Vadis??? 3 dagen lang'. Bron: NIOD *Beeldbank WO2*

Technische complexiteit

Alles digitaliseren klinkt eenvoudig, maar in de praktijk loopt men al snel tegen allerlei onvolkomenheden aan. Digitaliseren beperkt zich niet simpelweg tot het maken van een scan. Er komt veel meer bij kijken. Voordat een boek, krant, tijdschrift of document gescand kan worden, moet het van de plank worden gehaald, eventuele scheuren worden geplakt, vouwen gladgestreken en beschadigingen gerepareerd. Als het een archief betreft, moeten er soms nietjes worden verwijderd of stukken op volgorde worden gelegd, beide zeer arbeidsintensieve klussen, vooral als het om grote hoeveelheden gaat. Scanapparatuur is de afgelopen jaren goedkoper en beter geworden, maar alles staat of valt toch vooral bij de juiste instellingen en dus uiteindelijk een vakkundige operator. Scannen kan iedereen, maar *goed* scannen is nog altijd een vak apart en een grotere uitdaging in een situatie waarin grote hoeveelheden fabrieksmatig moeten worden verwerkt.

Bij tekstdigitalisering wordt er naast een scan ook vaak softwarematig een tekstbestand geproduceerd. Een dergelijk bestand is nodig om uiteindelijk gebruikers op elk woord in de tekst te laten zoeken, zoals men dit gewend is van bijvoorbeeld diensten als Google. De zogenaamde *Optical Character Recognition* (OCR)-software probeert letters en woorden in de scan te herkennen en om te zetten naar voor de computer begrijpelijke tekst. Deze transformatie verloopt doorgaans niet vlekkeloos; gotische drukletters, slecht gedrukte tekst (met weinig contrast, cursief of met onvoldoende ruimte tussen de letters), scheve regels, doordrukken en hobbelig papier kunnen ervoor zorgen dat woorden niet of verkeerd worden geïnterpreteerd door de computer en dus ook niet worden gevonden door gebruikers.

De inzet van OCR-software bij het digitaal toegankelijk maken van historisch tekstmateriaal heeft geleid tot een grote sprong voorwaarts. Veel onderzoekers zullen zich de dagen herinneren, waarin zij handmatig, van document tot document, uren- en soms dagenlang tevergeefs op zoek waren naar dat ene, summiere fragment. Dankzij OCR-technologie kan met één muisklik door kilometers archief worden gezocht, iets wat voorheen ondenkbaar was. Tegelijkertijd zijn de resultaten die met de huidige OCR-technologie worden bereikt, nog vaak ondermaats. Een meting van de KB in 2004, op basis van een steekproef in de kranten *Het Vaderland*, het *Centrum*, het *Volk* en de *NRC*, toont aan dat tussen 27% tot 52% van de woorden onjuist wordt weergegeven.[14] Een recent onderzoek binnen het 19th Century Online Newspaper Archive van de British Library wijst uit dat 22% van de woorden niet correct door de OCR-software was omgezet.[15] Een flinke foutmarge dus; door verbetering van deze software valt nog veel winst te behalen. Veel webgebruikers zullen er overigens maar weinig van merken; wat je niet vindt, zul je doorgaans ook niet missen.

Hoe kan de kwaliteit van de OCR-tekst verbeterd worden? Algemeen aanvaard uitgangspunt is dat de meeste winst valt te behalen op het niveau van de software. Slimme software kan regels rechtzetten, contrast aanpassen en op basis van bestaande lexica woorden herkennen. In vergelijking met de huidige software,

beschikt de mens nog altijd over een veel verfijnder perceptievermogen. Door menselijke intelligentie in te zetten bij het herkennen van woorden of letters kan OCR-software geleerd worden om bepaalde letters te herkennen en deze verworven kennis op grote gelijksoortige corpora toe te passen. In projecten als MONK (voor manuscripten)[16] en IMPACT (*Improving Access to Text*)[17] is hiermee uitgebreid geëxperimenteerd.

Andere OCR-verbeteringsacties richten zich op het inschakelen van het publiek bij het corrigeren van door de computer gegenereerde tekst. Zogenaamde public *collaborative OCR correction*, waarbij webgebruikers zelf OCR-tekst online kunnen verbeteren, resulteert bij grote hoeveelheden gedigitaliseerd materiaal in een verhoudingsgewijs marginale verbetering, maar zorgt tegelijkertijd ook voor een hoge mate van betrokkenheid van het publiek. 'A wonderful tool - the amount of user control is very surprising but refreshing', luidde het commentaar van een van de enthousiaste vrijwilligers van het krantenproject van de Australische nationale bibliotheek.[18] Overigens wordt het publiek ook wel eens ingezet om de metadata te verbeteren: het Nationaal Instituut voor Beeld en Geluid, in samenwerking met de Vrije Universiteit, heeft onlangs een proef gedaan om met behulp van een spel (Woordentikkertje) het beeldarchief van het populaire televisieprogramma *Man Bijt Hond* door gebruikers te laten ontsluiten.[19] Het project, dat tot veel positieve reacties van gebruikers leidde, zal een vervolg krijgen.[20]

De OCR-problematiek – en daarmee de kwaliteit van de doorzoekbaarheid van gedigitaliseerd erfgoedmateriaal – zal in de komende decennia een voortdurend aandachtspunt blijven. Hoe meer er wordt gedigitaliseerd, hoe belangrijker het wordt om goed en gericht te kunnen zoeken. OCR-technologie staat nog in de kinderschoenen, zeker wat betreft historische teksten. Een vraag die zich opdringt, is of het met het oog op de hoge foutmarge niet verstandiger is te wachten met OCR'en totdat de technologie betere resultaten oplevert. Een zeker pragmatisme is hierbij geboden: in de praktijk kost het maken van een OCR-bestand relatief weinig (3 à 4 cent per pagina). Als met deze investering de doorzoekbaarheid van de collectie aanzienlijk kan worden verbeterd, lijkt dit een legitieme keuze, zelfs als het gaat om een *voorlopig* resultaat. Wel moet er rekening mee worden gehouden dat er waarschijnlijk in de toekomst enkele malen opnieuw zal moeten worden ge-OCR'd.

Kosten

Andere beperkingen vloeien voort uit de aard van het oorspronkelijke materiaal en de kosten die met de voorbereiding en nazorg zijn gemoeid. Bij massadigitalisering draait alles om het inrichten van een sluitend logistiek proces. Hoe meer regelmaat en eenduidigheid, hoe minder kans op fouten en ook hoe lager de productiekosten. De grote valkuilen van elk massadigitaliseringsproject zijn, wat in de wandelgangen nog wel eens wordt genoemd: de meest voorkomende uitzonderingen. Hierbij kan worden gedacht aan originelen van buitengewone afmetingen, uitklapplaten in

boeken en ongedateerde afleveringen. Bij tekstdigitaliseringsprojecten wordt ongeveer de helft van het budget besteed aan het scannen, OCR'en en metadateren. Het overige deel gaat op aan de voorbereiding en de nazorg.[21] Bijna alle erfgoedinstellingen digitaliseren niet zelf, maar schakelen een van de circa vijf gespecialiseerde scanbedrijven in Nederland hiervoor in. De paginaprijs is sterk afhankelijk van het te verrichten maatwerk en de te verwerken hoeveelheden. Als het gaat om een losgesneden boek of een stapel A-4'tjes die met een doorvoerscanner kunnen worden verwerkt, bedraagt de paginaprijs ongeveer 10 cent (incl. BTW). In geval van niet-eenduidige, losbladige archiefstukken loopt de prijs al snel op naar 30 cent (incl. BTW) per pagina.[22] Het basaal digitaliseren van één meter archief kost ongeveer 2.100 euro. Dit lijkt te overzien, maar voor een middelgrote archiefcollectie zoals die van het NIOD, zou het ruim 10,5 miljoen euro kosten om de totale 2,5 kilometer aan analoog materiaal te digitaliseren. Om nog maar te zwijgen van alle andere archiefcollecties in heel Nederland.

Digitaliseren is dus prijzig. Fondsenwerving voor digitalisering van collecties is moeilijker dan ooit tevoren. Juist op een moment waarop massadigitalisering zo volwassen is geworden dat er efficiënt en snel grote hoeveelheden kunnen worden gedigitaliseerd, ontbreken met name bij de middelgrote en kleine erfgoedinstellingen de middelen om de grote slag te kunnen maken. De Nederlandse overheid beschouwt digitalisering vooral als een verantwoordelijkheid van de erfgoedinstellingen zelf. Die moeten digitalisering zelf bekostigen, omdat die deel uitmaakt van het reguliere dienstenpakket. Erfgoedinstellingen verbreden inmiddels hun gezichtsveld en richten zich steeds vaker op particuliere fondsen, Europese subsidies of publiek-private samenwerkingsverbanden. In de laatste categorie valt te denken aan de overeenkomsten van de KB met Google en Proquest of het Internationaal Instituut voor Sociale Geschiedenis (IISG) met SNS Reaal. Ook *Digitisation on Demand* – het digitaliseren op verzoek van klanten tegen betaling – is in opmars, maar het valt vooralsnog niet mee om hier een kostendekkende productiemethode voor in te richten, mits het opdrachten van een zekere omvang betreft.

Meer digitaliseren betekent ook meer structurele kosten voor onderhoud en beheer. Door te digitaliseren worden feitelijk nieuwe collecties gecreëerd. Net als hun fysieke equivalenten vragen deze collecties om zorg en aandacht. Maar daar waar een boek of papieren document altijd nog te lezen is zolang het materiaal zelf niet aan verval onderhevig is, zijn digitale bestanden kwetsbaarder. Om digitale collecties duurzaam te bewaren, moet niet alleen rekening worden gehouden met de bestanden, maar ook met de software waarmee de bestanden kunnen worden geopend. En deze software kent vaak een grillige levenswandel (Microsoft Word heeft inmiddels al zo'n twaalf versies doorlopen) of – in het ergste geval – een tragisch einde, wanneer software niet meer ondersteund wordt door de fabrikant (bijvoorbeeld het ooit veelgebruikte Word Perfect). Erfgoedinstellingen hanteren vaak twee tactieken of een combinatie daarvan om hiermee om te gaan; in eerste instantie migratie van 'bedreigde' bestanden naar een nieuw formaat of nieuwe versie, daarnaast emulatie van de oorspronkelijke software waarmee de bestanden kunnen

worden bekeken. In beide gevallen gaat het om ingrijpende maatregelen die vaak kostbaar zijn. De structurele kosten die gemoeid zijn met het 'in leven houden' van digitale collecties omvatten een veelvoud van de eenmalige digitaliseringskosten van dezelfde collecties.

Zoekmogelijkheden

Digitalisering levert bestanden op. Met alleen goede bestanden kan men nog niets. Om die uiteindelijk met een handige zoekinterface op het web te kunnen aanbieden, moet er nog het een en ander gebeuren. Er is software nodig die alle bestanden indexeert, zodat ze snel kunnen worden doorzocht, gesorteerd en gegroepeerd (later nodig voor de filters op de website). Met grote corpora telt elke milliseconde. Het efficiënt inrichten van de index kan een groot verschil maken voor de snelheid waarmee data op het web kunnen worden gepresenteerd. Naast een zoekmachine is ook software nodig om de data in een webomgeving te presenteren. Omdat de eisen voor presentatie vaak heel specifiek zijn, is dit maatwerk en dus kostbaar. Zowel zoektechnologie als webpresentatie is zeer onderhevig aan verandering. Webgebruik en zoekgedrag ontwikkelen zich razendsnel, zoals blijkt uit de opmars van het mobiele internet en de invloed die de zoekfunctie van iTunes heeft gehad op onder meer Google. Dergelijke ontwikkelingen sijpelen weer door in gebruikersverwachtingen voor bestaande applicaties. Een website is nooit af.

In grote lijnen zijn er drie bepalende factoren die ervoor zorgen dat goed kan worden gezocht in een digitale tekstcollectie, het gaat om de kwaliteit van de

- oorspronkelijke metadata;
- metadata en OCR-tekst die tijdens de digitalisering zijn geproduceerd;
- zoekmachine.

Als de bestanden die geproduceerd worden onder de maat zijn (veel spellingfouten, slechte OCR-tekst), zal een zoekmachine – hoe intelligent deze ook is – deze hoge foutmarge slechts voor een deel kunnen wegpoetsen met een handige *fuzzy search*[23] of anderszins. Door geïmporteerde thesauri, woordenboeken, lijsten met typografische varianten, spellingsvarianten of synoniemen, 'stemming' (algoritme dat gebruikt wordt om woorden terug te brengen naar 'stamwoorden') kan een zoekmachine vaak heel slim zoeken. Wat er precies achter de schermen gebeurt, is bij de huidige zoeksystemen voor gebruikers vaak moeilijk te achterhalen. Tegelijkertijd bepaalt dit wel voor een groot deel wat zij aan zoekresultaten gepresenteerd krijgen. De bepaling van relevantie van de treffers is bijvoorbeeld zoiets dat zich grotendeels in 'het zoeksysteem' afspeelt, terwijl vooral bij grote hoeveelheden data de schikking van de zoekresultaten heel belangrijk is. Zo geldt de Google-ranking van bedrijven inmiddels als een beursgevoelige, statusbepalende factor.

De website – in vakjargon geduid als de *frontend* of *user interface* – geldt vaak als het eindpunt van een digitaliseringsproces. Vooral bij grote projecten is de website slechts de buitenste schil van een geheel van op elkaar aangesloten hard- en soft-

warecomponenten. Het meeste werk wordt doorgaans verricht onder de motorkap, iets waar de gebruikers – als het goed is – niets van merken. De website is feitelijk het uiteindelijke virtuele dienblad waarmee de data aan de gebruiker wordt gepresenteerd. Een goede website beantwoordt aan de eisen en verwachtingen van de beoogde doelgroepen. Een grondig gebruikersonderzoek, liefst ook inclusief een *bèta*-test van de applicatie, mag eigenlijk niet ontbreken. Het is belangrijk om de functionele eisen van gebruikers te vertalen in technische oplossingen – en niet andersom. Hoewel de website vaak pas bij het sluiten van de markt wordt opgeleverd, begint elk goed digitaliseringsproject met een duidelijk idee over de functionaliteit van de te bouwen dienst in gebruikerstermen. Hoe gedetailleerder en nauwkeuriger de functionele eisen, hoe doelgerichter en efficiënter het digitaliseringsproces en de ondersteunende technologie kunnen worden ingericht.

Juridisch kader

Naast technologische beperkingen zijn er ook op juridisch vlak openbaarheidsbeperkingen die het vrij toegankelijk online aanbieden van erfgoedcollecties in de weg staan. Deze restricties vloeien voort uit een aantal wettelijke bepalingen waaraan Nederlandse erfgoedinstellingen zijn gebonden.[24] De Auteurswet schrijft voor dat auteursrechtelijk beschermd werk in principe alleen mag worden hergebruikt, als vooraf toestemming is verkregen van de maker(s). Het auteursrecht strekt zich uit tot zeventig jaar na het overlijden van de maker(s) en zeventig jaar na publicatie van het betreffende werk. Makers kunnen schrijvers of samenstellers zijn, maar ook illustratoren, fotografen, uitgevers of andere creatieve contribuanten.

In de dagelijkse praktijk van de massadigitalisering is het vaak ondoenlijk om vooraf elke rechthebbende te traceren en om toestemming te vragen. Erfgoedinstellingen proberen dit te ondervangen door bulkovereenkomsten af te sluiten met zogenaamde Collectieve Beheersorganisaties (CBO's), vertegenwoordigende koepelorganisaties zoals Pictoright (beeld) en BUMA/STEMRA (muziek). De monopoliepositie van de CBO's, hun representativiteit en hun transparantie zijn onderwerp van veel discussie, tot in politiek Den Haag aan toe.[25]

De Archiefwet 1995 schrijft voor dat de collectiehoudende instellingen alleen openbaarheidsbeperkingen mogen opleggen indien het gaat om:

a. de eerbiediging van de persoonlijke levenssfeer;
b. het belang van de Staat of zijn bondgenoten;
c. het anderszins voorkomen van onevenredige bevoordeling of benadeling van betrokken natuurlijke personen of rechtspersonen danwel van derden.[26]

De persoonlijke levenssfeer wordt ook nog beschermd door de Wet Bescherming Persoonsgegevens (WBP).[27] Deze richt zich op levende personen. De WBP maakt onderscheid tussen *gewone* en *bijzondere* persoonsgegevens. Gewone persoonsgegevens mogen worden gebruikt, als het gaat om onderzoek met historische, statisti-

sche of wetenschappelijke doeleinden. Wel moet de erfgoedinstelling dan voorzieningen treffen waarmee wordt gegarandeerd dat de verdere verwerking *uitsluitend* gebeurt onder genoemde voorwaarden. In geval van bijzondere persoonsgegevens – gegevens die betrekking hebben op iemands godsdienst of levensovertuiging, ras, politieke gezindheid, gezondheid, seksuele geaardheid, lidmaatschap van een vakvereniging of strafrechtelijke achtergrond – zijn de regels aanzienlijk strenger: deze mogen alleen beschikbaar worden gesteld voor wetenschappelijk onderzoek of statistiek als aan vier voorwaarden is voldaan:

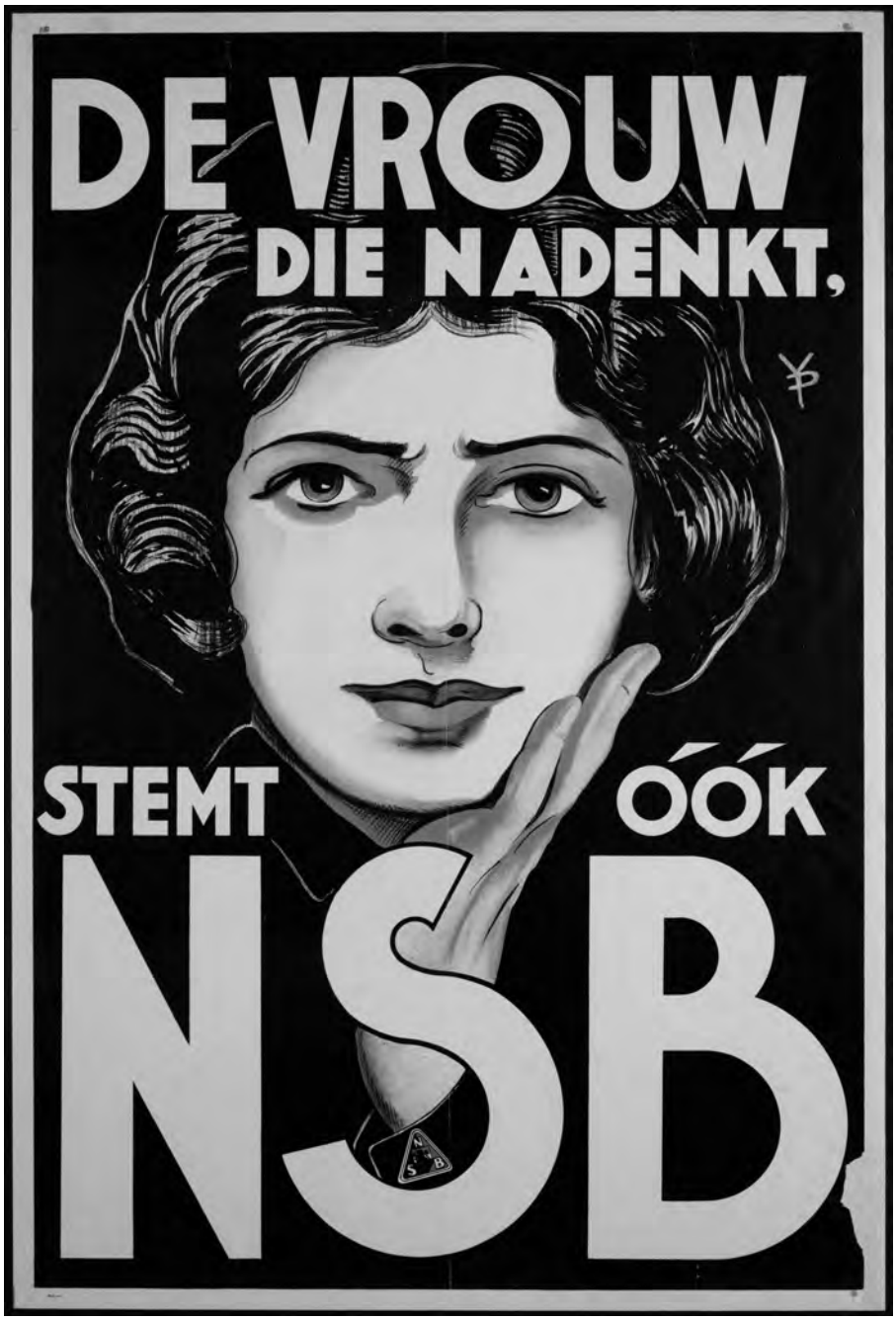
1. Het onderzoek dient een algemeen belang.
2. De verwerking van de gegevens is noodzakelijk voor het betreffende onderzoek.
3. Het vragen van uitdrukkelijke toestemming blijkt onmogelijk of vergt een onevenredige inspanning.
4. De uitvoering van het onderzoek voorziet in goede waarborgen dat de persoonlijke levenssfeer van de betrokkene niet onevenredig wordt geschaad.

Een flink aantal collecties in Nederland – in het bijzonder archieven over de Tweede Wereldoorlog, die veel persoonsgegevens bevatten – hebben te maken met de voorwaarden die door de Wbp worden gesteld. In geval van overleden personen is geen toestemming meer nodig om de persoonsgegevens te gebruiken en publiceren, zolang daarmee maar niet de privacybelangen van anderen worden geschaad (bijvoorbeeld kinderen of kleinkinderen).[28] Bij 'gevoelige' archieven, zoals het archief van de Nationaal-Socialistische Beweging, spelen privacyoverwegingen nog steeds een belangrijke rol bij de beschikbaarstelling.

Standaardisering en contextualisering

Wettelijke openbaarheidsbeperkingen, maar dus ook technologische en budgettaire beperkingen maken dat volledige online beschikbaarstelling van de 'collectie Nederland' vooralsnog toekomstmuziek lijkt. In de tussentijd zijn onderzoekers overgeleverd aan een onoverzichtelijk, versnipperd online erfgoedlandschap, waar digitale collecties op verschillende plekken op verschillende manieren beschikbaar worden gesteld. Dit 'digitale drama' is mede het gevolg van het gebrek aan standaardisering bij digitalisering en beschikbaarstelling van collecties.[29] Maar ook is het aan de onderzoekers om zich meer te bekwamen in zoektechnieken op het web.[30] Waar nu vooral behoefte aan is, zijn integrale zoekingangen, die gebruikers het overzicht bieden en op een eenduidige manier het materiaal ontsluiten. Om dit mogelijk te maken, is standaardisering van digitalisering een absolute randvoorwaarde.

Een andere reden voor meer standaardisering is de duurzame opslag van het digitale materiaal; hoe minder verschillende formaten, hoe efficiënter het beheer op termijn. Inmiddels is het leergeld betaald, en met nog zo veel niet-gedigitaliseerd materiaal in het verschiet, is het moment gekomen om deze standaardisering op



'De vrouw die nadenkt, stemt óók op de N.S.B.' 1935. In de denkbeelden van de NSB speelde de gehuwde vrouw maatschappelijk gezien totaal geen rol. De vrouw moest het huwelijk zien als een opdracht. Zij moest één zijn met de man die trouwde. Zijn moeilijkheden waren ook de hare. Door een leven van toewijding en zelfopoffering bewees de Nederlandse vrouw de gemeenschap een dienst. Het was de taak van de moeder haar kinderen op te voeden tot goede 'volksgenoten'. Tijdens de verkiezingen van 1935 richtte de NSB zich speciaal tot de Nederlandse vrouw. Bron: NIOD Beeldbank WO2

landelijk niveau aan te pakken. Een digitale infrastructuur, bijvoorbeeld door het instellen van enkele gespecialiseerde landelijke e-Depots (Nationaal Archief, Koninklijke Bibliotheek, Nationaal Instituut voor Beeld en Geluid en DANS-KNAW), kan ervoor zorgen dat in ieder geval al het materiaal dat vanaf nu overal wordt gedigitaliseerd, ook op lange termijn gevonden en gebruikt kan worden. Subsidiegevers zouden moeten investeren in een nationale digitale infrastructuur en standaardisering centraal moeten afdwingen.

Uit recent onderzoek blijkt dat studenten en onderzoekers in de universiteitsbibliotheken steeds minder zoeken naar boeken en gedrukte tijdschriften en steeds vaker in databases en op internet.[31] Er is een cultuurslag gaande, die vraagt om een heroriëntatie van erfgoedinstellingen. Zij die er niet in slagen hun collecties in het digitale domein zichtbaar te maken, zullen op termijn de aansluiting met hun beoogde doelgroepen verliezen. Het onderscheid tussen archieven, bibliotheken en musea zal verder vervagen. Toekomstige gebruikers zullen gedigitaliseerd erfgoed – boeken, documenten, foto's, films – vooral zien als 'informatie'. Erfgoedinstellingen zullen eraan moeten wennen dat volledige controle over 'hun' collecties in het digitale domein een illusie is.

Welke rol blijft er dan over voor de erfgoedinstellingen? Zoals informatiehistoricus James Gleick al in zijn recent verschenen boek *The Information* betoogt: '[I]nformation is not knowledge, and knowledge is not wisdom'.[32] We zagen al dat de gebruikers van Borges' Bibliotheek bovenal behoefte hadden aan een intelligente en betrouwbare hulp, die hen wegwijs kon maken in de wirwar van data. Het contextualiseren van de data door middel van adequate beschrijvingen, thesauri en soortgelijke toegangen op basis van gedegen inhoudelijke expertise kan onderzoekers helpen bij het vinden en interpreteren van materiaal. Juist archieven, bibliotheken en musea beschikken over veel kennis op dit gebied. Het is zaak dat erfgoedinstellingen zich bekwamen in nieuwe technologie en hun expertise aanwenden om inhoudelijk een bijdrage te leveren aan de technische uitvoering.

Naast contextualisering en duurzame opslag en toegang, ligt er ook op het gebied van authenticiteit een belangrijke verantwoordelijkheid weggelegd voor de erfgoedsector. Hoe weet je zeker dat er niet gerommeld is met het digitale materiaal dat wordt voorgeschoteld? Hoe kom je erachter waar, wanneer en door wie de 'bron' tot stand is gekomen? Hoe kun je erop vertrouwen dat de verwijzingen naar de bron niet over een paar maanden uitkomen bij een 'page not found'? Archieven, bibliotheken en musea staan doorgaans bekend als degelijke en betrouwbare organisaties. Duurzaam beheer van het gedigitaliseerde erfgoed en gestandaardiseerde beschikbaarstelling zijn taken die prima passen bij hun traditionele rol. Erfgoedinstellingen moeten zichzelf snel heruitvinden en hun sterke punten vertalen naar het digitale domein. En zoals het al eeuwen het geval is, resteert de onderzoeker de schone taak om uit alle informatie 'de wijsheid' te halen.

(Links laatst gecheckt op 26 oktober 2011.)

Noten

1 'De Bibliotheek van Babel' in: *De Aleph en andere verhalen*, Jorge Luis Borges, Ned. vertaling Barber van der Pol (Amsterdam 2010) p. 174.
2 *The Awesome Size of the Internet*, URL: http://theroxor.com/2010/10/28/the-awesome-size-of-the-internet-infographic.
3 Het recentelijk gestarte Europese project ENUMERATE (URL: http://enumerateproject.wordpress.com/) richt zich op het in kaart brengen van statistieken over digitalisering van erfgoedcollecties in Europa.
4 Nederlands Erfgoed: digitaal! Projectplan, p. 13, URL: http://www.faronet.be/files/bijlagen/blog/projectplan_NEDERLANDSERFGOEDDIGITAAL.pdf.
5 Beleidsplan 2010-2013: het weten waard. URL: http://www.kb.nl/bst/beleid/bp/2010/index.html en URL: http://nos.nl/audio/128431-koninklijke-bibliotheek-gaat-digitaliseren.html.
6 Uitgaande van 7.000 scans per meter (DEN rekenmodel, URL: http://www.den.nl/standaard/202/), moeten er 770 miljoen pagina's gescand worden.
7 *De Digitale Feiten. Onderzoek naar de omvang en kosten van gedigitaliseerd cultureel erfgoed. Eindrapportage* (Den Haag, januari 2009), p. 29, URL: http://www.den.nl/bericht/2274.
8 Zie de *Internet Archive Wayback Machine* (URL: http://wayback.archive.org/web/) voor de oude websites van Nederlandse erfgoedinstellingen.
9 URL: http://www.metamorfoze.nl.
10 URL: http://www.statengeneraaldigitaal.nl.
11 Metadata zijn 'data over data', bijvoorbeeld de beschrijving (titel, auteur, enz.) van een boek.
12 URL: http://www.beeldenvoordetoekomst.nl.
13 URL: http://kranten.kb.nl.
14 Astrid Verheusen en Rubrecht Zaat, 'KB en Haags Gemeentearchief. Tekstretrieval in krantencollecties' in: *Informatie Professional* [8] 11, p. 34-37, (2004), URL: http://igitur-archive.library.uu.nl/DARLIN/2005-0526-202104/UUindex.html.
15 Simon Tanner, Trevor Muñoz and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive' in: *D-Lib Magazine*, jrg. 15, no. 7/8, juli/augustus 2009, URL: http://www.dlib.org/dlib/july09/munoz/07munoz.html.
16 URL: http://www.ai.rug.nl/~lambert/Monk-collections-nl.html.
17 URL: http://www.impact-project.eu.
18 Rose Holley, *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers* (Canberra, maart 2009), URL: http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.
19 URL: http://woordentikkertje.manbijthond.nl.
20 URL: http://blog.waisda.nl over Video Labeling Game.
21 Edwin Klijn, *The quality of quantity: Newspaper digitization at the Koninklijke Bibliotheek* (Den Haag, 2009), URL: http://www.ifla.org/files/hq/papers/ifla75/99-klijn-en.pdf.
22 Gemiddelde prijs op basis van steekproef van de auteur bij drie scanbedrijven (juli 2011).
23 Zoektechnologie waarin niet alleen gezocht wordt op exacte treffers, maar meer algemeen ook op mogelijke andere relevante zoekresultaten.
24 Annemarie Beunen en Tjeerd Schiphof, *Juridische Wegwijzer Archieven en Musea online* (Taskforce Archieven/Museumvereniging 2006).
25 URL: http://www.rijksoverheid.nl/documenten-en-publicaties/wetsvoorstellen/2010/07/05/wetsvoorstel-versterking-en-verbreding-van-het-toezicht-op-collectieve-beheersorganisaties-auteursrecht-jus en ook URL: http://beeldenvoordetoekomst.nl/nl/research/digitalisering-audiovisueel-materiaal-erfgoed instellingen-modellen-voor-licenties-en.
26 URL: http://wetten.overheid.nl/BWBR0007376/.

27 Beunen, Schiphof, *Juridische Wegwijzer*, p. 40 e.v.
28 Zie URL: http://blog.iusmentis.com/2011/09/09/geldt-de-privacywet-ook-voor-overleden-personen/.
29 Zie: 'Het digitale drama', *NRC Handelsblad* 10 september 2011 voor discussie over slecht toegankelijke digitale collecties.
30 Zie Ewoud Sanders, Eerste hulp bij e-Onderzoek (Den Haag 2011) voor praktische tips om digitale collecties te doorzoeken, URL: http://www.kb.nl/nieuws/2011/sanders.htm.
31 'Bijna uitgeleend', URL: http://www.nrc.nl/boeken/2011/07/25/bijna-uitgeleend/.
32 J. Gleick, *The Information. A History, a Theory, a Flood* (Londen 2011), p. 409.