Open archieven door crowdsourcing

Ellen Fleurbaay

Alles online vindbaar en beschikbaar is wat steeds meer mensen doodgewoon vinden. Boeken, films, muziek, waarom dan geen archieven? Maar al krijgen we een veelvoud van de budgetten voor scanning waar we nu over beschikken en al krijgen we een verdubbeling van ons personeelsbestand, het zal nooit genoeg zijn om de vele kilometers archieven in onze depots online voor een groot publiek vindbaar en beschikbaar te maken. Willen we die volledige online toegankelijkheid bereiken dan zullen we een beroep moeten doen op de bijdrage en inzet van de grote massa internetgebruikers. Max J. Evans schreef twee jaar geleden een krachtig betoog voor brede publieksinzet in 'The American Archivist' getiteld 'Archives of the People, by the People, for the People[1]. Inmiddels heeft hij bewezen dat hij gelijk heeft. Familysearch.com heeft in een paar jaar tijd 250 miljoen documenten online geïndexeerd.

In Amsterdam zijn we na twee jaar ervaring met de Archiefbank tot een vergelijkbare conclusie gekomen. Onze scanning-on-demand procedure op basis van de archiefinventaris werkt prima voor onderzoekers die vertrouwd zijn met de wijze waarop archieven traditioneel worden ontsloten door middel van inventarissen. Het aanbieden van gescande inventarisnummers is bovendien een ideale manier om te voorkomen dat kwetsbare of diefstalgevoelige stukken naar de studiezaal moeten ter inzage en naarmate er meer is gedigitaliseerd zal het de druk op de (relatief dure) dienstverlening in de studiezaal ontlasten. Maar archiefinventarissen, ook als ze voorzien zijn van scans, zijn GEEN goede manier om het grote publiek te verleiden tot archiefonderzoek. Ook die bekende 28% potentieel geïnteresseerden haken af als ze de archiefstukken via een inventaris moeten raadplegen.[2] Er is maar één oplossing: behalve volledig scannen ook volledig indexeren. Bij moeilijk leesbare teksten hoort daar ook nog volledig transcriberen bij. En, willen we internationaal publiek bereiken, dan zullen we ten slotte nog volledig moeten vertalen.

Bij de beschrijving van digitaal geboren archiefdocumenten gaan we er van uit dat er veel meer metadata nodig zijn om de stukken te beheren en toegankelijk te maken en te houden. Filip Boudrez schrijft in één van zijn vele prima artikelen dat er voor elk digitaal archiefdocument metadata nodig zijn, terwijl dit voor papieren documenten lang niet altijd het geval is[3]. Hij heeft volkomen gelijk wat betreft de digitale documenten. Om die te beheren en te kunnen terugvinden zijn veel metadata nodig. Maar is dat bij de papieren documenten eigenlijk ook niet het geval? Zo niet voor het beheer, dan toch wel voor de toegankelijkheid. Tot op heden hebben we genoegen genomen met archieven die slechts voor een beperkt publiek toegankelijk zijn. Maar als we willen dat onze papieren documenten echt door een breed publiek gebruikt gaan worden, dan hebben we ook van papieren documenten een veelvoud aan metadata nodig. En van papieren documenten hebben we, net als van digitale documenten, ook meerdere representaties nodig. Om te beginnen digitale reproducties, en daarnaast dus liefst ook nog die transcripties en vertalingen.

---

[1] Max J. Evans, Archives of the People, by the People, for the People, in : The American Archivist, 70 (2007), ppp. 387-400. De titel is een vrije variant op de beroemde slotzin van het Gettysburg Adress van Abraham Lincoln.

[2] Koos van Dijken en Natasha Stroeker, Naar een publieksgericht archiefbestel, Zoetermeer, 2003, p.40

[3] Filip Boudrez, Beschrijven van digitaal archief, Antwerpen 2007, *www.edavid.be/docs/**Beschrijven_DigitaalArchief**.pdf*

Met scans van onze documenten online zal het mogelijk zijn grootschalig de metadata te genereren die we nodig hebben om volledige toegang te bieden tot onze archieven. Die metadata moeten op een eenduidige en structurele manier worden verzameld en de kwaliteit van de metadata moet in principe zonder menselijke tussenkomst gecontroleerd kunnen worden. Het heeft niet zo veel zin documenten online te zetten en vrijwilligers te vragen er naar eigen goeddunken een leuke beschrijving bij te maken. Deze werkwijze kan, bijvoorbeeld als het bijzonder beeldmateriaal betreft, via Flickr the Commons een nuttige opbrengst aan metadata genereren, maar voor de meeste archiefdocumenten zou het leiden tot heel ongelijksoortige, ongeordende en overbodige gegevens waar andere onderzoekers maar weinig mee kunnen. Op basis van dat soort gegevens kun je geen zoeksystemen bouwen die een volledige en uniforme opbrengst garanderen. Die gegevens zouden ook niet uitwisselbaar en nauwelijks beheersbaar zijn. De klassieke index biedt uitkomst: metadata in een vooraf bedacht en goed gestructureerd metadataschema, bij voorkeur in de vorm van een open standaard.

Willen we vervolgens op basis van dit schema al die metadata genereren dan hebben we aan tientallen of zelfs honderden vrijwilligers niet genoeg. Het moeten er duizenden zijn. Dat lijkt zo op het eerste gezicht misschien niet haalbaar. Maar als je ziet hoeveel geregistreerde gebruikers de Amsterdamse archiefbank telt, dan moeten dit soort aantallen, ook in een Nederlandstalig gebied, wel haalbaar zijn. Wat we nodig hebben is behalve de scans en de techniek een goed doordacht beleid voor crowdsourcing.

Crowdsourcing is een term die in 2006 is geïntroduceerd door Jeff Howe in een spraakmakend artikel in Wired: http://www.wired.com/wired/archive/14.06/crowds.html Crowdsourcing is een samenvoeging van de termen ‘outsourcing’ en ‘crowd’. Crowdsourcing wordt ingezet wanneer organisaties gebruik willen maken van de ideeën of het werk van de gemeenschap. Er kan een beloning tegenover staan in materiële of in immateriële vorm. ‘De eer van het werk’ is een beloning die misschien mager klinkt, maar voor veel mensen wel degelijk betekenis heeft. Zo ook het gevoel een zinvolle bijdrage te leveren aan onze informatiemaatschappij. Echter, crowdsourcing betekent niet per definitie belangeloos werk. Een beloning in de vorm van scans of scantegoed kan ook heel aantrekkelijk zijn. Dat is een materiële beloning voor de deelnemer aan het project maar kost de archiefdienst niets extra. Zelfs een financiële beloning kan overwogen worden, in de sfeer van de Mechanical Turk van Amazon, the marketplace for work. https://www.mturk.com/mturk/welcome. En zo zijn er meer creatieve manieren te bedenken om deelnemers te motiveren zich in te zetten voor de grootschalige ontsluiting van archieven.

Gaan we als archiefdiensten in Nederland onze collecties via crowdsourcing toegankelijk maken, dan zijn we niet de eersten. En behalve de genealogische sites, waar Familysearch, de site van de Latter Day Saints Church, waar ik eerder naar verwees, een van de bekendsten is, zijn er ook andere voorbeelden waar we van kunnen leren[4]. Een prachtig initiatief is de digitalisering van de archieven van The New York Times. Dat project loopt nu en moet in 2010 gereed zijn. Net als bij de projecten voor krantendigitalisering die we in Nederland kennen worden de scans eerst met OCR (optical character reading) voor het grootste deel leesbaar gemaakt. Daarna worden de gedeelten waar de OCR niet heeft gewerkt omdat de tekst fouten bevatte of simpelweg omdat de krant te vuil of gekreukt was, in de vorm van ‘recaptcha’ voorgelegd aan het grote publiek dat per geval het probleem bekijkt en oplost. Een

[4] een uitgebreide lijst voorbeelden van crowdsourcing is te ivnden in de Engelstalige Wikipedia: http://en.wikipedia.org/wiki/crowdsourcing

captcha is een ‘Completely Automated Public Turin test to tell Computers and Humans Apart’. Het wordt bijvoorbeeld gebruikt om te voorkomen dat computers die spam verzenden toegang krijgen tot systemen die daar niet op zitten te wachten. Recaptcha is een variant ontwikkeld aan de Carnegie Mellon Universiteit. Het selecteert de termen waar OCR-programma’s het niet eens zijn, legt die termen samen met controletermen voor ter interpretatie aan het publiek. Op basis van een puntentoekenning bij gelijke interpretatie wordt vastgesteld welke interpretatie van een probleemgeval de juiste is. Per september 2008 was met deze werkwijze 12.000 manuren per dag aan gratis arbeid gegenereerd.

Willen we de kilometers archieven in onze depots *echt* voor iedereen toegankelijk maken, dan zullen we niet alleen als instellingen beter met elkaar moeten gaan samenwerken, maar we zullen vooral moeten leren samen te werken met de grote hoeveelheid onbenutte maar geïnteresseerde arbeidskracht die via het internet *graag* ingezet wil worden. Dat samenwerken met de internetgebruikers vergt vooral intensieve communicatie. Daarmee bedoel ik geen dure voorlichtingscampagnes, maar een responsief 24/7 netwerk van professionals, deelnemers en klanten die bereid zijn elkaars vragen te beantwoorden, problemen op te lossen en te luisteren naar elkaars suggesties.

Een voorwaarde om het toegankelijk maken van onze archieven met behulp van crowdsourcing succesvol op te zetten is een gebruiksvriendelijk, snel en vlekkeloos werkend systeem met optimale backoffice faciliteiten. Zo moet het mogelijk zijn de ingevoerde metadata met minimale menselijke inspanning grondig te controleren omdat een klant van een overheidsinstelling betrouwbare informatie verwachten mag. Tegelijkertijd moet de productie ook snel online beschikbaar zijn omdat een te traag proces demotiverend bij de deelnemers kan werken. En vanzelfsprekend moet de intensieve communicatie tussen deelnemers, klanten en instellingen efficiënt ondersteund kunnen worden door het systeem. Om een dergelijk te systeem te ontwikkelen is er misschien nóg een nieuwe vorm van samenwerking nodig: met de leverancier.

Max Evans opende bovengenoemd artikel met de opmerking ‘archival institutions have to reinvent themselves’. Ik denk dat hij gelijk heeft.