Open archieven door crowdsourcing: Vele handen maken licht werk
Een project voor grootschalige archiefontsluiting

Projectvoorstel
Stadsarchief Amsterdam
Januari 2010

Inhoudsopgave

Inleiding

In 1999 besloot het Stadsarchief zich te richten op het internet om het publiek toegang te geven tot zijn archieven en collecties. Een traject werd ingezet om alle analoge ontsluitingssystemen te vervangen door digitale systemen. Onder de koepel van het traject *Vernieuwing van de Ontsluitingssystemen van de Archieven en Collecties* zijn vervolgens een aantal projecten georganiseerd waarvan de realisatie van de Archiefbank het sluitstuk was. Het project dat we nu willen starten kan worden gezien als een logisch vervolg op dit traject. Nu de fundamenten zijn gelegd voor volledige online toegang tot systemen is het mogelijk de gebruikers van die systemen ook te betrekken bij het genereren van nieuwe content voor de archieven en voor toekomstige gebruikers.

2. Probleemstelling

2.1 Aanleiding voor het project: publieksgebruik Archiefbank toont aan dat nieuwe en grote gebruikersgroepen alleen worden bereikt met diepgaande ontsluiting via indexen, transcripties en vertalingen

De Archiefbank van het Stadsarchief is een groot succes. De Archiefbank bevat nu ruim 6 miljoen scans en telt circa 30.000 geregistreerde gebruikers. Op verzoek van klanten worden wekelijks gemiddeld 10.000 scans van archiefstukken toegevoegd aan de Archiefbank. De Archiefbank is bekroond met de Computable Award 2008 in de categorie *ICT-project van het jaar*. Voor deze Award waren in totaal vijf projecten genomineerd. Het Stadsarchief werd verkozen boven 4 grote commerciële bedrijven: de NS, de KLM, chemiegigant DSM en bierbrouwer Inbev. Internationaal heeft de Archiefbank erkenning gekregen met de toekenning van de *Best Archives on the Web-award*, op voordracht van een jury van Amerikaanse vakgenoten.

Nu de Archiefbank ruim anderhalf jaar online is, is ook de klantenkring van de Archiefbank duidelijk in beeld. Het gaat enerzijds om een kleine groep onderzoekers die diepgaand historisch onderzoek verricht en intensief gebruik maakt van de ‘scanning-on-demand-formule’. Deze groep is vertrouwd met de wijze waarop archieven traditioneel worden

ontsloten door middel van inventarissen. Anderzijds gaat het om een veel grotere groep gebruikers die vrijwel uitsluitend gebruik maakt van scans geproduceerd in de grootschalige projecten waarbij het archiefmateriaal niet alleen is gescand, maar ook door middel van indexen is ontsloten op namen van personen, plaatsnamen of op onderwerpen. Zeer veel mensen zijn geïnteresseerd in scans van archiefstukken met informatie over hun voorouders, over het huis waar ze in wonen of over bijvoorbeeld de voetbalclub, de paardentram of de Bijlmerramp. Maar die informatie moeten ze eenvoudig en snel kunnen vinden door te zoeken op achternaam, adres of onderwerp en als het gaat om oude handschriften wil men liefst ook een transcriptie en een hertaling of vertaling.

Was de scheiding tussen enerzijds de kleine groep diepgravers en anderzijds de grote groep snuffelaars in het verleden al in beeld gebracht bij marktonderzoeken, de gebruiksgegevens van de Archiefbank bevestigen dit beeld overtuigend. De conclusie die getrokken kan worden uit deze gegevens is dat *als* we de in potentie grote groep belangstellenden voor de historische informatie in archieven willen bereiken, dan *moeten* we er voor zorgen dat die gegevens eenvoudig te vinden zijn en gemakkelijk gelezen kunnen worden.

Dat wil zeggen eenvoudig te vinden door middel van simpele zoeksystemen volgens de inmiddels ingeburgerde methode van zoeken op ‘wie-wat-waar-wanneer’, of te wel door middel van indexen op naam, adres of onderwerp en gemakkelijk te lezen dankzij transcriptie en vertaling.

2.2 Probleem: héél veel werk dat alleen uitgevoerd kan worden bij systematische en grootschalige aanpak

Echter, de kosten voor het maken van indexen, transcripties en vertalingen van archiefbronnen zijn extreem hoog omdat het heel veel werk is en soms een gedegen en specialistische kennis van oud schrift vergt. De budgetten van archiefdiensten gaan uit van de wettelijke taak van het beheren van de archieven in goede, geordende en toegankelijke staat. De budgetten zijn toereikend voor het conserveren en inventariseren van de archieven, maar niet voor het scannen en indexeren, laat staan voor transcripties en vertalingen. Vinden van extra geld voor het scannen van de bronnen lukt wellicht met enige moeite, maar vinden van een veelvoud van dat bedrag voor de nog veel kostbaarder indexering, transcriptie en vertaling is onmogelijk. Dat vergt grootschalige inzet van vrijwilligers.

Binnen archiefdiensten wordt al jaren van dit zelfde arbeidspotentieel van vrijwilligers gebruik gemaakt bij projecten. Dat zijn projecten waar groepjes van een handjevol tot enkele tientallen mensen aan werken. Zo’n project omspant meestal niet meer dan enkele jaren. Resultaat is dat veel archiefdiensten beschikken over talloze ladebakjes met grotere en kleinere kaartsystemen, allemaal met een eigen systematiek. En in het digitale tijdperk groeit het aantal databases met ieder een eigen systematiek en eigen vormen van standaardisering nog sneller. De kwaliteit van de indexen is bovendien zeer wisselend omdat intensieve controle ondoenlijk is. Transcripties en vertalingen van archiefbronnen zijn alleen beschikbaar van een aantal van de historisch meest bekende en waardevolle documenten.

Dat grootschalige archiefontsluiting de enige mogelijkheid is om nieuwe en brede groepen potentiële archiefgebruikers te bereiken en dat we grootschalige archiefontsluiting alleen voor elkaar kunnen krijgen met dito grootschalige inzet van de gebruikers is een conclusie die ook

getrokken werd door Max J. Evans in ‘Archives of the People, by the People, for the People (American Archivist, 2007, pp. 387-400)

2.3 Oplossing: online indexeren en bewerken met grootschalige inzet van vrijwilligers

Dankzij het internet is het mogelijk echt grootschalige projecten te organiseren. Wikipedia is daarvan het meest bekende bewijs. Indexen, transcripties en vertalingen kunnen worden gemaakt vanaf de scans van archiefbronnen en daar kunnen heel veel mensen tegelijkertijd aan werken.

Deze zeer voor de hand liggende oplossing van online indexeren is vanzelfsprekend al eerder bedacht en toegepast. Het Gemeentearchief in Den Haag heeft als eerste microfiches gescand en via de website kunnen gebruikers daar een index bij maken. De website voor genealogisch onderzoek Genlias werkt op vergelijkbare wijze.

Maar de systemen voor de gegevensinvoer bij deze voorbeelden zijn onderwerpsgebonden en niet gebaseerd op een open standaard die bruikbaar is voor alle archiefbronnen en beschikbaar voor alle archiefdiensten. En de bij deze projecten gebruikte tools voor gegevensinvoer zijn gekoppeld aan de eigen website en de geproduceerde gegevens zijn niet eenvoudig uitwisselbaar met andere websites. Bovendien laten de tools veel te wensen over op het gebied van snelheid en gebruiksvriendelijkheid, waardoor veel vrijwilligers afhaken. *Echt* grootschalige ontsluiting, transcriptie en vertaling van archiefbronnen door een menigte aan webgebruikers via internet is in Nederland daardoor tot nu toe nog geen gangbare praktijk. In Amerika heeft de auteur van bovengeciteerd artikel, Max Evans, intussen wel een dergelijk systeem van grootschalige ontsluiting gerealiseerd waarbij inmiddels 250 miljoen documenten zijn geïndexeerd.

Toch is het Stadsarchief er van overtuigd dat het realiseren van systematische en grootschalige inzet mogelijk is, mits aan een aantal voorwaarden is voldaan. Eerste voorwaarde is het beschikbaar zijn van faciliteiten die optimaal gebruiksvriendelijk zijn en vlekkeloos werken. Een tweede essentiële voorwaarde is een goed doordacht beleid voor crowdsourcing met de juiste mix van materiële en immateriële beloning voor de vrijwilligers. Via dit project kunnen deze voorwaarden gerealiseerd worden.

3. Projectvoorstel

3.1 Doelstelling: grootschalige diepteontsluiting van archieven door inzet van de menigte webgebruikers

Het doel van dit project is om faciliteiten te creëren en beleid te ontwikkelen om de vele kilometers archieven in onze archiefbewaarplaatsen beschikbaar te maken met behulp van grootschalige inzet van vrijwilligers. Onderdeel van het project moet zijn een pilot van voldoende omvang om zowel de technische faciliteiten te testen als te experimenteren met verschillende vormen van crowdsourcing.

3.2 Voorwaarde: open uitwisselingsformaat

Voorwaarde om de technische faciliteiten te kunnen creëren is een goed gedocumenteerd open uitwisselingsformaat voor indexgegevens. Voor het beschrijven van archieven en het maken van archiefinventarissen is een internationale standaard vastgesteld in XML-formaat, genaamd EAD: encoded archival description. Voor het maken van indexen zijn er nog geen vergelijkbare standaarden. Willen we komen tot grootschalige archiefontsluiting dan is een standaard uitwisselingsformaat een vereiste. XML is daarbij de meest voor de hand liggende keuze.

In de praktijk zijn er veel overeenkomsten in de wijze waarop indexen gemaakt worden. In grote lijnen gaat het om 4 categorieën van gegevens: personen, topografie, onderwerpen en datums (wie, waar, wat, wanneer). Daarnaast is de relatie (rol) van de gegevens onderling en de relatie van de gegevens tot de bron relevant. Uitgaand van deze overeenkomsten is het goed mogelijk een formaat vast te stellen dat deze basisgegevens bevat. Het formaat moet voldoende flexibel zijn om uitbreidingen toe te staan wanneer men meer gegevens wil toevoegen.

Het Stadsarchief werkt inmiddels al met een uitwisselingsformaat voor indexgegevens omdat we afzonderlijke systemen hebben voor het beheren van metadata intern en voor het raadplegen via de website. Dit is een XML-formaat gebaseerd op de eerder genoemde eigenschappen en voorwaarden. Op basis van de positieve ervaring die we hierbij hebben opgedaan denken we dat het relatief eenvoudig is te komen tot een uitwisselingsstandaard dat een brede toepassing kan krijgen.

Binnen het project wil het Stadsarchief op basis van de eigen ervaring onder leiding van een XML-deskundige een uitwisselingsformaat opstellen dat, wanneer het formaat in de pilot voldoet, zal worden gedocumenteerd en gepubliceerd en vrij beschikbaar komt. Onderdeel van het project is communicatie over en presentatie van dit formaat.

3.3 Eisen: snel en gebruiksvriendelijk platform

De technische omgeving voor het grootschalig online aan de hand van scans creëren, verwerken, beheren, uitleveren en belonen van indexgegevens, transcripties en vertalingen moet voldoen aan een groot aantal hoge eisen, zoals:

- er moet tegelijkertijd met zeer veel mensen aan zeer veel scans, snel en doeltreffend gewerkt kunnen worden
- de door deze mensen geproduceerde gegevens moeten optimaal, en waar mogelijk zonder menselijke tussenkomst, gemonitored en ge-audit kunnen worden
- betrokkenen, zowel medewerkers als deelnemende vrijwilligers, moeten in staat zijn elkaar hulp te bieden en elkaars hulp in te roepen
- resultaten moeten in XML geëxporteerd kunnen worden
- resultaten moeten snel beschikbaar en leverbaar zijn omdat het motiverend werkt voor de deelnemers en omdat levering van resultaten aan derden deel uitmaakt van de projectfinanciering

Specifieke onderdelen waar het backoffice van de technische omgeving over moet beschikken:

- een tool voor import, beheer en export van scans, indexgegevens, transcripties en vertalingen conform de eisen van het uitwisselingsformaat
- een viewer met faciliteiten voor inzoom, contrastregeling, download van scans

- opties voor samenstelling van groepen en toekennen van functies / autorisaties aan deelnemers van groepen
- optie voor dubbele of meervoudige invoer van dezelfde gegevens en geautomatiseerde controle van de invoer, te bepalen per (deel)project door projectleider
- opties voor diverse, nader te bepalen vormen van beloning van vrijwilligers
- diverse periodieke rapportages van de projectresultaten en van de verkoop aan klanten
- een gebruiksvriendelijk zoeksysteem op de indexgegevens van het uitwisselingsformaat
- een webwinkel voor de online verkoop van scans, transcripties en vertalingen
- optimale communicatie mogelijkheden tussen beheerders, vrijwilligers en gebruikers

Als de pilot positief wordt beoordeeld moeten deze faciliteiten op de markt beschikbaar komen niet alleen voor de aan het project deelnemende instellingen, maar ook voor andere partijen. Gezocht wordt naar een leverancier die wil optreden als partner in het project door zelf risico te dragen in de ontwikkeling van deze faciliteit. Een leverancier die optreedt als partner zal bij het tot stand komen van het product actief meedenken over de toekomst van zijn product en het geschikt willen maken voor meer klanten, wellicht ook uit andere branches. De kans dat het daardoor een zelfstandig inzetbaar en sterker product wordt is groot. De leverancier krijgt wel de garantie dat er een serviceovereenkomst voor gebruik van het product wordt afgesloten voor de volledige projectperiode van 3 jaar. Voordeel voor de leverancier is de inbreng van kennis van zijn toekomstige klanten in de ontwikkeling van een product en de gegarandeerde afname van het product gedurende een periode van 3 jaar.

3.4 Pilot: minimaal 1 miljoen scans en 10 miljoen gegevens

In overleg met de projectpartners wordt bepaald welke archiefbestanden deel gaan uitmaken van de pilot. De archiefbestanden moeten bij voorkeur landelijk uniform aanwezig zijn.

Een geschikt archiefbestand voor het aspect van de indexering zijn bijvoorbeeld de 19^{e}-eeuwse inschrijvingsregisters voor de militaire dienstplicht die na de Franse tijd is ingevoerd, de zogenaamde militieregisters. Deze registers zijn gedurende bijna 100 jaar in heel Nederland op redelijk uniforme wijze gevormd. Geregistreerd in deze bron is de gehele mannelijke Nederlandse bevolking van 19 jaar. De bron is voor genealogen bijzonder interessant, maar wordt nog niet breed gebruikt omdat hij zelden geïndexeerd is en dus onbekend. Met name voor de eerste helft van de 19de eeuw zal veel vastgelopen stamboomonderzoek dankzij beschikbaar komen van deze bron kunnen worden vlotgetrokken. Belangrijke motivatie voor vrijwillers! Bij de inschrijving zijn bovendien persoonsgegevens opgenomen die verder moeilijk of niet te vinden zijn.

Scannen en indexeren van deze bronnen is niet bijzonder ingewikkeld omdat het voor het merendeel vastbladige registers zijn en de tekst goed leesbaar is. De militieregisters zijn bovendien in zeer veel overheidsarchieven in Nederland beschikbaar. Voordeel daarvan is dat een werkelijk grootschalig project opgezet kan worden zodat in principe iedere Nederlandse archiefinstelling aan het project kan deelnemen.

Op één pagina uit de registers staan meerdere inschrijvingen zodat meer indexrecords aan één scan gekoppeld zullen worden. Daarmee is de scan potentieel interessant voor meerdere kopers, hetgeen gunstig is voor de financiering van het project.

In de eerste fase van het project wordt onderzocht welke overige archiefbestanden in aanmerking komen voor opname in het project. Criteria zijn diverse typen bestanden zodat behalve indexering op persoonsnamen ook indexering op topografie en op onderwerp plaatsvindt. Daarnaast moeten proeven met transcriberen / vertalen worden genomen. Vrijwilligers met verschillende soorten belangstelling moeten bij het project betrokken kunnen worden.

Van belang is dat de pilot van voldoende omvang is en een voldoende ruime periode betreft om faciliteiten en keuzes te testen. Een omvang van 1 miljoen scans en 10 miljoen gegevens is niet overdreven.

3.5 Pilot: honderden vrijwilligers

Diverse projecten hebben bewezen dat er een groot arbeidspotentieel is dat bereid is op vrijwillige basis online kennis en arbeid in te zetten, maar wanneer daartegenover een vorm van beloning staat neemt het aantal deelnemers aan projecten exponentieel toe. De beloning kan materieel of immaterieel zijn. Een voor de hand liggende vorm van beloning is scans in ruil voor gegevens, bijvoorbeeld indexeer 10 scans in ruil voor 1 scan. Kost het indexeren 5 minuten en zou het kopen van 1 scan 25 cent kosten, dan is de beloning een aantal scans ter waarde van anderhalve euro per uur. Dit is geen ‘dik betaald’ werk, maar in 5 minuten een scan verdienen kan toch aantrekkelijk zijn.

De beloning kan ook immaterieel zijn. Mensen schrijven een product graag op hun naam, of publiceren uit naam van hun samenwerkingsverband, bijvoorbeeld ‘Index op de dienstregeling van de Amsterdamse tram in de 19^{e} eeuw, gemaakt door de Vereniging voor behoud van de Historische Tram in Amsterdam’. Wellicht is het ook mogelijk ‘beloningen’ in de zin van certificaten of getuigschriften te verstrekken in samenwerking met opleidingen of reïntegratieprojecten.

Onderzocht kan worden of er ook sprake kan zijn van ‘semi-vrijwillig’ werk. Ook daar is online inmiddels veel ervaring meer, denk bijvoorbeeld aan de Mechanical Turk van Amazon.

Voor het onderzoek naar en de besluitvorming over het vrijwilligersbeleid wordt samenwerking gezocht met projectpartners die ervaring hebben met online communicatie met de doelgroepen en / of met vrijwilligerswerk binnen historische context. Het Stadsarchief heeft op dit moment verschillende voorbereidende contacten aangesproken die enthousiast meedenken.

4. Projectorganisatie

4.1 Professionele organisatie

Voor een succesvol verloop van een langdurig project is een heldere organisatiestructuur met duidelijk vastgelegde verantwoordelijkheden een eerste vereiste. En er moet gewerkt worden binnen een vastgesteld kader van projectfasering en projectplanning met vooraf vastgestelde oplevering van deelproducten en periodieke rapportages. Het Stadsarchief heeft op dit gebied een goede naam: projecten worden conform doelstelling afgerond binnen het beschikbare budget en zonder extreme vertraging.

Dit project is in verhouding tot de voorgaande projecten van het Stadsarchief bijzonder omvangrijk en gecompliceerd, met name omdat er bij dit project hopelijk zal worden samengewerkt met een groot aantal externe projectpartners. Een goede organisatie van de productontwikkeling binnen het project is daarom van het grootste belang, maar nóg belangrijker is goede organisatie van en structurele aandacht voor de communicatie. Communicatie is misschien wel de belangrijkste succesfactor van dit project.

4.2 Samenwerkingsverbanden met meerdere projectpartners

Het Stadsarchief wil bij de uitvoering van dit project graag meerdere externe projectpartners betrekken in diverse soorten samenwerkingsverbanden voor de verschillende onderdelen van het project.

Bij de ontwikkeling van het uitwisselingsformaat bijvoorbeeld wordt samenwerking gezocht met archiefbeherende instellingen die in de toekomst wellicht met het formaat gaan werken. Daarnaast is, ook voor dit product, inbreng van de gebruikers van groot belang. Daarom zal, zeker ook voor de inhoudelijke aspecten van het uitwisselingsformaat, samenwerking met gebruikers worden gezocht. Het Stadsarchief heeft goede ervaring met inbreng van zijn gebruikersraad bij dit soort beslissingen.

Voor de ontwikkeling van de technische faciliteiten willen we met de leverancier van dat product bij voorkeur geen klassieke verhouding opdrachtgever – opdrachtnemer, maar een verhouding die wij omschrijven als projectpartner. We willen zoeken naar een leverancier die bereid is op eigen risico het product financieel te ontwikkelen, te onderhouden en in de markt te zetten op basis van service overeenkomsten. De leverancier kan rekenen op intensieve en kwalitatief hoogwaardige inbreng in de ontwikkeling van het product en een gegarandeerde afname van het product gedurende minimaal de 3-jarige duur van het project. Dankzij deze samenwerkingsvorm komt het product breed beschikbaar na afloop van het project. Ontwikkeling in open source context lijkt voor deze technische faciliteiten niet haalbaar omdat het product daarvoor te complex is en verdere ontwikkeling dan niet gepland kan worden.

Bij het opzetten van het vrijwilligersbeleid voor de pilot zoeken we samenwerking met online gebruikersgroepen of individuele gebruikers die online service aan collega onderzoekers aanbieden. Op gebied van genealogie kan gedacht worden aan vertegenwoordigers van bijvoorbeeld voorouders.net, stamboomforum.nl en geneaknowhow.net. Afhankelijk van de keuze van de overige archiefbestanden worden overige bijpassende projectpartners

uitgenodigd. Zo heeft het Stadsarchief bijvoorbeeld goede ervaringen met maritieme onderzoekers en met kerkelijke onderzoekers.

Aan de pilot van het project worden alle daarvoor in aanmerking komende Nederlandse archiefinstellingen uitgenodigd deel te nemen. Voorwaarde voor deelname aan de pilot is dat instellingen binnen een afgesproken termijn de voorbereidende werkzaamheden voor het scannen van de eigen archiefbestanden realiseren op een binnen het project af te spreken wijze. Van de instellingen wordt een bijdrage gevraagd per scan die minder is dan de gemiddelde kostprijs van een scan. Het levert de instelling bovendien een archiefbestand op dat niet alleen is gescand, maar ook geïndexeerd. De kostprijs van de scans kan laag zijn bij aanbesteding van 1 miljoen scans en dankzij het vrijwilligerswerk zijn er geen kosten voor indexering. De inkomsten van de verkoop van de scans dragen gedurende de projectperiode bij aan de financiering van het project.

4.3 Projectteam

Het Stadsarchief wil voor de aansturing van het project een projectteam instellen met een drietal permanente teamleden: een projectleider, een projectsecretaris en een productiebegeleider. In sturing van het project en vakinhoudelijke ondersteuning voorziet het Stadsarchief.

Afhankelijk van de fase van het project en de daarbij benodigde inzet en kennis kan het team worden uitgebreid met verschillende medewerkers of adviseurs. Voor de (technische) controle van de scans bijvoorbeeld zijn gedurende korte periodes meerdere mensen tegelijkertijd nodig en voor de ontwikkeling van het uitwisselingsformaat is advies van een XML-deskundige gewenst.

4.4 Projectfasering

Het project kent 4 fasen:

1. Voorbereidingsfase: ca. 6 maanden
    a. Fondsenwerven en vaststellen definitieve begroting en financiering
    b. Werven vaste bemanning van het projectteam
    c. Werven projectpartners en contracten aangaan met externe leveranciers
    d. Nader uitwerken projectplanning
2. Eerste uitvoeringsfase: ca. 1 jaar
    a. Ontwikkelen uitwisselingsformaat
    b. Ontwikkelen technische omgeving
    c. Ontwikkelen vrijwilligersbeleid
    d. Vaststellen inhoud pilot en productie 1 miljoen scans
    e. Externe communicatie gebruikers en vrijwilligers
3. Tweede uitvoeringsfase: ca. 2 jaar
    a. Technische omgeving beschikbaar voor productie
    b. Grootschalige indexering
4. Afronding en evaluatie: ca. 3 maanden