7 vragen & antwoorden over crowdsourcing

Vanuit het Meertens Instituut heeft Nicoline van der Sijs jarenlange ervaring opgedaan met crowdsourcingprojecten. De antwoorden van zeven veelgestelde vragen heeft ze op een rij gezet, zodat anderen van het voortschrijdend inzicht kunnen profiteren.

Nicoline van der Sijs

Sinds juni 2007 heb ik zes crowdsourcingprojecten georganiseerd, die telkens draaiden rond het transcriberen van oudere Nederlandse teksten. Dat was avant la lettre, want in 2007 was het woord *crowdsourcing* in het Nederlands nog onbekend – we spraken toen nog van *vrijwilligersproject*. De term *crowdsourcing* vierde op dat moment in het Engels net zijn eerste verjaardag: in juni 2006 had journalist Jeff Howe de term gemunt in zijn artikel 'The Rise of Crowdsourcing', dat verscheen in het Amerikaanse tijdschrift *Wired Magazine*. Inmiddels is de term crowdsourcing onder wetenschappers en erfgoedinstellingen als musea en bibliotheken gemeengoed. Dat is niet uit idealisme, maar uit bittere noodzaak. Subsidiegevers verstrekken nauwelijks nog geld voor het verzamelen en rubriceren van data, terwijl wetenschappers juist steeds meer behoefte hebben aan omvangrijke en betrouwbare gegevens: big data. Erfgoedinstellingen willen hun collecties graag ontsluiten, maar hebben daarvoor geen menskracht. Op dat moment komt de vraag in beeld of crowdsourcing kan worden ingezet: kunnen buitenstaanders een deel van de taken uitvoeren?

Toch zijn er nog maar relatief weinig onderzoeks- en erfgoedinstellingen die daadwerkelijk overgaan tot crowdsourcing, zeker als het gaat om het maken van betrouwbare transcripties, wat mijn corebusiness is. Onderzoekers en instellingen blijken namelijk op te zien tegen de organisatie van een crowdsourcingproject en vrezen dat het erg tijdrovend is en een zware wissel trekt op de instelling. Vanwege de ervaring die ik inmiddels heb opgedaan, vragen ze me regelmatig om advies. Hieronder heb ik de antwoorden van enkele veelgestelde vragen eens op papier gezet, volgens het beproefde principe van wie, wat, waar, wanneer, waarom en hoe. Waar dat relevant is, zal ik aan de hand van praktijkvoorbeelden laten zien welke lessen we op het Meertens Instituut hebben geleerd, zodat anderen van het voortschrijdend inzicht kunnen profiteren. Er zijn natuurlijk meer Nederlandse instellingen die crowdsourcen, bijvoorbeeld Beeld en Geluid, streekarchieven, Wikipedia en VeleHanden (zie het artikel 'Vele handen maken licht werk' in *IP* 5, 2015). Ik beperk mij hier tot de opgedane ervaring met het transcriberen van moeilijk leesbare teksten.

1 WAAROM?

De vraag naar het waarom heb ik al beantwoord: in principe ligt aan de basis van ieder crowdsourcingproject een concrete maar onvervulbare wens naar bepaalde gegevens. Die wens garandeert een grote motivatie bij de organisator om van de crowdsourcing een succes te maken. Zo'n wens was voor mij ook de reden om in 2007 met crowdsourcing te beginnen: ik wilde dolgraag onderzoek doen naar de taalkundige invloed van de Statenvertaling uit 1637, maar die tekst was niet digitaal beschikbaar en er was geen subsidiegever te vinden die de 2,6 miljoen in gotisch schrift gezette woorden door professionals wilde laten overtikken. Met honderd vrijwilligers was het overtikken, corrigeren en op een website plaatsen binnen een jaar gepiept. Mijn collega Hans Beelen en ik deden dat door instructies en porties Wordbestanden per e-mail heen en weer te sturen – een methode die werkt, maar inmiddels weten we dat het praktischer is om het werk via een webapplicatie op internet aan vrijwilligers aan te bieden.

Inspirerende voorbeelden van crowdsourcing

> **easy.dans.knaw.nl/ui/datasets/id/easy-dataset:34380**, met 86 getranscribeerde oude teksten

> **www.gekaaptebrieven.nl**, met transcripties van 8000 gekaapte brieven uit de zeventiende en achttiende eeuw

> **www.meertens.knaw.nl/ewnd**, met 35 getranscribeerde en taalkundig verrijkte dialectwoordenboeken

2 WAT?

Een lastiger vraag is wat: wat voor werk kunnen vrijwilligers doen? Het maximalistische antwoord is dat er voor iedere denkbare klus geschikte vrijwilligers te vinden zijn, het minimalistische antwoord is dat crowdsourcing alleen geschikt is voor heel eenvoudige klussen.

De balans ligt in het midden: hoe specialistischer het werk, hoe kleiner de potentiële crowd, en hoe meer tijd de werving van geschikte vrijwilligers en de begeleiding kosten. Maar te eenvoudig en eentonig werk is ook minder geschikt, want dan wordt het snel saai. En dat leidt weer tot verloop onder de vrijwilligers: veel van hen doen mee omdat ze een intellectuele uitdaging zoeken. Het meest geschikt is een project dat een beroep doet op algemene ontwikkeling, inzicht en vaardigheden, en waar de vrijwilligers ook nog wat van kunnen leren.

Voor ingewikkelder klussen kan worden overwogen om de vrijwilligers een cursus aan te bieden, maar zorg er wel voor dat het werk ook gedaan kan worden zonder dat men die cursus heeft gevolgd, want lang niet alle vrijwilligers zullen in staat of bereid zijn een cursus bij te wonen.

3 WIE?

Iedereen kan vrijwilliger worden, maar niet iedere vrijwilliger kan elke klus aan. Sommige organisaties willen daarom vrijwilligers van tevoren selecteren, maar mijn ervaring is dat een dergelijke voorselectie onnodig en overbodig is: vrijwilligers kennen hun eigen beperkingen heus wel, en bij een voorselectie wordt noodzakelijkerwijs alleen geselecteerd op cv en praktijkgerichte ervaring, terwijl juist de algemene kennis, het leervermogen en de motivatie van vrijwilligers doorslaggevend is in de bijdrage die zij leveren. Sterker nog: de ervaring leert inmiddels dat vrijwilligers van wie je dat qua achtergrond niet zou verwachten, dikwijls de omvangrijkste en belangrijkste bijdrages leveren.

Bij moeilijker werk treedt automatisch een selectiemechanisme op, waardoor vanzelf de vrijwilligers overblijven die het werk aankunnen. Dat is gebleken toen vrijwilligers handgeschreven brieven van zeevarenden uit de zeventiende en achttiende eeuw moesten ontcijferen. Na aanmelding kreeg iedere vrijwilliger automatisch een portie toebedeeld. Dat was achteraf gezien te snel: ongeveer een derde van de vrijwilligers haakte direct af toen ze de complexiteit van het werk zagen. Belangrijk is daarom vrijwilligers voordat ze zich aanmelden goede informatie te geven over het werk dat ze staat te wachten. Ook bij goede informatie vooraf kent ieder crowdsourcingproject een percentage afvallers. Dat is geen probleem: als de crowd voldoende massa heeft, loopt een project gewoon door. Sowieso geldt als vuistregel dat tien procent van de vrijwilligers negentig procent van het werk doet. Dat is absoluut geen bezwaar. Het is trouwens de vraag of dat in een normale werksituatie heel anders is...

Iedereen kan dus in principe vrijwilliger worden, maar hoe bereik je als organisatie zoveel mogelijk potentiële vrijwilligers? Veel instellingen en onderzoekers zijn geneigd daarvoor in hun eigen netwerk te zoeken, maar dat is in tegenspraak met de term crowdsourcing: je eigen netwerk ken je al, maar je wilt nu juist de (onbekende) menigte, de crowd, bereiken. Daarvoor moet je het project via zoveel mogelijk kanalen bekendmaken. Met nieuwsbrieven, internet en sociale media kan dat tegenwoordig vrij eenvoudig.

Op het moment dat potentiële vrijwilligers zich aanmelden, is het nuttig hen enige vragen te stellen over hun achtergrond en ervaring. Op basis daarvan kan de organisatie namelijk geschikte vrijwilligers in een later stadium vragen het werk van anderen te corrigeren. Alternatieve werkwijzes zijn: alles door twee vrijwilligers te laten verwerken en een controleur naar de verschillen te laten kijken, of verwerkte porties te laten rouleren tussen de vrijwilligers.

4 WAAR?

De essentie en de sterkte van crowdsourcing is dat mensen zelf kunnen kiezen waar, wanneer en hoeveel ze werken: de menigte is te groot om naar een instelling te halen, dus mensen werken van huis uit. Dat betekent wel dat de organisatie extra aandacht moet besteden aan de communicatie met en tussen de vrijwilligers: er moet een alternatief gezocht worden voor het gebrek aan direct persoonlijk contact. Correspondentie kan uiteraard plaatsvinden via e-mail, maar voor de organisatie kost dat relatief veel tijd. Voor het project rond de digitalisering van de historische kranten van de Koninklijke Bibliotheek hebben we daarom een forum ingericht op de website. En dat blijkt een prima oplossing: iedereen die meewerkt kan berichten in het forum lezen en versturen. Bovendien maken we zo optimaal gebruik van de wijsheid van de crowd: er is bijna altijd wel iemand die het juiste antwoord op een vraag kent.

In een klein jaar zijn er 774 berichten gewisseld, waaronder ook faits divers; zo geeft een verordening uit 1647 een tip over hoe bankiers in toom te houden: 'Alle personen welcke voordachtelijcken [= opzettelijk] Bancqueroet maken, sullen een lichte groene snoer op de mantelkrage genaeyt werden'. Het uitwisselen van dergelijke leuke weetjes is goed voor het groepsgevoel.

De motivatie van de vrijwilligers wordt versterkt door bijvoorbeeld één keer per jaar een bijeenkomst te organiseren. Dat maakt het werk minder anoniem: mensen leren de gezichten achter de namen kennen. Omdat natuurlijk niet alle vrijwilligers kunnen of willen komen bij zo'n bijeenkomst, werkt het goed om daarnaast een jaarlijkse kleine attentie voor de vrijwilligers te verzorgen, bijvoorbeeld rond de jaarwisseling, als symbolisch dankgebaar.

Figuur 1: De voorkant van de editor voor het digitaliseren van kranten van de Koninklijke Bibliotheek; de homepage voor inloggen

Figuur 2: De voorkant van de editor voor het digitaliseren van kranten van de Koninklijke Bibliotheek; de pagina na inloggen

Figuur 3: De achterkant van de editor voor het digitaliseren van kranten van de Koninklijke Bibliotheek

5 WANNEER?

Een crowdsourcingproject kan op ieder moment starten, maar daaraan moet een gedegen voorbereiding voorafgaan. Mijn ervaring is dat het meeste werk en de grootste kosten bij crowdsourcing zitten in de voorbereiding: het nadenken over de workflow, de taak die van vrijwilligers wordt verwacht, de manier waarop het werk wordt aangeboden, verdeeld en gecontroleerd.

Een belangrijke voorwaarde om te kunnen starten is dat er scans, foto's, beelden of geluidsbestanden op internet beschikbaar zijn voor bewerking door de vrijwilligers. Een les die ik daarbij heb geleerd, is dat er ruim voldoende scans beschikbaar moeten zijn, zodat vrijwilligers niet hoeven te worden teleurgesteld omdat het werk door een onverwacht grote toeloop van vrijwilligers al op is. Het blijkt dat organisaties de belangstelling en werklust van vrijwilligers nogal eens onderschatten. Dat is zonde van de moeite en kosten die het opzetten van een crowdsourcingproject met zich meebrengt.

6 HOE?

Uiteindelijk draait alles om de vraag van het *hoe*: hoe zet je het werk voor de vrijwilligers op? Om te beginnen moet er een trekker zijn die het aanspreekpunt is voor zowel de vrijwilligers als de organiserende instelling. De trekker schrijft instructies voor de vrijwilligers en beantwoordt vragen.

Een webapplicatie neemt veel werk uit handen. Maar die moet wel worden ontwikkeld (en dat kost geld), ze moet foolproof zijn, en ook digibeten moeten er intuïtief mee kunnen werken: veel vrijwilligers zijn op leeftijd en voor computergebruik afhankelijk van hun kleinkinderen. De applicatie moet het werk in hapklare brokken serveren. Daarvoor moet van tevoren de grootte van een portie (zoals de hoeveelheid scans) worden bepaald, en de volgorde waarin het te bewerken materiaal wordt uitgedeeld, bijvoorbeeld chronologisch of thematisch.

De voorkant van de applicatie wordt ingericht voor de vrijwilligers en bestaat uit verschillende modules (zie als voorbeeld de figuren 1 en 2). Ik noem de belangrijkste. Een bladermodule biedt belangstellenden de mogelijkheid te bladeren door scans en instructies, om op die manier vertrouwd te raken met het project, en te beslissen of ze mee willen werken. Een inlogmodule geeft vrijwilligers toegang tot de applicatie en houdt voor de organisatie automatisch de voortgang van het project bij. De forummodule zorgt voor de onderlinge communicatie.

Een centrale rol speelt de bewerkingsmodule; deze kan verschillende soorten werk aanbieden, zoals:

> het bewerken (bijvoorbeeld roteren) van scans of het markeren van delen van een scan;
> het toevoegen van steekwoorden (metadata) bij een scan;
> het toevoegen van een transcriptie of vertaling van de tekst op een scan of de transcriptie van een geluidsband;
> het toevoegen van annotaties bij getranscribeerde gegevens.

Tot slot biedt de correctiemodule de mogelijkheid om het werk voor correctie te laten rouleren.

Aan de achterkant biedt de applicatie een hulpmiddel voor de organisatie: hier worden de voortgang van het project en versiebeheer getoond (zie figuur 3). Van het project van de gekaapte brieven, dat relatief veel afvallers kende, heb ik geleerd dat het belangrijk is dat de applicatie na een bepaalde periode automatisch porties laat rouleren als vrijwilligers ze onaangeroerd hebben gelaten, zodat de voortgang van het werk is verzekerd.

7 KOSTEN-BATENANALYSE

In welke situaties loont crowdsourcing nu de moeite? En wat zijn de exacte kosten? Voor die vragen zijn geen precieze vuistregels te geven. Aan de kostenkant moet in ieder geval gedacht worden aan het maken van scans en het bouwen van een applicatie. De kosten voor het laatste kunnen beperkt zijn als men uitgaat van een al bestaande open source-applicatie. Dan is er natuurlijk de tijd die nodig is voor de coördinatie van het project: sommige organisaties schatten die erg hoog in, maar als het project goed is doordacht en opgezet, blijkt de benodigde tijd erg mee te vallen; alleen in het begin is er aardig wat tijd nodig voor het opzetten van het project, het werven van vrijwilligers en het in goede banen leiden van de eerste stroom vragen. Als die fase achter de rug is, gaat een crowdsourcingproject stationair draaien en kost de coördinatie relatief weinig tijd. Een aantal van tussen de honderd en tweehonderd vrijwilligers per project is in mijn ogen ideaal: er is altijd gestage voortgang, ook als een deel van de vrijwilligers in de zomer tegelijkertijd op vakantie gaat. En het aantal vragen dat beantwoord moet worden, is goed te behappen.

Aan de batenkant staat natuurlijk dat het werk zonder vrijwilligers niet zou worden uitgevoerd. Bovendien zorgt een crowdsourcingproject ervoor dat een instelling direct contact krijgt met belangstellenden en zichtbaarder wordt in de maatschappij. De instelling zelf kan er ook nog wat van leren: crowdsourcing vergt een heel andere manier van werken, waarin bijvoorbeeld geen managementlagen zitten – een situatie waar velen van dromen. Al het werk draagt onmiddellijk bij aan het netto eindresultaat. Vrijwilligers kiezen zelf welke taken ze uitvoeren en hoeveel tijd ze daaraan besteden. Door die keuzevrijheid doen ze het werk altijd met plezier. Organisaties zijn soms bang dat de kwaliteit van het vrijwilligerswerk onvoldoende is. Dat is niet mijn ervaring. Je kunt de correctie veilig aan een groep speciaal daarvoor geselecteerde vrijwilligers overlaten: ze stellen er een eer in foutloos werk af te leveren, en anders dan normale werknemers hoeven ze geen concessies aan de kwaliteit te doen vanwege deadlines: zij nemen er alle tijd voor, maar dan krijg je ook wat.

Meer weten?

Wie na het lezen van dit stuk meer wil weten over het starten met een crowdsourcingproject, stuurt een berichtje naar post@nicolinevdsijs.nl. De applicaties die op het Meertens Instituut zijn ontwikkeld, zijn allemaal open source en kunnen – tegen kostprijs – voor anderen worden aangepast. Voorbeelden van die applicaties zijn te vinden op www.meertens.knaw.nl/kranten_editor/ en op www.meertens.knaw.nl/vragenlijsten/sessie.

Nicoline van der Sijs is hoogleraar Historische taalkunde van het Nederlands aan de Radboud Universiteit Nijmegen en senioronderzoeker bij het Meertens Instituut.