Crowdsourcing

Voor het verzamelen van gegevens is mankracht nodig. Veel mankracht. James Murray, die vanaf 1879 hoofdredacteur was van het grootste Engelse woordenboek, de *Oxford English Dictionary*, schakelde een legertje vrijwilligers in om citaten uit de Engelse literatuur te verzamelen. Die citaten schreven ze, met bijbehorend trefwoord en bronvermelding, op losse fiches. Iedere dag leverde de postbode pakjes met fiches af. Murray verdeelde de pakjes onder zijn elf kinderen, die met het alfabetisch sorteren hun zakgeld verdienden. Ze hielden er een enorme woordenschat en prachtige thuistaal aan over: favoriet onder de kinderen was de bespotting *you dirty toe-rag*. *Toe-rag* (letterlijk 'teenlap, voetlap') is een ouderwetse benaming voor een schooier.

Mede dankzij Murrays kinderen verliep het werk aan de *Oxford English Dictionary* voorspoedig: de eerste aflevering verscheen in 1884, de laatste in 1928. Vergelijk dat eens met ons eigen *Woordenboek der Nederlandsche taal*, waarvan de publicatie maar liefst 135 jaar duurde (1863-1998). Maar ja, hoofdredacteur Matthias de Vries had slechts vier kinderen en zijn collega L.A. te Winkel is nooit getrouwd.

Tegenwoordig is kinderarbeid verboden, maar vrijwilligers inschakelen bij wetenschappelijke projecten gebeurt nog steeds. Sterker nog, dat neemt de laatste jaren een hoge vlucht dankzij de jongste technologische ontwikkelingen. Via internet kunnen mensen gezamenlijk aan een project werken. Op die manier komt sinds 2001 de internetencyclopedie Wikipedia tot stand. Daarin wordt de kennis van vele duizenden mensen samengebracht: iedereen die iets weet over een onderwerp, kan dat toevoegen. Zo wordt voor het eerst een naslagwerk door vrijwilligers samengesteld zonder de strakke leiding van een hoofdredacteur als Murray.

Onderzoekers, die altijd hongerig zijn naar data en chronisch last hebben van tijdgebrek, hebben op dit idee voortgeborduurd. Voor veel wetenschappelijk onderzoek moeten gegevens worden verzameld of gerubriceerd, voordat ze kunnen worden geanalyseerd. Dit verzamelen en rubriceren kost veel tijd. Als het werk goed wordt gedefinieerd en geleid, kunnen niet-specialisten erbij helpen. In het precomputertijdperk echter viel de kosten-batenanalyse voor het inzetten van vrijwilligers vaak negatief uit. Zo schreef Murray dagelijks tussen de 30 en 40 brieven aan zijn vrijwilligers – met de hand, want een typemachine had hij niet. Dat kan tegenwoordig beter en sneller, met computerprogramma's waarmee vrijwilligers via internet kunnen samenwerken en gestructureerd gegevens kunnen aanleveren.

Sinds 2006 heeft deze nieuwe collectieve werkwijze een aparte naam: *crowdsourcing*. De van oorsprong Engelse term is bedacht door de Amerikaan Jeff Howe, redacteur van *Wired Magazine*. Crowdsourcing is het nieuwe *outsourcing*: activiteiten worden uitbesteed aan de *crowd*, de menigte. Een andere term, die meer de nadruk legt op de toepassing binnen de wetenschap, is *citizen science* oftewel *burgerwetenschap*.

Bètawetenschappers liepen voorop met crowdsourcingprojecten: een van de oudste is Galaxy Zoo, dat als doel heeft sterrenstelsels te classificeren. Een bekender project is de jaarlijkse tuinvogeltelling.

Inmiddels hebben ook geesteswetenschappers het idee omhelsd. Ze stellen namelijk steeds vaker kwantitatieve vragen, en ook daarvoor zijn veel, heel veel data nodig. Vragen als: hoe vaak citeert Joost van den Vondel de Bijbel, welke passages zijn bij hem favoriet en heeft zijn overgang tot het katholicisme in 1641 gevolgen voor zijn keuzes? Iemand met érg veel tijd kan deze vragen misschien beantwoorden door het volledige werk van Vondel te lezen en te turven, maar de meeste onderzoekers zullen toch naar de computer hollen en digitale teksten van Vondel gaan doorzoeken. Zeker als ze de antwoorden in een breder kader willen plaatsen en bijvoorbeeld ook willen weten of Vondel zich onderscheidt van andere – protestantse en katholieke, literaire en non-fictie – auteurs. Pas als je dat weet, krijg je immers inzicht in de veelomvattende vraag naar de invloed van de Bijbel op het denken in de 17de eeuw.

Onderzoekers stellen vaker kwantitatieve vragen als: hoe vaak citeert Joost van den Vondel de Bijbel?

Om dergelijk kwantitatief onderzoek mogelijk te maken, worden steeds meer oude boeken gescand en met optische tekenherkenning omgezet in een tekst die je kunt doorzoeken. Hoe ouder de werken echter zijn, hoe moeilijker de computer de schrifttekens herkent. Gotisch schrift en handgeschreven tekst zijn voor de computer voorlopig nog een brug te ver.

Hier kunnen vrijwilligers inspringen. Sinds 2007 heb ik zelf ervaren hoe enorm groot de bijdrage van vrijwilligers kan zijn aan projecten als het digitaliseren van oude Bijbels, handgeschreven gekaapte brieven en dialectvragenlijsten van het Meertens Instituut. Er is sprake van een win-winsituatie: de onderzoekers krijgen een enorme hoeveelheid gegevens en maatschappelijke feedback; de vrijwilligers vergroten hun horizon, doen nieuwe kennis op, leveren een zinvolle bijdrage aan de wetenschap en leren gelijkgestemden kennen.

Het grootste geesteswetenschappelijke crowdsourcingproject dat ik ken, is opgezet door de Nationale Bibliotheek van Australië onder de naam Trove. Trove biedt, net als de Koninklijke Bibliotheek bij ons, historische kranten en tijdschriften aan. Die kranten zijn gescand. Doordat krantenpapier vaak van slechte kwaliteit is, bevat de door de computer gelezen tekst veel fouten. Trove laat deze fouten corrigeren door vrijwilligers. Het project is een groot succes. Het zou toe te juichen zijn als de KB dit initiatief overneemt. In 2012 werden de historische kranten van de KB door ruim een kwart miljoen unieke bezoekers geraadpleegd, potentieel een enorme vijver aan vrijwilligers. Met hun hulp kunnen de kranten worden getransformeerd tot betrouwbaar onderzoeksmateriaal.