NICOLINE VAN DER SIJS

# *Less is more* in digitaliseringsland

Nooit eerder in de geschiedenis van de mensheid is informatie zo gedemocratiseerd: iedereen heeft via internet toegang tot enorme hoeveelheden digitale gegevens, en kan daaruit halen wat van zijn gading is. Daarvoor is wel wat handigheid nodig. In de ideale wereld zorgt het onderwijs ervoor dat scholieren leren hoe ze geschiedkundige feiten, literaire teksten en recensies kunnen vinden, wat het belang is van fact-finding, en dat niet alles wat op internet blinkt, goud is. Die ideale wereld is nog ver: de staatssecretaris van OCW heeft het advies van de KNAW om Informatica als verplicht schoolvak in te voeren, in de wind geslagen, ondanks de waarschuwing dat Nederlandse scholieren digitaal ongeletterd dreigen te worden.

Gegevens van de laatste decennia zijn hun leven digitaal begonnen - ze zijn 'digital born'. Er zit wel eens een tikfoutje in, maar ze zijn relatief 'schoon'. Er komen ook steeds meer oude teksten op internet te staan, teksten die ooit zijn gedrukt en nu worden gedigitaliseerd: gescand en automatisch gelezen met optische tekenherkenning. Die teksten zijn inhoudelijk buitengewoon interessant voor iedereen met historische belangstelling. Tot nu toe waren deze teksten moeilijk te vinden, beperkt toegankelijk en veel te omvangrijk om handmatig te doorzoeken. Denk aan de 9 miljoen krantenpagina's, 1,5 miljoen tijdschriftpagina's en 2 miljoen boekpagina's uit 1781-1800 die via Delpher, het toegangsportaal van de wetenschappelijke bibliotheken, doorzoekbaar zijn.

Delpher geeft eenvoudig toegang tot primaire bronnen over historische gebeurtenissen en personen. Dat levert interessante observaties op. In het *Algemeen Handelsblad* uit 3 juni 1888 kun je lezen dat in het jaar daarvoor in Rusland ('Finland niet meegerekend') van de werken van Poesjkin 'ruim 1 1/2 millioen exemplaren [werden] gedrukt; 667.600 exemplaren van Leo Tolstoi's werken; van Gogol 40000, van den fabeldichter Kryloff 50.000.' In *De Amsterdammer* van 11 september 1896 staat de opmerkelijk onderkoelde mededeling:

> Bij S.J. Veen verscheen van Louis Couperus, *De verzoekingen van den St. Antonius* naar Gustave Flaubert (Fragmenten). Het is zeker geen kleinigheid dit beroemde werk van Flaubert in Nederlandsche proza over te zetten. Menig bladzijde in de vertaling getuigt daar dan ook van, hoewel er ook te vinden zijn, die bewijzen dat Couperus zijn taak begrepen heeft.

Dat we nu dergelijke pareltjes kunnen vinden, is pure winst. En dat het digitaliseren enorm veel geld heeft gekost, hebben we daar graag voor over. Toch? Er is echter een keerzijde aan het massadigitaliseren: de kwaliteit van de optische tekenherkenning waarmee de teksten zijn gelezen, laat te wensen over. Zeer te wensen over. Het gevolg is dat de informatie die je vindt, onbetrouwbaar is. Als je Delpher mag geloven, dateert *radio* van 1671, *televisie* van 1853, *fietsen* van 1676, en de eerste *tweet* van 1645. *Facebook* wordt al in 1729 genoemd, en *verzuiling* in 1781. Zou het echt?

Dit kun je checken in het origineel. Voor veel nieuw onderzoek geldt dat niet. De digitalisering heeft geleid tot een nieuwe geesteswetenschappelijke discipline: e-humanities. Onderzoekers analyseren teksten met behulp van de computer. Veel onderzoek draait om de vraag hoe vaak een bepaald begrip voorkomt in een bepaalde periode ten opzichte van andere periodes (wanneer culmineert de discussie rond de evolutietheorie?), of in een bepaalde tekstsoort (wordt het debat over dekolonisatie in juridische werken, het parlement of de krant gevoerd?), of in een bepaalde regio (stamt *bajes* uit Amsterdam?). Belangrijk is dus de relatieve frequentie van een begrip. Daarover valt echter niets zinnigs te zeggen als de data onbetrouwbaar zijn en je – volkomen ontraceerbare – aantallen vals positieven en vals negatieven krijgt. Met name de vals negatieven storen enorm; de informatie die je niet kunt vinden omdat die verminkt is.

Om de slechte data te verbeteren, is en wordt er opnieuw veel geld uitgegeven: er worden – matig werkende – correctieprogramma's ontworpen, en zoekroutines die fouten omzeilen. Dat is het spreekwoordelijke paard achter de wagen spannen. In plaats van met slechte data te beginnen en die proberen op te poetsen, kun je beter direct zorgen voor goede data. Je koopt toch ook geen jurk met gaten die je vervolgens door een kleermaker voor veel geld laat oplappen? Mijn oproep aan iedereen die digitaliseert – particulieren en instellingen – is: digitaliseer kwalitatief en niet kwantitatief. *Less is more!* Scholieren en onderzoekers zullen ermee geholpen zijn.