Vakblad voor public folklore & public history

Levend Erfgoed

Jaargang 8
nummer 1 2011
www.volkscultuur.nl

Immaterieel erfgoed en nieuwe sociale media

Computationele methoden in erfgoed-onderzoek

Een nieuw perspectief op historische data

Peter van Kranenburg
onderzoeker aan het Meertens Instituut

Veel van de huidige activiteit in 'digital humanities' is gericht op het digitaliseren en het ontsluiten van erfgoedcollecties. Dit biedt niet alleen nieuwe perspectieven voor de toegankelijkheid van de collecties, maar ook voor wetenschappelijk onderzoek naar relaties en verbanden tussen de artefacten in de collecties. Omdat het hierbij om grote hoeveelheden digitale gegevens gaat, wordt het interessant om computationele onderzoeksmethoden te gebruiken om deze collecties hun geheimen te ontfutselen.

Dit artikel gaat over mogelijkheden die computationele methoden bieden voor onderzoek naar culturele artefacten, toegespitst op lopend en afgerond onderzoek dat op het Meertens Instituut plaatsvindt. Het Meertens Instituut heeft een aantal etnologische databanken in beheer die een schat aan informatie bevatten over alledaagse Nederlandse cultuur.

Tools en modellen

Allereerst wil ik graag een onderscheid maken tussen 'tools' en 'modellen', beide vanuit het perspectief van een erfgoedonderzoeker. Dit onderscheid hangt samen met de mate van interdisciplinariteit in het onderzoek. We gaan uit van de situatie dat een erfgoedonderzoeker op basis van een collectie artefacten een theorie over die artefacten wil opstellen.

Stel dat we een verzameling handgeschreven historische brieven hebben die we digitaal zouden willen doorzoeken op bepaalde trefwoorden. We willen bijvoorbeeld alle brieven vinden waarin de plaatsnaam 'Rotterdam' wordt genoemd. Dan dienen eerst alle teksten te worden gedigitaliseerd. Dat kunnen we doen door de brieven in te scannen, maar dan nog is het niet mogelijk om de tekst van de brieven te doorzoeken. Daarvoor is een computertoepassing nodig die op een betrouwbare manier handgeschreven letters en woorden kan herkennen en omzetten naar doorzoekbare tekst. Dit is een zeer ingewikkeld probleem, zeker als de handschriften in de brieven niet consistent van vorm zijn. Daar is gedegen onderzoek voor nodig. Echter, voor de erfgoedonderzoeker is de ingewikkeldheid van dit probleem niet relevant. Hij heeft slechts een hulpmiddel, een 'tool', nodig om die omzetting betrouwbaar te doen. Als deze tool zijn werk heeft gedaan, wordt hij weer 'opgeborgen'. In de meeste gevallen is zo'n tool voor de erfgoedonderzoeker een black box. Als 'het werkt' is het voldoende. Hoe de tool werkt is niet interessant, de motorkap mag dicht blijven. Je zou dit computer-ondersteund onderzoek kunnen noemen. De theorie over de brieven die de erfgoedonderzoeker uiteindelijk produceert bevat niets van de technologie die het mogelijk maakte om tot die theorie te komen. We hebben hier dus te maken met een tamelijk 'losse' koppeling tussen geesteswetenschappelijk onderzoek en informatica, waarbij de informatica voornamelijk een dienstverlenende rol heeft. Gebruikmaking hiervan is waardevol en kan tot resultaten leiden die zonder computer onmogelijk te bereiken waren. Daarom is het ontwikkelen van tools belangrijk.

We zouden met deze bijdrage van de informatica aan geesteswetenschappelijk onderzoek tevreden kunnen zijn, maar de integratie van computationele methoden in het onderzoek kan nog een stap verder. Cruciaal daarvoor is dat de kennis die uiteindelijk wordt geproduceerd, wordt geformuleerd in termen van een computationeel model dat relaties en verbanden tussen culturele artefacten verklaart en voorspelt. In dit geval is de theorie zelf van computationele aard, niet enkel de wijze waarop de theorie is verkregen. Dit soort onderzoek vereist een ander soort samenwerking tussen geesteswetenschappers en informatici. In computer-ondersteund onderzoek levert de informatica hulpmiddelen, terwijl in computationeel onderzoek de informatica de 'taal' levert waarin de resulterende kennis wordt gevat. In het laatste geval zal de samenwerking tussen de twee disciplines veel intensiever zijn. In de volgende paragrafen zullen we deze interdisciplinaire onderzoeksbenadering verder verkennen.

Voorbeelden van beide soorten onderzoek vinden we in projecten die deel uitmaken van het NWO CATCH-programma, een onderzoeksprogramma dat al een aantal jaren loopt, waaruit projecten gefinancierd worden die toegankelijkheid van erfgoedcollecties voor publiek en onderzoekers verhogen. Elk van deze projecten is een samenwerking tussen een kennisinstelling (een universiteit of onderzoeksinstituut) en een erfgoedinstelling (musea, bibliotheken, enzovoort). Deze opzet dwingt tot interdisciplinair onderzoek en biedt daarmee een uitstekende basis om computationele methoden voor erfgoedonderzoek te verkennen.
Eén van die projecten was het WITCHCRAFT[1] project (2006-2010), waarin de Universiteit Utrecht en het Meertens Instituut samenwerkten om computationele modellen van gelijkenis tussen melodieën te ontwikkelen. De basisvraag in dit project was: hoe kunnen we berekenen in hoeverre twee melodieën op elkaar lijken?

Computationeel onderzoek

Het woord 'computationeel' veronderstelt dat er gerekend wordt. Dit is inderdaad het geval. De kern van een computationele benadering is dat een rekenprocedure (een algoritme) wordt gebruikt om een bepaald probleem op te lossen.

Het totaalplaatje van computationeel onderzoek dat ik hier wil uitwerken ziet er als volgt uit. De informatica levert abstracte modellen en methoden om abstracte problemen op te lossen. Computers zijn in staat deze methoden uit te voeren. Onderzoekers kunnen hiervan gebruik maken door hun onderzoeksdata te formaliseren en hun onderzoeksvragen te formuleren in termen van zulke abstracte modellen en methoden. In het vervolg van deze paragraaf zal ik de elementen uit dit totaalplaatje van enige toelichting en van voorbeelden voorzien.

Fundamenteel onderzoek in de informatica richt zich op het vinden van abstracte oplossingen voor abstracte problemen. Een eenvoudig voorbeeld van zo'n abstract probleem is hoe je efficiënt een reeks elementen kunt sorteren. Er zijn verschillende sorteeralgoritmes ontworpen die dat met relatief weinig operaties (efficiënt) kunnen doen.[2] Voor een aantal van deze algoritmes is de enige voorwaarde dat voor elk paar van elementen bepaald kan worden of het ene element kleiner is dan het andere. Om zo'n sorteeralgoritme te gebruiken hebben we dus slechts drie dingen nodig:
- Een reeks te sorteren elementen.
- Het sorteeralgoritme, dat de sortering uitvoert en daarbij gebruikt maakt van
- een methode om voor twee willekeurige elementen te bepalen of het ene element kleiner is dan het andere.

Met behulp van de melodieënzoek-machine kunnen medewerkers van het Meertens Instituut onbekende melodieën identificeren.

Dit is een abstracte beschrijving. De elementen kunnen van alles zijn en ook de wijze om te bepalen welke van twee elementen kleiner is, kan op allerlei manieren worden ingevuld. Dat laatste is cruciaal. Dat maakt allerlei toepassingen mogelijk. Een concrete toepassing op een reeks getallen ligt voor de hand omdat de relatie 'kleiner dan' een duidelijke betekenis heeft voor twee getallen. 7 is kleiner dan 10, waardoor 7 altijd vóór 10 zal komen in een oplopend gesorteerde reeks getallen. Maar als we bijvoorbeeld een reeks mensen sorteren, wordt de vraag hoe we de vergelijkingsmethode definiëren interessanter. We kunnen bijvoorbeeld zeggen dat persoon A 'kleiner' is dan persoon B als hij een kleinere

schoenmaat heeft. Maar we kunnen ook zeggen dat persoon A 'kleiner' is dan persoon B als hij jonger is. Deze twee mogelijkheden leiden (hoogstwaarschijnlijk) tot verschillende sorteringen. Zo kunnen we door het veranderen van de definitie van de kleiner-dan-relatie het resultaat van de sortering veranderen.

In dit eenvoudige voorbeeld zien we hoe we een abstracte oplossing voor een abstract probleem op verschillende manieren kunnen inzetten voor een concreet doel. Het maakt dus niet zoveel uit wat de elementen van zo'n reeks precies zijn, zolang ze maar een kleiner-dan relatie tot elkaar kunnen hebben. Voor bepaalde soorten culturele artefacten zal het mogelijk zijn om op één of meerdere manieren een kleiner-dan-relatie te definiëren. In zo'n definitie kan allerlei kennis over die artefacten verwerkt worden.

Stel nu dat we een manier hebben om de 'afstand' tussen twee artefacten te berekenen. Samen met het sorteeralgoritme hebben we dan twee belangrijke onderdelen om een zoekmachine te maken. Dan kunnen we namelijk alle elementen sorteren volgens de afstand tot een zoekterm: element A is kleiner dan element B als de afstand van element A tot de zoekterm kleiner is dan de afstand van element B tot de zoekterm. Het resultaat van de sortering is een zogenaamde 'ranked list'. Een lijst waarbij het meest gelijkende element bovenaan staat. Hoe lager je op de lijst kijkt, des te minder lijken de elementen op de zoekterm. Een dergelijke lijst wordt bijvoorbeeld door Google geretourneerd als je een zoekvraag ingeeft en op de zoek-knop klikt.

WITCHCRAFT

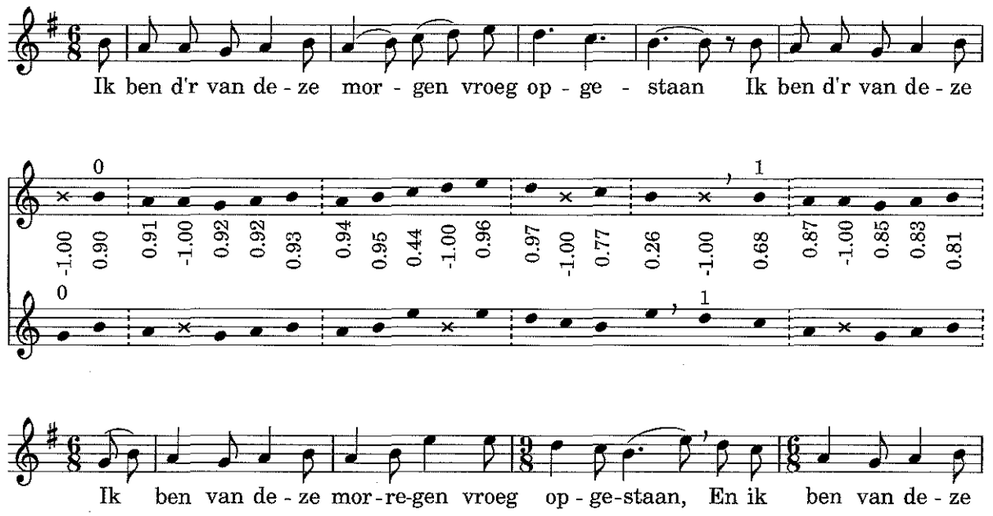
Voor de Nederlandse Liederenbank van het Meertens Instituut is binnen het WITCHCRAFT project een dergelijke zoekmachine gemaakt. De bedoeling is dat aan de zoekmachine een melodie als zoekvraag wordt gegeven, waarna de zoekmachine die melodieën vindt die het meest op de zoekvraag lijken. Hiermee kun je bijvoorbeeld andere teksten vinden die op dezelfde melodie worden gezongen, of je kunt aan de hand van de zoekresultaten een onbekende melodie identificeren. Een complicerende factor daarbij is dat de collectie veel melodieën uit de mondelinge overlevering bevat, liedjes die nergens op papier staan, maar die door mensen vanuit hun geheugen zijn gezongen en opgenomen op band. Tijdens het mondeling aanleren en het reproduceren vanuit het geheugen kunnen er allerlei veranderingen optreden. Daarom dient de zoekmachine in staat te zijn ook melodieën te vinden die niet letterlijk hetzelfde zijn als de zoekvraag, maar die wel als variant beschouwd kunnen worden.

De belangrijkste onderzoeksvraag voor het WITCHCRAFT-project was dus hoe de gelijkenis tussen twee melodieën in een getal kan worden uitgedrukt. Hiervoor is een computationeel model nodig van de gepercipieerde gelijkenis van melodieën.

Het model dat gekozen is om deze probleemstelling te benaderen is dat van de uitlijning: de mate waarin twee melodieën uitgelijnd kunnen worden bepaalt in hoeverre ze op elkaar lijken. 'Uitlijnen' betekent hier dat de melodieën zodanig onder elkaar geplaatst worden dat de overeenkomende gedeelten onder elkaar staan. Het zal dus nodig zijn om hier en daar in de melodieën wat ruimte in te voegen om het vervolg weer te laten corresponderen. Een voorbeeld van zo'n uitlijning staat in de afbeelding. Er worden twee melodiefragmenten getoond met in het midden de uitlijning van die melodiefragmenten. De melodiefragmenten komen uit varianten van hetzelfde liedje. Om de corresponderende noten van de twee melodieën onder elkaar te krijgen is op verschillende plaatsen een 'gat' (x) tussengevoegd.

Zo'n uitlijning kan 'met de hand' gemaakt worden – dat is in volksliedonderzoek dan ook veelvuldig gedaan – maar de uitlijning in de afbeelding is berekend door een uitlijningsalgoritme, een abstracte procedure die in de verzameling van alle mogelijke uitlijningen op een efficiënte wijze de optimale uitlijning vindt. De abstracte beschrijving is als volgt: gegeven twee reeksen symbolen een manier om de gelijkenis tussen twee symbolen te berekenen en het uitlijningsalgoritme, wordt de optimale uitlijning van de twee reeksen gevonden, waarbij de optimale uitlijning die uitlijning is die de hoogste totaalscore heeft. De totaalscore van een uitlijning wordt bepaald door individuele scores van elk paar met elkaar uitgelijnde symbolen bij elkaar op te tellen. Als een symbool niet met een ander symbool is uitgelijnd maar met een 'gat', geldt ook daarvoor een score. Voor de individuele scores gebruiken we de gelijkenismaat voor symbolen. In de afbeelding zijn de scores voor individuele symbolen tussen de notenbalken weergegeven.

De abstracte onderdelen die we concreet in moeten vullen zijn dus enerzijds de symbolen en anderzijds een gelijkenismaat die de score van de uitlijning van twee symbolen en van een symbool met een gat berekent. In deze concrete invulling kan allerlei kennis over muziek en mondelinge overlevering verwerkt worden. Deze concrete invulling is daarmee de cruciale stap die dit onderzoek tot 'computational humanities' maakt door domeinkennis te verbinden met de abstracte methoden van de informatica.



Twee melodie fragmenten, met in het midden de uitlijning van die melodiefragmenten. Hierdoor kun je gelijkenissen meten van verschillende melodieën.

In de uitlijning in de afbeelding zijn de noten de symbolen. We hebben dus twee reeksen van noten. De scores worden berekend met behulp van eigenschappen van de noten zoals de plaats van de noot binnen de frase, het metrisch gewicht en de toonhoogte.
Door de berekening van deze scores te variëren, kan het effect van allerlei muzikale parameters op de melodische gelijkenis bestudeerd worden. We zouden bijvoorbeeld een score kunnen berekenen die enkel gebaseerd is op de tijdsduur van de noten. Dan krijgen we een gelijkenismaat voor ritmes.

Deze methode is in de Nederlandse Liederenbank geïmplementeerd in een melodieënzoekmachine. Met behulp van deze zoekmachine kunnen medewerkers van het Meertens Instituut onbekende melodieën identificeren. Zo kan gevonden worden welke andere teksten op de melodie van een bepaald liedje worden gezongen. Als bijvoorbeeld wordt gezocht met de melodie van *Daar was laatst een meisje loos* vinden we ook *Daar was laatst een oude soldaat, Elf november is de dag* en *Daar was laatst een turrefboer*.

Zowel binnen als buiten het domein van de muziek zijn vele andere toepassingen van het uitlijningsalgoritme mogelijk. Als de symbolen letters zijn, kunnen we woorden met elkaar uitlijnen. Dit is precies wat gebeurt in de spellingscontrole van moderne tekstverwerkers. Als de symbolen woorden zijn, kunnen we varianten van een bepaalde tekst met elkaar uitlijnen. In de biologie wordt deze methode gebruikt om corresponderende gedeelten van DNA reeksen te vinden. We kunnen in principe alles met alles uitlijnen zolang we het maar kunnen representeren als een reeks symbolen en we voor die symbolen een gelijkenismaat kunnen definiëren.

Beproeving

Wanneer we eenmaal een model hebben ontwikkeld, is het belangrijk om dat model te beproeven. We willen tenslotte weten hoe goed het model is. Dit is meestal geen eenvoudige opgave.
In de exacte wetenschappen wordt een model beproefd door het te confronteren met meetresultaten, met empirische observaties die onomstotelijk vast staan. Er wordt dan een model gezocht dat die meetresultaten op een zo elegant mogelijke manier beschrijft. Het zou daarom erg fijn zijn als we voor ons geesteswetenschappelijk probleem voor een aantal gevallen de juiste uitkomst al zouden weten. Dan kunnen we voor die gevallen de uitkomsten van het algoritme vergelijken met wat eruit zou moeten komen. Hoe beter het algoritme presteert op die bekende gevallen, des te betrouwbaarder zijn de uitkomsten voor gevallen waarvoor we nog niet de juiste uitkomst weten. Dit is een werkwijze die vaak gekozen wordt door informatici die geen specialistische kennis hebben over het domein waarvoor ze algoritmes ontwerpen. De verzameling juiste uitkomsten wordt een 'gouden standaard' of 'ground-truth' genoemd. De kwaliteit van het algoritme wordt dan uitgedrukt in het percentrage van deze 'ground-

truth' waarvoor het algoritme het correcte antwoord geeft. In de praktijk blijkt dat een nauwkeurigheid van meer dan 80% als succesvol gezien wordt.
Uiteraard is het in veel gevallen zeer problematisch zo'n gouden standaard samen te stellen. Zeker in het geesteswetenschappelijk domein geldt dat er weinig onomstotelijke kennis is. Een voorbeeld waar deze benadering denkbaar is, is auteurschapsonderzoek. Als we een tekst hebben waarvan de auteur onbekend is, maar we hebben wel twee serieuze kandidaten, auteurs A en B, dan kunnen we zoveel mogelijk teksten van auteurs A en B verzamelen (onze gouden standaard), een model ontwikkelen dat zoveel mogelijk van die bekende teksten correct herkent en vervolgens dat model toepassen op de onbekende tekst. Maar meestal is de waarheid minder eenduidig dan ze lijkt. Complicaties bij auteurschapsonderzoek zijn bijvoorbeeld dat verschillende personen aan een tekst gewerkt hebben (de auteur, een redacteur, een ghostwriter, etc.), terwijl er toch maar één naam boven de tekst staat, of dat auteurs verschillende stijlen ontwikkelen voor verschillende genres of elkaar imiteren, etc. Een zorgvuldige deconstructie van het begrip 'auteurschap' laat zien dat dit begrip verre van eenduidig is.[3]
Een ander probleem van de gouden standaard is dat alle domeinspecifieke vragen achter de standaard verdwijnen. Er wordt verondersteld dat definitieve antwoorden beschikbaar zijn, terwijl er in de geesteswetenschappelijke praktijk over vrijwel alles discussie bestaat. Precies dat wat interessant is vanuit geesteswetenschappelijk perspectief wordt dus 'weggemoffeld' waardoor de gouden standaard een soort hermetische scheiding tussen de disciplines wordt en een zinvolle integratie in de weg staat.

Als we het idee van de gouden standaard opgeven, lijken we vanuit empirisch perspectief de grond onder onze voeten te verliezen. We hebben immers geen 'observaties' meer om ons model aan te toetsen. Ik heb in dit verband ooit iemand de term 'moeras' horen gebruiken. Toch zou ik hiervoor willen pleiten, want juist hierdoor kunnen uitkomsten van algoritmes betekenis krijgen in het geesteswetenschappelijk domein. De vraag is dan niet meer in hoeverre de gouden standaard door een algoritme *gereproduceerd* kan worden, maar wat de uitkomsten van een algoritme zeggen over het geesteswetenschappelijk probleem. Dit ontneemt voor geesteswetenschappers ook het bedreigende karakter van computationele methoden. Het idee dat de computer ons wel even zal vertellen hoe het zit wordt hiermee ontkracht en de computationele methode wordt één beschikbare methode naast andere om een bepaalde probleemstelling te benaderen.

Het laatste woord is hierover ongetwijfeld nog niet gesproken. Er wordt zeker nagedacht over de methodologische consequenties die computationele benaderingen in de geesteswetenschappen hebben, maar de verkenning van dit interdisciplinaire onderzoeksgebied is nog maar net begonnen.

Andere formalisaties en methoden

Terug naar de computationele methoden en modellen. We hebben een voorbeeld gezien van een datarepresentatie (een reeks symbolen) en we hebben twee voorbeelden gezien van algoritmes (een sorteeralgoritme en een uitlijningsalgoritme), maar er zijn uiteraard talloze andere voorbeelden. Andere representaties zijn bijvoorbeeld vectoren in een ruimte, grafen, bomen, weighted point sets, enovoort. Er zijn allerlei methoden die op zulke formele representaties kunnen worden losgelaten. De ruimte ontbreekt hier om voorbeelden uit te werken. Maar wat wel duidelijk zal zijn is dat de keuze van een bepaalde formalisatie en een bepaalde methode bepaald wordt door zowel kennis van die formalisaties en methoden als kennis van het geesteswetenschappelijk domein en het geesteswetenschappelijk discours betreffende een bepaalde onderzoeksvraag. De creativiteit in dit soort onderzoek bestaat hierin dat een passend model wordt ontworpen voor een bepaalde onderzoeksvraag. Hoe beter een concreet probleem inpasbaar is in de gebruikte abstracte methode, des te waardevoller de resultaten voor het onderzoek zullen zijn.

De computer

Waar is nu de computer in dit geheel? Uit de manier waarop er over computationele onderzoeksmethoden gesproken wordt lijkt het soms of 'de computer' centraal staat en alles doet: de computer denkt en beslist, en wij hebben dat maar te accepteren.
Wellicht is het een verrassend inzicht dat computationeel onderzoek in principe zonder computer gedaan kan worden, zij het dat het in de meeste gevallen zeer lang zal duren totdat het eindantwoord bereikt is. De computer is puur een uitvoerende instantie. Wel geldt dat soms de uitkomsten van algoritmes onnavolgbaar zijn omdat de berekeningen die eraan ten grondslag liggen onmogelijk geheel overzien kunnen worden. Het is aan de onderzoeker om te bepalen of dat voor de betreffende onderzoeksvraag wenselijk is of niet.

De (weerbarstige) praktijk

Wanneer men computationeel onderzoek wil doen in een erfgoedinstituut brengt dat allerlei (wellicht onvoorziene) praktische problemen met zich mee. De communicatie tussen informatici en geesteswetenschappers kan zeer moeizaam verlopen. Ik heb hoog oplopende discussies meegemaakt waarin een programmeur een bepaald concept ondubbelzinnig gedefinieerd wilde hebben, zodat hij het kon implementeren in een computerprogramma, terwijl de musicoloog die geïnteresseerd was in dat concept dat niet leek te willen doen. Het betreffende concept was 'gesture'. De musicoloog leverde steeds een andere omschrijving, terwijl de door de programmeur gewenste duidelijkheid uitbleef. Vanuit beide perspectieven werd een zinvolle bijdrage geleverd, maar toch kwam een vruchtbare samenwerking niet tot stand.
Een andere praktische kwestie is dat algoritmes en datastructuren geïmplementeerd moeten worden in computersystemen om ze daadwerkelijk in werking te zetten. Dit vereist deskundigheid. Computers zijn gecompliceerde machines. Er zijn dus programmeurs nodig. Bovendien is een goede infrastructuur onontbeerlijk: data-opslag en beheer blijken in de praktijk zeer bewerkelijk te zijn. Dat moet op een doordachte manier gebeuren, zeker als de hoeveelheid data groeit.

Toepasbaarheid

Juist daar waar veel gegevens voorhanden zijn en waar onderzoeksvragen een duidelijke kwantitatieve component hebben, zijn computationele methoden een goede keuze. In het kader van het Meertens instituut zijn het de etnologische databanken die zich lenen voor dergelijk onderzoek in het erfgoeddomein. De liederenbank is al genoemd, maar ook de verhalenbank bevat een enorme hoeveelheid gegevens waar patronen in ontdekt kunnen worden.
Voor het automatisch classificeren van volksverhalen en voor het herkennen van varianten is een computationeel model nodig van de inhoud van een verhaal. Een andere databank die door het Meertens Instituut wordt beheerd is de boedelbank. Hierin zijn duizenden inventarissen van inboedels opgenomen. Met behulp van computationele methoden kunnen bijvoorbeeld innovatietrends en ontwikkelingen in de tijd zichtbaar gemaakt worden (de gegevens omspannen enkele eeuwen), maar ook een automatische inventarisatie van voorwerpen die altijd samen voorkomen – of juist niet – behoort tot de mogelijkheden. We kunnen zelfs onderzoeken of het mogelijk is om een basisgrammatica van het interieur uit de data af te leiden.
Uiteraard zijn er ook buiten de muren van het Meertens Instituut vele zinvolle toepassingen. Reeds genoemd is het NWO CATCH-programma. Hierin vinden we bijvoorbeeld projecten waarin aan automatische classificatie van archeologische voorwerpen is gewerkt (RICH), of waarin radio-archief doorzoekbaar is gemaakt (CHORAL), of waarin gewerkt wordt aan computationele modellen van historische gebeurtenissen en hun samenhang (AGORA).
Voor al dit soort onderzoeksvragen geldt dat ze onmogelijk 'met de hand' zijn te benaderen vanwege de enorme hoeveelheid gegevens en de enorme hoeveelheid verwerkingsstappen. Zonder computer komt men vaak niet verder dan 'proof by example', waarbij niet alle beschikbare data worden gebruikt om theorieën te onderbouwen.

De moeilijkheidsgraad van de techniek compliceert de samenwerking tussen informatici en geesteswetenschappers.

Tot slot

Niettegenstaande de praktische hobbels die overwonnen dienen te worden, wil ik hier benadrukken dat computationele methoden een waardevolle toevoeging zijn aan het arsenaal van onderzoeksmethoden dat beschikbaar is voor de geesteswetenschapper; zeker als het gaat om onderzoeksvragen die duidelijk kwantificeerbare aspecten hebben. De mogelijkheid die de computer biedt om in enorm tempo een enorme hoeveelheid gegevens te gebruiken geeft empirische basis aan onderzoeksresultaten en stelt in staat patronen zichtbaar te maken die anders verborgen zouden blijven. Bovendien kunnen computationele modellen een geheel nieuw perspectief op bestaande onderzoeksvragen toevoegen. De geesteswetenschappen zouden zich tekort doen door die mogelijkheden onbenut te laten. ■

Met dank aan Louis Grijp (Meertens Instituut) en Frans Wiering (Universiteit Utrecht) voor kritische lezing en suggesties.

Noten

1 Het acroniem staat voor: What Is Topical In Cultural Heritage: Content-Based Retrieval Among Folk-song Tunes.
2 Zie bijvoorbeeld T.H. Cormen (redactie), *Introduction to Algorithms* (Cambridge, Massachusetts 2002).
3 H. Love, *Attributing Authorship: An Introduction* (Cambridge 2002).