CLARIAH: BIG DATA, GRAND CHALLENGES

Geesteswetenschappers krijgen van NWO 12 miljoen Euro subsidie om een digitale infrastructuur te bouwen. Daarmee kunnen ze grote data bestanden van teksten, audiovisueel materiaal en archiefgegevens interpreteren en ontsluiten. Maar welke nieuwe onderzoeksvragen worden hiermee mogelijk?

Geesteswetenschappers onderzoeken sinds jaar en dag de menselijke cultuur. Ze stellen fundamentele vragen als: waarom zijn sommige regio's in de wereld al zo lang rijk en andere arm? Hoe komt het dat in publieke debatten hardnekkige beeldvorming over bepaalde minderheden blijft bestaan? Hoe verandert taal onder invloed van migratie? Die vragen worden al decennia lang gesteld en beantwoord door historici, media-wetenschappers, taalkundigen, en vele andere onderzoekers. Wetenschappers uit de geesteswetenschappelijke disciplines zijn heel goed in het interpreteren van *inhoud*, en dan vooral van a*fzonderlijke data* ieder *op hun eigen terrein.* Historici werken (naast teksten) met gestructureerde gegevens uit archieven. Media-experts gebruiken tekst, beeld en geluid uit kranten, radio, televisie en andere (sociale) media als bronmateriaal. En taalkundigen putten uit grote geschreven en gesproken tekstbestanden.

Alfa-wetenschappers, zou je kunnen zeggen, bestuderen ieder op eigen wijze *bouwstenen* van cultuur en *patronen* van cultuurverandering. Die bouwstenen waarmee ze van oudsher werken (tekst, beeld, geluid, en historische data) waren talrijk en versnipperd. Daarom leggen veel geesteswetenschappers zich meestal toe op één puzzelstukje om dat zo goed mogelijk te interpreteren en analyseren. Bijvoorbeeld, het werk van één schilder, de romans van één schrijver, de cijfers uit gemeentearchieven in één historische periode of het taalgebruik van één sociale groep. Waar de humaniora tot nu toe nog minder in slaagden, was om deze talrijke vragen aan elkaar te knopen en ze in samenhang te zien. Een grote vraag als: Hoe leiden verschillen tussen Europese culturen tot processen van in- en uitsluiting? Zo'n immens vraagstuk vereist een samenhangend inzicht in veranderingen in taal, sociaal-economische positie, en beeldvorming van specifieke groepen. Niet als afzonderlijke fenomenen, maar als complex geheel.

Big Data

Die grote uitdaging kan nu worden opgepakt. Sinds een kleine tien jaar beschikken onze erfgoedinstellingen (archieven, bibliotheken, kenniscentra) over steeds meer digitale data bestanden. De omvang is zodanig toegenomen dat we over Big Data zijn gaan spreken. Maar om deze rijkdom aan materiaal te ontginnen, moeten nieuwe instrumenten ontwikkeld worden: instrumenten om de data te bevragen op betekenisvolle inhoud. Daarmee veranderen niet alleen de werkwijzen van geesteswetenschappers, maar kunnen ze vragen op nieuwe manieren stellen en hopelijk beantwoorden. Om al die data te ontginnen en de interpretaties in samenhang te zien, hebben onderzoekers het plan opgevat een gezamenlijke infrastructuur te ontwikkelen. De puzzelstukjes op het gebied van tekst, afbeeldingen, bewegend beeld, geluid, en

historische gegevens moeten in elkaar kunnen worden geschoven en dus moeten experts van elkaar leren hoe ze deze data-bestanden bij hun onderzoek kunnen inzetten.

CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities) is een gezamenlijk project van een kernteam van wetenschappers, gesteund door een consortium van 40 kennis- en erfgoedinstellingen, publieke organisaties en bedrijven dat door NWO beloond is met een forse subsidie van 12 miljoen euro. Met dit geld voor een gezamenlijke infrastructuur kunnen geesteswetenschappers niet alleen digitale instrumenten ontwikkelen om grote databestanden te ontginnen; door deze instrumenten met elkaar te laten "praten", leren geesteswetenschappers ook samenwerken om die complexe vragen te beantwoorden. Drie deelgebieden vervullen een voortrekkersrol in CLARIAH: taalkunde, mediastudies en sociaal-economische geschiedenis. Taalkundigen richten zich met name op het ontginnen van digitale tekstbestanden. Mediastudies experts ontwikkelen vooral tools voor het interpreteren van audiovisuele bronnen (beeld en geluid). En sociaal-economische historici concentreren zich op gestructureerde databestanden uit archieven. Het is echter nadrukkelijk de bedoeling dat de te ontwikkelen tools bruikbaar zijn voor alle onderzoekers binnen de geesteswetenschappen die met verschillende typen digitale data werken. Uit een recente enquête in de Groene Amsterdammer onder 200 geesteswetenschappers-ers in Nederland bleek dat men *Digital Humanities*, zoals deze omslag ook wel genoemd wordt, de belangrijkste ontwikkeling in hun wetenschapsgebied vond.

Behalve dat *Digital Humanities* de alfa-wetenschappen vooruit helpen, beoogt CLARIAH ook iets heel essentieels bij te dragen aan de grote wetenschappelijke vragen rondom Big Data *buiten* de humaniora. Dit project levert namelijk ook bouwstenen die complementair zijn aan het werk van bèta's en gamma's op het terrein van data-ontginning. Waar informatici heel goed zijn in het ontwerpen van zoekalgoritmes, en sociale wetenschappers alles willen weten over het gedrag van gebruikers, is de kracht van alfa's het *interpreteren van menselijke boodschappen*. Big Data in de geesteswetenschappen zijn vooral "rich data": ze zitten vol ruis, net als cultuur vol ruis zit. Cijfers over armoede zijn niet loutere feiten, maar vragen om duiding. Meningen in een publiek debat zijn talrijk maar ook diffuus—ze hebben een verschillend soortelijk gewicht. En beelden of teksten kunnen ironisch zijn of ambigu. Wie cultuur bestudeert, weet dat inhoud interpretatie behoeft en dat boodschappen pas in hun samenhang betekenis krijgen. Die *complexiteit van content* begrijpen—dat is de bijdrage van geesteswetenschappers aan het onderzoek naar grote hoeveelheden digitale data. CLARIAH betekent dan ook een nog intensievere samenwerking tussen alfa, bèta (vooral informatica) en gamma waar het gaat om het begrijpen van culturele complexiteit.

Grote uitdagingen

Maar niet alleen wetenschappers hebben iets aan de toekomstige infrastructuur. Ook het bedrijfsleven toont interesse in de projecten die met behulp van CLARIAH uitgebouwd gaan worden. Misschien hebben bedrijven niet direct belang bij het oplossen van de Grote Vragen die de geesteswetenschappers stellen, zoals het onderzoek naar migratie en minderheden. Ze tonen echter wel interesse voor het *type*

kennis dat hier wordt ontwikkeld: het interpreteren van complexe boodschappen in digitale omgevingen. In de aanloop naar deze infrastructuur hebben onderzoekers al gewerkt met middelgrote en kleine bedrijven op het gebied van taal- en spraaktechnologie en beeldherkenning. Ook een groot bedrijf als IBM toont zich geïnteresseerd in de bijdrage van geesteswetenschappers. Besprekingen tussen IBM, de KNAW en de Amsterdamse universiteiten over samenwerking in Digital Humanities zijn in een ver gevorderd stadium. En deze vormen van publiek-private samenwerking sluiten heel goed aan bij de Topsector Creatieve Industrie, waar de Geesteswetenschappen in de eerste ronde van aanvragen zeer goed gescoord hebben. Kortom: deze infrastructuur geeft een enorme impuls aan de samenwerking tussen onderzoekers en het bedrijfsleven.

Het is vrij uniek in de geschiedenis dat Nederlandse geesteswetenschappers zich zo unaniem achter CLARIAH hebben geschaard. In de afgelopen vijf jaar hebben onderzoekers uit afzonderlijke disciplines—de taalwetenschap, sociaal-economische geschiedenis en mediastudies voorop—grote vorderingen gemaakt op hun deelgebied, bijvoorbeeld met projecten als CLIO-Infra, HSN, NederLab en EU Screen. Door deze investeringen heeft Nederland een buitengewoon sterke positie verworven in Europa. CLARIAH verbindt niet alleen de Nederlandse deelprojecten met elkaar, maar garandeert ook een gecoördineerde bijdrage aan de meer omvattende Europese infrastructuren DARIAH en CLARIN— infrastructuren waarin Nederland, ondersteund door OCW en NWO, een hoofdrol speelt.

Op korte termijn gaat het vooral om de ontwikkeling van instrumenten en het opleiden van onderzoekers in de voor velen nieuwe wereld van grote databestanden. Vervolgens stellen de verworven inzichten en vaardigheden onderzoekers in staat om grote vragen over cultuur en cultuurverandering aan te pakken. De vraag naar de Europese identiteit is niet langer “te groot” om te onderzoeken. Integendeel, door de rijkdom aan digitale data en door ze in samenhang te interpreteren komen antwoorden in zicht. Het eerder genoemde voorbeeld van in- en uitsluiting in Europa is ook niet willekeurig. Eén van de Grand Challenges die de Europese Commissie heeft geformuleerd voor de wetenschapsagenda *Horizon 2020* is “Europe in a changing world: Inclusive, innovative, and reflective societies.” Met ondersteuning van de CLARIAH infrastructuur staan de Nederlandse Geesteswetenschappers samen in de startblokken om deze grote uitdaging aan te nemen, en dat is een unicum in de geschiedenis.

Grote en kleine ambities

CLARIAH zelf is niet het antwoord op het grote vraagstuk over de complexe menselijke cultuur, net zomin als Big Data het antwoord vormen op allerlei maatschappelijke uitdagingen. CLARIAH is voor de geesteswetenschappen een onmisbare infrastructuur. Zoals de Hubble telescoop noodzakelijk is om de raadsels van de kosmos op te lossen, en de MRI-scanner om de menselijke textuur te ontrafelen, zo is CLARIAH een hulpmiddel om de complexe betekenis van de menselijke cultuur te onderzoeken. En terwijl geesteswetenschappers deze ambitieuze doelen nastreven, hopen ze meer down-to-earth dat hun zoektocht naar de samenhang tussen tekst, beeld, geluid, en andere data heel veel bruikbare instrumenten gaat opleveren.

http://www.clariah.nl

Prof. dr. José van Dijck
Professor of Comparative Mediastudies
University of Amsterdam
Department of Mediastudies
Turfdraagsterpad 9
1012 VT Amsterdam
The Netherlands
Phone: 31 20 5252980
E-mail: j.van.dijck@uva.nl
http://home.medewerker.uva.nl/j.f.t.m.vandijck/