Toine Pieters doet samen met collega's in Utrecht grootschalig transnationaal onderzoek. Big data maken dat mogelijk. Maar er doemen nieuwe grenzen op: gaten in collecties, slecht leesbare teksten. Hoe gaat de wetenschap daarmee om? *Inge Angevaare*

Toine Pieters in zijn werkkamer *foto Inge Angevaare*

Toine Pieters over grootschalig transnationaal onderzoek:

'We moeten nog leren werken met big data'

"Wij bestuderen referentieculturen, dat zijn culturen die over nationale grenzen heen een voorbeeldfunctie hebben. In de zeventiende en achttiende eeuw fungeerde de Republiek der Nederlanden als een belangrijke referentiecultuur voor Europa en de Verenigde Staten. In de twintigste eeuw groeiden de VS zelf uit tot een invloedrijke voorbeeldcultuur. Wat ons interesseert, is hoe wij onze eigen (Nederlandse of Europese) identiteit ontwikkelen met verwijzing naar voorbeeldculturen. De grote databanken met gedigitaliseerde kranten zijn een onmisbare bron voor dit onderzoek. Kranten kunnen beschouwd worden als een klankbord voor wat wij 'de publieke sfeer' noemen. Niet alleen artikelen, maar ook ingezonden brieven, cartoons en advertenties leveren ons belangrijke informatie over de wisselwerking tussen wetenschap, cultuur en economie."

"In het Europese HERA-project *AsymEnc* (*Asymmetrical Encounters*) zoeken we naar de historische dynamiek van culturele concepten zoals 'metropolis' in grote digitale krantenbanken in Nederland, Duitsland en Engeland. Voor het eerst proberen we systematisch een grootschalig longitudinaal onderzoek te doen in diverse landen. Zonder big data en text mining zou dat bijzonder lastig zijn."

Big data, new questions

"Maar we lopen ook tegen grote nieuwe vragen aan. Over hoe we dit materiaal wetenschappelijk kunnen en mogen gebruiken. Want er zitten grote gaten in de digitale krantencollecties. Wat hebben die voor effect op onze onderzoeksresultaten? Ook de uitgelezen teksten (OCR) zijn vaak van slechte kwaliteit. Kunnen de bibliotheken daar iets aan doen of moeten we daarmee leren leven? En hoe dan?"

"Het gaat om 'digital empowerment' in de geestes-wetenschappen"

"Afgelopen april organiseerde ik met collega Jaap Verheul de workshop '*Mining digital repositories: challenges and horizons*' in de Koninklijke Bibliotheek. Daar zijn die vragen uitgebreid aan de orde gekomen. Tussen onderzoekers onderling, maar ook in een dialoog met de bibliotheken en commerciële partijen die de krantencollecties beheren en ontwikkelen."

"Tijdens die workshop hebben we natuurlijk niet alle vragen kunnen beantwoorden. Maar wat mij betreft was het een uitstekend begin van de dialoog die moet leiden tot een nieuwe wetenschappelijke cultuur rond big data. En ook tot goede afspraken met bibliotheken. Zo willen wij als onderzoekers méér informatie over wat er in de databanken zit en wat niet, en over de wisselende kwaliteit van de OCR. Zodat we daar rekening mee kunnen houden bij ons onderzoek."

Crowdsourcing belangrijk

"Bibliotheken zijn openbare instellingen met beperkte budgetten. Dat is natuurlijk een gegeven. En technisch verandert de situatie voortdurend. De machines waarmee nu OCR wordt gemaakt, zijn veel beter dan een jaar of tien geleden. Maar dat oudere OCR-spul zit nog steeds in de databanken. Ik denk dat er maar één betaalbare manier is om die situatie te verbeteren: *crowdsourcing*. Onderzoekers, studenten en amateurhistorici toestaan om de uitgelezen teksten te verbeteren. Bibliotheken hebben daar niet altijd open voor gestaan. Maar nu lijkt het tij voorzichtig te keren. Ik bespeurde tijdens de workshop bij de KB althans een positieve grondhouding."

"En dan natuurlijk het auteursrecht. Er is zoveel materiaal waar we als onderzoekers nog niet bij kunnen, dat is doodzonde. Het krantenarchief van de British Library, bijvoorbeeld, is helemaal in commerciële handen. Persoonlijk vind ik dat je individuele onderzoekers niet lastig moet vallen met al die auteursrechtenkwesties. Dat bibliotheken het op zich moeten nemen om toegang voor wetenschappelijk onderzoek te regelen met de rechthebbenden.

Liefst in Europees verband. En als het moet onder strikte voorwaarden dat het materiaal niet verder verspreid wordt. Daar staan we best open voor."

INTERVIEW

"Hoe ziet uw digitale utopia eruit?"

Omgaan met taalkwesties

"Taalkwesties zijn een andere factor in transnationaal onderzoek waarmee we moeten leren werken. Je kunt concepten niet zomaar vertalen en er dan van uitgaan dat die vertaalde termen dezelfde lading hebben als het origineel. Woordbetekenissen veranderen ook voortdurend. Wat op het ene moment een neutraal woord is, kan tien jaar later een positieve of een negatieve bijklank hebben. We moeten tools ontwikkelen om die finesses uit de databestanden te kunnen halen. Tijdens onze workshop vroeg iemand of je kunt analyseren waarover niet gesproken wordt in kranten. Tja, dat gaat nog weer een stap verder."

"We vroegen tijdens de workshop ook aan collega's hoe hun 'digitale utopia' eruit zou zien. De een wilde vooral de OCR verbeteren, de ander wilde méér materiaal. Het mooiste vergezicht kwam denk ik van mijn Utrechtse '*digital history*' collega Joris van Eijnatten. Hij sprak van een 'Globiana 5.0': een wereldwijd, transnationaal digitaal archief vol materiaal in standaardformaten en gekoppeld aan tweetalige en meertalige woordenboeken. Die moeten we dan wel kunnen onderzoeken met een netwerk van wetenschappers die goed kunnen omgaan met big data. Collega Rens Bod van het Amsterdamse Centre for Digital Humanities benadrukte dat hier nog veel werk te verrichten is. Veel onderzoekers missen nog big data-vaardigheden, en dat zijn lang niet alleen de ouderen onder ons. Het gaat hier letterlijk om *capacity building* met als doel '*digital empowerment*' in de geesteswetenschappen."

Datalabs

"Onderdeel van die '*digital empowerment*' is ook dat bibliotheken een andere rol nemen ten opzichte van de data die ze beheren. De British Library heeft daarin het voortouw genomen en de Koninklijke Bibliotheek volgt binnenkort. Ze stellen zich niet langer op als passieve leverancier van hun digitale data, maar zoeken actief de samenwerking met wetenschappers om méér met die data te kunnen doen. In zogenaamde 'datalabs' werken wetenschappers samen met data-experts van de bibliotheken. Met voornamelijk opensourcetools worden de mogelijkheden van big data verkend. Van die samenwerking worden beide partijen rijker."

tinyurl.com/ketqa9m

Toine Pieters
Toine Pieters is verbonden aan het Descartes Centrum voor Wetenschapsgeschiedenis en -filosofie van de Universiteit Utrecht. In april maakte hij de overstap van de afdeling Metamedica van het VU Medisch Centrum naar een aanstelling als hoogleraar/directeur van het Freudenthal Institute for Science and Mathematics Education.
Hij is projectcoördinator van Translantis, ASYMENC en projectleider van het digitale erfgoedproject Timecapsule.