COLUMN

Tools: ik wil wel, maar waar moet ik precies beginnen?

In het vorige nummer van E-data schreef ik dat men in de e-humanities door de bomen het bos niet meer ziet. Te veel tools, te weinig samenwerking.

Er wordt wel degelijk geprobeerd om een geesteswetenschappelijke e-infrastructuur aan te leggen, luidt de samenvatting van een ingezonden brief in het huidige nummer. Zie de initiatieven en plannen voor de komende jaren van Virtual Language Observatory, CLAPOP, Bamboo Dirt en DASISH.

Vanzelfsprekend juich ik dergelijke initiatieven van harte toe, maar je kunt vier initiatieven om duidelijkheid te scheppen ook zien als meer van hetzelfde. Voor de gemiddelde gebruiker zal het er niet snel helderder op worden.

Onlangs gaf ik een gastcollege aan een paar van die gemiddelde gebruikers. In dit geval ging het om masterstudenten boekwetenschap. De vraag van de docent was of ik wilde laten zien hoe ik bepaalde tools gebruikte.

Ik liet onder meer zien hoe je met indexeringssoftware snel door grote datacollecties kunt zoeken – een van de weinige technieken die ik zelf redelijk beheers. Wildcards, jokertekens, woorden of namen in elkaars nabijheid zoeken, ik heb er dagelijks veel baat bij.

Vervolgens demonstreerde ik hoe je in Evernote heel makkelijk data kunt ordenen en van metadata kunt voorzien.

Omdat dit alles bij veel masterstudenten tot glazige blikken leidde, stapte ik over op technieken om snel relevante bronnen op internet te vinden. Zoals via Google zoeken binnen een site, gebruikmaken van de url, zoeken naar bepaalde bestandstypen en filtertechnieken. 'Had ik dat maar eerder in m'n studie geleerd, dat had me echt enorm veel tijd gescheeld', verzuchtte een studente. Anderen beaamden dit.

Voor de duidelijkheid: voor zo-

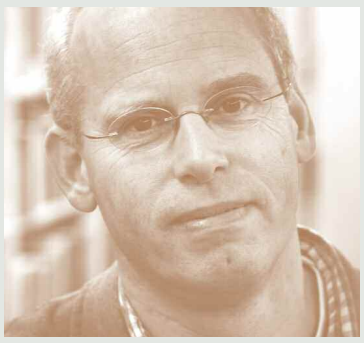


foto Leo van Velzen

ver ik dat kon beoordelen waren het allemaal leergierige en slimme mensen. Maar tegelijk was duidelijk dat ze niet erg vertrouwd waren met wat in mijn ogen inmiddels tot de basisvaardigheden van de e-humanities zou moeten behoren. Of eigenlijk: tot de basisvaardigheid van iedere student op een universiteit, hogeschool en middelbare school.

Natuurlijk heb ik gevraagd welke slimme tools ze gebruikten of kenden. Hun antwoorden, kort samengevat: van sommige tools hadden ze wel gehoord, maar ze gebruikten ze niet of nauwelijks want ze waren te moeilijk en het ontbrak aan duidelijke uitleg.

Dat is ook mijn ervaring. Het is fijn dat er straks – over enkele jaren – websites zijn waarin je kunt zoeken naar honderden tools en honderdduizenden datacollecties. Maar ik zou het al erg fijn vinden als er nu een site zou bestaan met de tien beste en meest gebruikte tools voor bijvoorbeeld historisch en letterkundig onderzoek.

En niet alleen een lijstje met een korte beschrijving, maar iedere tool zou moeten worden voorzien van een heldere handleiding. Plus, nog belangrijker, een korte videocursus die je stap voor stap laat zien: 1. Hoe je de tool installeert, 2. wat je er in grote lijnen mee kunt, 3. hoe je een tool op maat kunt maken. En dan graag een paar datasets erbij zodat je zelf kunt ervaren hoe ongelooflijk nuttig zulk gereedschap kan zijn. Eis van de subsidiegever: de site moet minstens een keer per jaar grondig worden geactualiseerd.

Hoezeer dit ook voor de hand ligt, bij mijn weten ontbreekt het aan zo'n platform. Ik zou er zelf onmiddellijk gebruik van gaan maken, want er kan inmiddels veel meer dan ik weet. Maar net als voor die studenten geldt voor mij: ik wil wel, maar ik weet niet of nauwelijks waar ik moet beginnen.

Ewoud Sanders
Taalhistoricus en journalist. Sanders is vaste medewerker van onder meer NRC Handelsblad en Onze Taal.

Reactie op column Ewoud Sanders

In het woud van tools en data

Ewoud Sanders klaagt in zijn column 'In de e-humanities ziet niemand door de bomen het bos meer' (E-data, oktober 2014) dat er 'nergens […] een overzicht te vinden [is] van alle tools die zijn of worden ontwikkeld'. Hij heeft daar volkomen gelijk in.

Feitelijk is de situatie nog erger: er is ook geen overzicht van alle data die er zijn of die er gaan komen. Maar er zijn ook lichtpuntjes, zoals de Virtual Language Observatory (VLO), een poging van CLARIN (één van de initiatieven voor een geesteswetenschappelijke e-infrastructuur) om een dergelijk overzicht van alle data te maken. Dat is niet eenvoudig, omdat het onder andere vereist dat iedereen dezelfde standaarden gebruikt voor de beschrijving van de data. Er dient nog veel verbeterd te worden aan het overzicht en aan de manieren om in de data te zoeken, maar het is een goed begin. De portal bevat op dit moment ruim 800.000 metadatabestanden. VLO wordt systematisch verder uitgebreid en verbeterd met nieuwe zoekmogelijkheden.

Portal voor services

Daarnaast biedt CLARIN-NL met CLAPOP een online overzicht van inmiddels 40 software services die uit CLARIN-NL voortgekomen zijn, waaronder TICCLops, MigMAP, OpenSONAR en War in Parliament. Daarnaast is er ook een overzicht met tot nu toe 25 datacollecties, zoals Discan en INTER-VIEWS. De portal stelt een gebruiker in staat naar deze data en services te zoeken op basis van vakgebied, taal, functionaliteit, en nog enkele criteria.

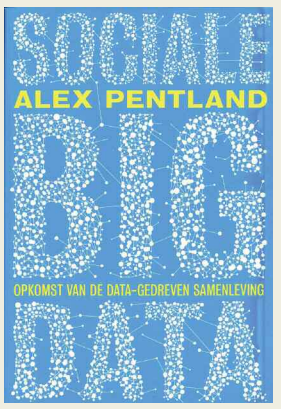
Internationaal

Ook internationaal wordt gewerkt aan brede catalogi voor tools en data in de geesteswetenschappen, zoals Bamboo Dirt. En ook in DASISH, een samenwerkingsverband van vijf grote onderzoeksinfrastructuren, worden dergelijke voorzieningen ontwikkeld. De zoekinterface en de kwaliteit van de metadata in zulke grootschalige overzichten moeten echter aan hoge eisen voldoen om nuttig te zijn voor onderzoekers uit de geesteswetenschappen. En dat is precies de uitdaging die vanuit CLARIN en CLARIAH de komende jaren wordt aangegaan. Ook al is er bij lange na nog geen volledig overzicht van 'alle tools die zijn of worden ontwikkeld', langzaamaan gaan we wel door de bomen het bos weer zien.

Jan Odijk - Directeur CLARIN-NL
Daan Broeder - Technisch directeur CLARIN-NL
portal.clarin.nl

GELEZEN

Sociale Big Data - Opkomst van de data-gedreven samenleving, Alex Pentland, 2014
Erica Renckens



De digitale revolutie biedt veel mogelijkheden voor de sociale wetenschappen, zo blijkt uit het boek Sociale Big Data van Alex Pentland, hoogleraar aan het MediaLab van het Massachusetts Institute of Technology (MIT), Verenigde Staten. Waar onderzoekers vroeger hun conclusies moesten trekken uit laagfrequente, veelal subjectieve metingen onder een select gezelschap, kan nu iedereen continu dienen als proefpersoon. De hele dag door laten we overal concrete sporen van ons gedrag achter, via sociale media, gps, telefoon- en betaalverkeer. Pentland beschrijft in zijn boek hoe hij in deze big data zoekt naar patronen om menselijk gedrag beter te kunnen beschrijven en beïnvloeden. 'Sociale fysica', zoals Pentland het noemt. Na een ietwat taaie introductie neemt Pentland de lezer in verschillende delen mee op een boeiende reis naar de hypothetische toekomst van organisaties, steden en samenlevingen, onder invloed van deze sociale fysica. Hij laat zien wat de uitkomsten van deze sociaal-fysische analyses kunnen betekenen voor de organisatie van bedrijven, wijken, steden en de gehele samenleving. De mogelijkheden klinken veelbelovend, maar er is één essentieel struikelblok: het privacyvraagstuk. Pentland pleit voor nieuwe wetgeving waarin de burger zelf zeggenschap heeft over zijn eigen dataprofiel. Bepalen we straks zelf met wie we welke data delen?
mavenpublishing.nl/boeken/sociale-big data