E-data & Research bestaat 10 jaar

Op deze twee pagina's blikken we terug en vooruit met toonaangevende personen en projecten binnen de alfa- en gammawetenschappen en laten we de uitkomsten van ons lezersonderzoek zien.

E-data & Research wordt mogelijk gemaakt door CentERdata, CLARIAH, DANS, Huygens ING, de KB en het RIVM. Meedoen kan! Als stakeholder en als redactielid. Neem contact met de hoofdredacteur (zie colofon pagina 2) op voor de mogelijkheden.

In gesprek met programmadirecteuren Jaap van den Herik (CATCH) en Jan Odijk (CL

'Datamanagement is reger

Twee grote programma's op het snijvlak van geesteswetenschappen en informatica zijn recent afgerond. De programmadirecteuren maken zich zorgen over de toekomst van data en software. *Erica Renckens*

"Toen we in 2009 begonnen, was er nog een grote kloof tussen de geesteswetenschappers en de taaltechnologen die grote corpora samengesteld hadden," memoreert Jan Odijk, programmadirecteur van CLARIN-NL. "De data waren er wel, maar onderzoekers moesten die zelf downloaden, query's maken, stukjes programmeren. Ons hoofddoel was om dat makkelijker te maken." In 2015 werd CLARIN-NL afgesloten - de verdere ontwikkeling van een digitale onderzoeksinfrastructuur voor de geesteswetenschappen is inmiddels overgenomen door CLARIAH.

"Bij ons was er nog niks toen we begonnen," aldus Jaap van den Herik, programmadirecteur van CATCH, dat afgelopen januari werd afgerond. "Ik vond dat kunstmatige intelligentie niet alleen bij wiskunde en informatica thuishoorde, maar juist ook het toepassingsgebied moest bestrijken. Daarom hebben we een samenwerking tussen de exacte wetenschappen en de gamma's bewerkstelligd, uitgevoerd op de plaats van het cultureel erfgoed door een team van een aio, een postdoc en een programmeur." In tien jaar tijd zijn zo bijna twintig projecten bij erfgoedinstellingen afgerond.

Voorlopers als paradepaardjes

"In het begin moesten we de erfgoedinstellingen natuurlijk nog wel enthousiast maken," vertelt Van den Herik. "Van nature hadden ze geen affiniteit met kunstmatige intelligentie. Maar met twee voorbeeldprojecten kwam de boel al snel op gang."

Odijk: "Wij hebben precies dezelfde ervaring. Wij gingen iets doen voor geesteswetenschappers die daar helemaal niet om gevraagd hadden. Maar elk vakgebied kent wel een paar voorlopers, daar moet je overtuigende projecten mee doen om de rest mee te krijgen. Zo hebben we naast de taalkunde ook veel andere disciplines in de geesteswetenschappen weten te bereiken, zoals geschiedenis, literatuurstudies, filosofie. Daar ben ik wel trots op."

Van den Herik: "Helemaal mee eens. Je moet van de voorlopers in een vakgebied je paradepaardjes maken."

Proefschrift als hoogtepunt

Zowel CATCH als CLARIN-NL kunnen terugkijken op een mooi programma met innovatieve resultaten. "Van 2000 tot 2015 hebben CATCH en CLARIN-NL een duidelijke landmark gezet op de mondiale onderzoekskaart," aldus Van den Herik. "Ik ben erg blij met wat we bereikt hebben, maar dat neemt niet weg dat je je af kunt vragen wat beter had gekund. Dat is wat mij betreft met name de valorisatie. Het proefschrift was het hoogtepunt in de afronding van een project. Daarna zeiden ze 'dankjewel' en dat was het dan. Musea wilden wel verder met wat we ontwikkeld hadden, maar daar was aanvankelijk geen geld voor." Dit inzicht leidde tot het project CATCH-Plus, waarin de succesvolste resultaten uit CATCH werden doorontwikkeld. Daarnaast stapte Van den Herik naar NWO, die vervolgens elk project een half jaar extra gaf voor valorisatie.

Managen van data en software

"Vervolgens had je natuurlijk in Tilburg de Stapel-affaire, waardoor het duidelijk werd dat er iets aan datamanagement gedaan moest worden," stelt Van den Herik. "Inmiddels moet dat wel in elke NWO-aanvraag zijn opgenomen, maar daar was toen nog geen geld voor beschikbaar." Odijk: "Dat is ook een van mijn grootste zorgen. We hebben nu een netwerk van datacentra in Nederland die voldoen aan allerlei voorwaarden om data goed en veilig te bewaren en ter beschikking te stellen. Maar die doen dat nu eigenlijk gratis of in de context van een lopend project. Gelukkig zitten we dankzij CLARIAH wat dit betreft de komende tien jaar goed, maar voor daarna moet er wel een concreet plan liggen.

CATCH
CATCH staat voor Continuous Access to Cultural Heritage. Met de methoden en technieken die zijn ontwikkeld binnen CATCH worden erfgoedcollecties beter toegankelijk voor publiek, professionals en wetenschappers. Een voorbeeld hiervan is SCRATCH, een gebruikersinterface voor software die handgeschreven teksten doorzoekbaar kan maken. Of ZieOok, een applicatie die bezoekers aanbevelingen doet op basis van hun persoonlijke profiel.

CLARIN-NL
CLARIN-NL staat voor Common Language Resources and Technology Infrastructure. Binnen CLARIN-NL zijn bestaande datasets en tools omgezet naar duurzame standaarden, zodat het geheel interoperabel wordt en nieuw onderzoek mogelijk maakt. Dit heeft bijvoorbeeld in het War in Parliament-project tot een applicatie geleid waarin de Handelingen der Staten-Generaal doorzoekbaar zijn. De Mimore-tool maakt het mogelijk om in drie verschillende dialectologische databanken tegelijk te zoeken. De resultaten van CLARIN-NL zijn beschikbaar via de portal.clarin.nl.

Online dataverzameling in Nederland

Een kwart eeuw survey-onderzoek

25 jaar geleden stapte ik bij Stichting Telepanel aan de UvA binnen. De bakermat van online survey-onderzoek in Nederland, zeggen velen nu. We werkten met een Sinclair-computer zonder harde schijf met modem, het scherm was de eigen TV van de respondent. *Corrie Vis*

De surveys werden elke vrijdag via het modem opgestuurd aan de panelleden. Men deed graag mee, uit nieuwsgierigheid en om de wetenschap te helpen. Het interviewprogramma was robuust en klein genoeg voor een floppy. Voor het aansluiten van de apparatuur kwam een medewerker bij de panelleden langs. Onderzoek dat schriftelijk eigenlijk niet of nauwelijks te doen was, kwam met deze nieuwe techniek binnen handbereik. Een voorbeeld hiervan is onderzoek naar de levensloop, met onderwerpen als wonen, relaties, opleidingen, arbeid en inkomen. Deze computergestuurde surveys waren gebaseerd op de werking van het geheugen met herhaalde vraagblokken en complexe *routing*. Ze waren weldoordacht, maar de resulterende databestanden hielden in het begin geen gelijke tred. Soms zaten ze zo ingewikkeld in elkaar dat ze niet zomaar te analyseren waren.

De technische verandering in dataverzameling leverde ook nieuwe aandachtsgebieden op zoals: de tijdbelasting van de respondent, de cognitieve eisen die worden gesteld en het experimentele design. Computergestuurd survey-onderzoek heeft zich bewezen en is inmiddels een zelfstandige richting in de sociale wetenschappen.

Te veel, te lang

Inmiddels is het afnemen van een survey via internet eenvoudig. De doe-het-zelf-softwarepakketten zijn legio. Wat vond u van onze service? Heeft u kleding op internet gekocht? Wat vindt u van onze website? Deze wijze van dataverzameling is snel en kostenefficiënt, maar het is geen wonder dat steeds minder mensen zin hebben om vragen te beantwoorden.

Ook voor sociaalwetenschappelijk onderzoek worden te veel lange, computergestuurde surveys ontwikkeld. Er zijn echter goede initiatieven gaande om het aantal en de lengte van surveys te verminderen. Bijvoorbeeld door surveydata te koppelen aan administratieve data, verzamelde data te archiveren en toegankelijk te maken voor andere onderzoekers en *data scientists* op te leiden om big data te analyseren en te interpreteren.

Op waarde schatten

Surveys in de sociale wetenschappen blijven nodig. Zelfs als iedereen een chipje heeft om zijn gezondheidstoestand te monitoren, zullen we willen weten hoe iemand zich voelt. Met name longitudinaal onderzoek vereist dat we als onderzoekers de respondenten weer op waarde moeten schatten. Een effectieve manier hiervoor is aandacht geven aan de mensen van wie we iets willen weten en een betrouwbare relatie met hen opbouwen. Zorg ervoor dat het beantwoorden van vragen interessant en zinvol is. Alleen dan kunnen we ook in de toekomst goede antwoorden verwachten.

credits nachtblauw.com