Focus

CLARIN-NL – Common Language Resources and Technology Infrastructure in Nederland

INGE ANGEVAARE

Bestaande digitale bestanden en de *tools* om ze te gebruiken zo goed zichtbaar en toegankelijk maken dat vele onderzoekers in de geesteswetenschappen er nieuw onderzoek mee kunnen doen. Dat is de missie van het project CLARIN-NL. *e-data&research* sprak met bestuursleden Jan Odijk (Utrecht) en Arjan van Hessen (Twente) op het Programmabureau van CLARIN-NL aan de Trans in hartje Utrecht.

In de alfawetenschappen behoorden taal- en spraakwetenschappers bij de eerste groepen die de grote mogelijkheden van het digitale tijdperk ontdekten. Ze begonnen al snel met het samenstellen van digitale tekstcorpora en het ontwikkelen van gereedschappen om die te kunnen analyseren. Maar omdat ieder dat op zijn eigen manier deed, kon men elkaars data niet goed gebruiken; de benodigde interoperabiliteit ontbrak. Jan Odijk: 'Zelfs een simpel overzicht van wat er allemaal in Nederland beschikbaar is, ontbreekt.' CLARIN-NL wil dat veranderen door, zoals het officieel heet, een *e-science* infrastructuur te bouwen voor talige data en tools in de geesteswetenschappen. 'Wij richten ons daarbij op alle onderzoekers uit de geestes- en sociale wetenschappen die met talig materiaal werken, niet alleen op de taal- en spraakwetenschappen,' benadrukt Jan Odijk. 'Ook veel onderzoekers in de geestes- en sociale wetenschappen kunnen gebruik maken van de corpora en de tools om die te exploreren.' Als ze tenminste goed vindbaar zijn, goed gestructureerd zijn, en duurzaam beheerd worden. Het is de kern van het CLARIN-project om alles zo te structureren dat tools en data van verschillende herkomst goed in combinatie gebruikt kunnen worden.

Metadata spelen hierbij een cruciale rol. CLARIN-NL werkt niet met één vast metadataschema, maar met een flexibel, modulair systeem. Onderzoekers kunnen daarin zo nodig zelf componenten aanmaken als zij die nodig hebben. Daarnaast moeten de data voldoen aan algemeen geaccepteerde en binnen CLARIN ondersteunde technische standaarden.

INGE ANGEVAARE

Jan Odijk (links) en Arjan van Hessen: 'Het lukt ons steeds beter om ook minder technisch georiënteerde onderzoekgroepen bewust te maken van de mogelijkheden'

Duurzaamheid niet vanzelfsprekend

Duurzaam databeheer is nog lang niet overal ingeburgerd. Daarom schrijven de CLARIN-regels voor dat onderzoeksgegevens na afloop van projecten moeten worden ondergebracht bij één van de vijf CLARIN-centra: het Max Planck Instituut voor Psycholinguïstiek, het Meertens Instituut, het Instituut voor Nederlandse Lexicologie, het Huygens Instituut en DANS. CLARIN-NL hoopt dat dit netwerk uitgebreid zal worden en dat op korte termijn ook beheerders van enorme digitale corpora als de Koninklijke Bibliotheek, het Nationaal Archief en het Nederlands Instituut voor Beeld en Geluid data zullen gaan leveren die passen in het CLARIN-systeem.

Maar ook de studenten en onderzoekers zelf moeten bewust worden gemaakt van de mogelijkheden. Van Hessen: 'Vooral de opleiders spelen hier een sleutelrol. Door cursussen te ontwikkelen en gastcolleges te geven, maken we opleiders en studenten bewust van de mogelijkheden en leren we ze te werken met grote digitale bestanden. We streven ernaar de technieken onderdeel te laten worden van het standaardcurriculum in de geesteswetenschappen, zowel in de bachelor- als in de masterfase.'

Onderzoekers niet altijd informatici

CLARIN-NL wil ook onderzoekers helpen bij het omzetten van hun gegevens naar CLARIN-standaarden. Van Hessen: 'Je mag niet van elke onderzoeker verwachten dat hij technisch onderlegd is'. De hulpschermen naar allerlei functies binnen de CLARIN-infrastructuur zoals zoek- en browsefuncties of het exploreren of bewerken van data, moeten daarom zeer intuïtief en gebruiksvriendelijk zijn.

Het totale budget is negen miljoen euro. Odijk: 'Daar kunnen we heel wat mee bereiken, vooral op het vlak van bewustwording. We ontwikkelen in elk geval overtuigende *showcases* om aan onderzoekers te laten zien wat er allemaal mogelijk is, zodat het voor wetenschappers vanzelfsprekend wordt om de CLARIN-regels te volgen omdat je anders de boot mist.'

www.clarin.nl;
www.isocat.org

Interedition: interoperabiliteit voor duurzaamheid

Op een recente bijeenkomst van Interedition op de Ludwig Maximilians Universität in München telden samengeschoolde literair-historici en IT-onderzoekers in nauwelijks vijf minuten niet minder dan achttien projecten om software te ontwikkelen voor de transcriptie en annotatie van gedigitaliseerde teksten. In de geesteswetenschappen, waar budget en capaciteit voor IT-ontwikkeling steevast beperkt zijn, is het opmerkelijk dat digitale instrumenten met hetzelfde doel in veelvoud ontwikkeld worden. Er schuilt ook een aanzienlijk risico in, omdat het onderhoud dat nodig is om de ontwikkelingen bij te benen vaak niet te financieren is. De toegankelijkheid en het behoud van de instrumenten en gerelateerde data komen daardoor snel onder druk te staan.

Sleutel tot het ontwikkelen van duurzamere digitale gereedschappen is interoperabiliteit, aldus de werkgroep 'Strategic IT Recommendations' van het Interedition-project. De gedachte is dat *tools* die elkaars gegevens en processen kunnen gebruiken - dus interoperabel zijn - meer gedefinieerde en gedeelde werkprocessen vereisen. Die kunnen op hun beurt leiden tot efficiëntere spreiding van de verantwoordelijkheden voor het bouwen en onderhouden van de software voor zulke processen. Interoperabiliteit wordt in het Interedition-model breed opgevat. Het is niet alleen een technische eigenschap die zorgt dat programma's met elkaar kunnen praten, maar heeft ook een sociaal aspect: zorg dragen dat betrokken specialisten kennis kunnen uitwisselen en kunnen samenwerken. Tenslotte betekent interoperabiliteit op methodologisch vlak de identificatie van congruente werkprocessen.

Dat dit alles niet slechts theorie is laten de *proof of concept* producten van Interedition zien. CollateX is van deze tools met een actuele 1.0 release het verst gevorderd. Het is software die met literair-kritische precisie variatie in verwante teksten opspoort. Dit is bijvoorbeeld relevant bij de analyse van Darwins *Origin of Species*. Dat werk verscheen in achttien opeenvolgende edities en kende nogal wat mutaties, wat tot debat leidde of Darwin stelliger dan wel onzekerder werd in de loop van de tijd. De analyse van de variatie in extenso die nu mogelijk is, leidt tot de conclusie dat Darwin steeds overtuigder werd van zijn inzichten.

De ontwikkelingsgeschiedenis van CollateX toont de relevantie van Interedition's interoperabiliteitsmodel aan. Gezamenlijke bijeenkomsten voor tekstonderzoekers en ontwikkelaars mondden uit in de definitie van een methodisch werkproces voor digitale collatie van variante teksten, het zogenaamde Gothenburgmodel. Daarnaast werden *bootcamps* voor IT-ontwikkelaars georganiseerd. Deze boden gelegenheid om het methodische model in een coproductie tussen ontwikkelaars wereldwijd te implementeren. Daarmee legde Interedition de basis voor een Open Source development community in de geesteswetenschappen. Zo'n community kan op termijn leiden tot een beter geïntegreerde aanpak van software-ontwikkeling voor de geesteswetenschap. Op technisch niveau leidde het interoperabiliteitsdenken tot een model van microservices. In dit model wordt een groter werkproces zoals de tekstcollatie door CollateX opgedeeld in een aantal kleinere services die onafhankelijk op een gedistribueerde infrastructuur (de *cloud*) kunnen bestaan. De implementatie van CollateX als een set van dergelijke gedecentraliseerde webservices maakt het delen en onderhouden van technische resources praktischer. Mede daardoor maken nu meerdere Europese, Amerikaanse en Canadese projecten gebruik van de CollateX webservice, en dragen zij actief bij aan de ontwikkeling ervan. (Joris van Zundert)

www.interedition.eu
http://collatex.sourceforge.net/

CATCHPlus maakt doorstart

Na de verhuizing van het projectbureau CATCHPlus van het Instituut voor Beeld en Geluid naar het Meertens Instituut heeft het met een grotendeels nieuw team een goede doorstart gemaakt.

CATCHPlus is in het leven geroepen om de onderzoeksresultaten van CATCH (*Continuous Access To Cultural Heritage*) te verzilveren door bruikbare tools en diensten voor de hele Nederlandse erfgoedsector op te leveren. 'Het is het knooppunt tussen ICT en erfgoed,' volgens de nieuwe projectleider Patricia Alkhoven. 'Het gaat om een unieke samenwerking tussen grote erfgoedinstellingen, universiteiten en bedrijfsleven. Dat is ingewikkeld en het is dan ook een uitdaging om resultaten te realiseren. We hebben nu in elk geval de zekerheid dat het project in zijn geheel volbracht kan worden, want we kregen onlangs het bericht dat de aanvraag bij het ministerie van Economische Zaken, Landbouw en Innovatie (EL&I) voor het laatste gedeelte van de financiering gehonoreerd is.'

CATCHPlus werkt aan een centrale infrastructuur om informatie en software te delen, met daarbinnen een repository voor het gemeenschappelijk opslaan en beheren van annotatiedata. Het project wordt gefinancierd door het ministerie van Onderwijs, Cultuur en Wetenschap, de Nederlandse Organisatie voor Wetenschappelijk Onderzoek en het ministerie van EL&I. Het loopt tot medio 2012. (Douwe Zeldenrust)

www.catchplus.nl