Multifunctioneel informatiewetenschapper Herbert Van de Sompel:

‘Iedereen wil een oplossing om webarchieven toegankelijk te maken, en wij hebben die oplossing!’

Tijdreizen op het web is een van de opvallende oogmerken van de in de Verenigde Staten werkende Belg Herbert Van de Sompel. Maar de veelzijdige informatiewetenschapper heeft meer spraakmakende internetprojecten op zijn naam staan. *e-data&research* verkende de ruime horizon van de geboren innovator tijdens een verblijf in Nederland.

PETER BOOT EN MARTIJN DE GROOT

Vanuit zijn werkplek bij het befaamde Los Alamos National Laboratory in New Mexico (VS) streek Van de Sompel deze zomer twee maanden als *visiting professor* neer bij DANS. Het duurde niet lang of de informatiewetenschapper had bij het gastheer-instituut en bij de Koninklijke Bibliotheek zijn plannen besproken voor de internet-teletijdmachine Memento, en bij de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) zijn visie op Open Annotatie voor de geesteswetenschappen.

Alfalab, het KNAW-project dat het gebruik van digitale gereedschappen in de humanities bevordert, biedt een aanknopingspunt om dat laatste onderwerp te belichten: ‘Zij hebben twee demonstratieprojecten. Textlab werkt aan de annotatie van historische manuscripten, Spacelab aan die van kaarten met microtoponiemen, veldnamen dus. Die annotaties wil je kunnen delen tussen gebruikersgroepen en collecties, en dat moet Open Annotatie mogelijk maken door een gemeenschappelijk datamodel te definiëren. Daarmee kunnen verschillende projecten hun annotaties publiceren, waarna robots en harvesters de data eruit kunnen gaan zuigen en daarop nieuwe diensten aanbieden. Maar daarvoor zijn wel interoperabiliteitsspecificaties nodig.’

WIEBE KIESTRA

Van de Sompel, in Den Haag neergestreken vanuit New Mexico, had al snel oog voor de efficiency van het vervoer per fiets in zijn tijdelijke verblijfplaats.

Waarom interoperabiliteit?

‘De gedachte is dat je mensen niet wilt dwingen hun interne systemen te veranderen. Wat je wel wil is een gemeenschappelijke interface die met alle partners kan samenwerken. Je moet dus specificaties opstellen die het voor de meeste annotatiesystemen mogelijk maken om mee te doen.’

En is er ook een verband met het populaire Firefox-hulpmiddel Zotero?

‘Ook Zotero, dat onderdeel is van het Open Annotation Project, zou van het gedeelde datamodel gebruik moeten gaan maken. Zotero is vooral ontwikkeld om gegevens te verzamelen over bronnen, zoals webpagina’s en wetenschappelijke literatuur. Daarvoor wordt nu een annotatiecomponent ontwikkeld, maar tot nog toe hebben ze zich geconcentreerd op de *user interface*, niet op de interoperabiliteit. Nu houden ze nog geen URI’s bij. Dat moet natuurlijk veranderen. De Open Annotation Collaboration is geheel webgecentreerd.’

Uw project voor toegankelijkheid van webarchieven hangt hier eigenlijk direct mee samen.

‘Ja, het is erg belangrijk dat annotaties in de tijd constant blijven. Je gaat internetbronnen annoteren, maar zo’n bron kan er morgen al anders uit zien. Je zou dus willen verwijzen naar een URI op een bepaald moment in de tijd, maar dan moet je die vroegere toestand van de URI wel terug kunnen vinden. Dat is het doel van Memento: tijdreizen op het web. En dat is natuurlijk niet alleen van belang voor annotaties. Het gaat ook om persberichten van het Witte Huis, commerciële en juridische gegevens, wetenschappelijke literatuur, en *what have you*. En het bewaren van versies van andere bronnen, zoals data, onderzoeksdata ook, dat is ècht belangrijk! Iedereen weet al zo lang dat we een oplossing nodig hebben voor het toegankelijk maken van archieven, en wij hebben die oplossing!

Je systeem heeft dus een archiveringskant nodig. Dat kan op verschillende manieren. Websites kunnen een eigen archief bijhouden, zoals de British Library al doet. Je kunt vertrouwen op externe archivering, bijvoorbeeld door het Internet Archive of door een nationale bibliotheek. Combinaties kunnen ook. De architectuur die wij voorstellen voorziet ook een plaats voor *aggregators*: instellingen die informatie over beschikbare archieven verzamelen.

In het beste geval maakt een website zelf duidelijk waar het archief te vinden is. Dat gebeurt met een http header, extra informatie die een web server naar de browser of client terugstuurt, extra informatie die met een webpagina wordt meegestuurd. De universiteitsbibliotheek van Gent doet dat al. De client, bijvoorbeeld een browser zoals Firefox, maakt dan uit de Memento-header op waar hij het archief voor een bepaalde datum moet zoeken. Van DBPedia, de database-versie van Wikipedia, bestaat intussen al een Memento-compatibele versie. Die hebben wij opgezet, als voorbeeld. Maar meestal zullen die Memento-headers er niet zijn. Dan kan de client op zoek gaan naar een aggregator die de metadata al heeft verzameld. Wij hebben daar een proefopstelling van gemaakt, en het werkt.’

Is Memento het stadium van het experiment dus al voorbij?

‘Het is een experiment in de zin dat er nog geen officiële specificaties bestaan. Maar het slaat wel aan. Bij de laatste World Wide Web conferentie heb ik het verhaal over Memento verteld met Tim Berners-Lee op de eerste rij. Die had fenomenaal lovende woorden (glimlacht trots).

De invoering van Memento hoeft niet ingewikkeld te zijn. De meeste webarchieven draaien allemaal dezelfde software, WayBack. Dat hebben we bekeken en geconstateerd dat het een fluitje van een cent is om ze Memento-compatibel te maken. En wat betreft de clients: nu werken we met plug-ins, in te bouwen hulpprogramma’s in de browser. We hopen dat na verloop van tijd de browsers standaard Memento-mogelijkheden zullen aanbieden. We praten nu met Microsoft over die mogelijkheid in Internet Explorer.

Er wordt vaak gevochten om de status van iso-standaard of een W3C recommendation. Uw standaarden hebben die status vaak niet maar ze worden wel op grote schaal gebruikt.

‘We hebben er wel eentje: de OpenURL. Maar ik heb geen duidelijk antwoord hoe dat komt. Eén denkbare oorzaak is dat het om echte problemen gaat die mensen graag opgelost willen zien. Bijvoorbeeld het Protocol for Metadata Harvesting PMH, dat was echt een *burning platform*! OpenURL was ook zoiets. Het ging erom te kunnen verwijzen naar wetenschappelijke artikelen in systemen van bibliotheken en wetenschappelijke uitgevers. Daarbij moet je rekening kunnen houden met de abonnementen die een bibliotheek heeft afgesloten. We hebben daar een elegante oplossing voor bedacht.’

U heeft met cognitief psycholoog Johan Bollen ook een systeem ontwikkeld voor het aanbevelen van wetenschappelijke artikelen?

‘Bollen en ik kwamen toevallig allebei uit België naar Los Alamos. We realiseerden ons dat de OpenURL software precies kan bijhouden hoe vaak er op links naar wetenschappelijke artikelen wordt geklikt. Dan weet je dus ook in welke context wetenschappers in bepaalde artikelen geïnteresseerd zijn. Daarmee kun je dezelfde soort dingen doen als Amazon: *kopers die dit boek kochten, kochten ook deze boeken*. Toen ik later opnieuw in Los Alamos werkte heb ik Johan teruggehaald om het idee uit te werken. Dat heeft geleid tot de *bX Recommender Service*. Die is overgenomen door het bedrijf Ex Libris.’

U werkt ook aan Object Reuse and Exchange, ORE.

‘Daarbij gaat het om samengestelde objecten. Bij een artikel horen onderzoekgegevens, een video, afbeeldingen, en die moet je aan elkaar kunnen koppelen. Elk een eigen identiteit, maar je moet ook de aggregatie ervan, die een machineleesbare beschrijving geeft van metadata en componenten, een eigen identiteit kunnen geven.

Er was veel belangstelling voor het probleem maar het was geen *burning platform*. Daardoor kwam het wat langzamer op gang maar er wordt toch op veel plekken mee gewerkt, bijvoorbeeld in de DataNeT projecten in de VS, in de Europese digitale bibliotheek Europeana, maar ook hier bij DANS in het werk aan verrijkte publicaties.’

Is ORE bedoeld voor de wetenschappelijke uitgevers, die de gegevens over artikelen met ORE ter beschikking moeten stellen?

‘Met uitgevers weet je het nooit. Vrijwel geen enkele uitgever heeft ooit het Protocol for Metadata Harvesting ingevoerd, dus... Ik hoop dat het gebeurt want dat zou veel dingen oplossen. Maar veel uitgevers verstoppen nog altijd de metadata van publicaties. Het is echt een enorme blunder van de Open Access Movement dat men dat heeft laten liggen. Metadata vormen het *core* niveau, dat moet op nummer 1 staan. Het is *fuel for the engine*. Uitgevers weten dat ook. In hun nadruk op open access voor de inhoud van de artikelen zijn activisten en bibliotheken de metadata vergeten. Dat is kwalijk.’

Herbert Van de Sompel studeerde in 1979 in Gent af als wiskundige en twee jaar later als computerwetenschapper, waarna hij in die plaats zijn loopbaan begon als hoofd bibliotheekautomatisering. Na onder meer een kort verblijf als research fellow aan het Los Alamos National Laboratory (LANL) promoveerde hij in 2000 op een proefschrift over het dynamisch linken van onderzoekbronnen. Na korte verbintenissen met The British Library en Cornell University werd hij in 2002 opnieuw aangesteld, nu als stafonderzoeker, bij het LANL waar hij tegenwoordig nog steeds aan is verbonden.