Van de Sompel, Chief Innovation Officer bij DANS:

‘Automatisch artefacten vinden, ophalen en duurzaam archiveren’

“Onderzoekers willen onderdeel uitmaken van hun sociale netwerk op het web, ook voor hun onderzoek.”
E-data interviewt Herbert van de Sompel, per 1 januari Chief Innovation Officer bij DANS.

Marion Wittenberg

Herbert van de Sompel
Van de Sompel (1957) studeerde wiskunde en computerwetenschap aan de universiteit van Gent. Hij promoveerde in 2000 op een proefschrift over contextgevoelig en dynamisch linken van wetenschappelijke informatie.
Vanaf 2002 was hij leider van het onderzoeksteam ‘Digital Library Research and Prototyping’ aan het Los Alamos National Laboratory (New Mexico, USA). Hij is één van de grondleggers van veel gebruikte informatiestandaarden (onder meer OAI-PMH, OpenURL, OAI-ORE, Memento, NISO/OAI ResourceSync, Web Annotation) en nam deel aan invloedrijk onderzoek naar alternatieve metrieken voor wetenschappelijke publicaties (MESUR-project) en ‘reference rot’ in wetenschappelijke communicatie (Hiberlink project).
In 2017 ontving hij de prestigieuze Paul Evan Peters Award voor zijn bijdragen aan duurzame digitale infrastructuren die een diepgaande en blijvende impact hebben gehad op de wetenschappelijke communicatie.
Van de Sompel was in 2010/2011 en in 2013/2014 visiting fellow bij DANS.
hvdsomp.info/bio/

“Onderzoekers zetten overal op het web artefacten neer van hun onderzoeksactiviteiten in voor hen aantrekkelijke systemen”.
credits Bart van Vliet

INTERVIEW

Na bijna twintig jaar in de USA gewerkt te hebben, is Herbert van de Sompel, grondlegger van onder andere het OAI-PMH-protocol, op 1 januari aan de slag gegaan bij DANS. DANS stimuleert onderzoekers om hun digitale onderzoeksgegevens vindbaar, toegankelijk, interoperabel en herbruikbaar te maken. Welke plannen heeft van de Sompel bij DANS? “Ik wil eerst een goed overzicht krijgen van de ontwikkelingen en projecten bij DANS en daarna bekijken wat ik vanuit mijn expertise kan inbrengen. Er zijn een aantal concepten uit mijn recente werk bij het Los Alamos National Laboratory, onder andere het ‘Scholarly Orphans Project’, die hierbij mogelijk interessant kunnen zijn”.

Een spoor van artefacten

“Onderzoekers zetten overal op het web artefacten neer van hun onderzoeksactiviteiten in voor hen aantrekkelijke systemen: presentaties in SlideShare, code in GitHub, data in Figshare of Zenodo. Ze kiezen voor deze globale portals omdat deze zich binnen hun sociale netwerk bevinden en hun zichtbaarheid vergroten. Onderzoekers verplichten om al die bijdragen ook bij hun instelling te deponeren, is problematisch: de toegevoegde waarde voor de onderzoeker is niet meteen duidelijk. Het resultaat is dat instituten geen volledig zicht hebben op wat hun onderzoekers doen, ze zien het uiteindelijke paper, maar niet alle stappen die tot dat paper geleid hebben. En artefacten die op commerciële platforms gedeponeerd worden kunnen ook zomaar verdwijnen”. Het ‘Scholarly Orphans Project’ onderzoekt dit probleem en heeft een prototype ontwikkeld dat automatisch artefacten vindt, ze ophaalt en duurzaam archiveert. “In plaats van de onderzoeker verplicht te stellen hun artefacten bij hun instelling te deponeren, draait de instelling processen waarmee ze die volautomatisch binnenhalen.”

Myresearch.institute

Van de Sompel legt uit dat het project uitgaat van twee perspectieven. Ten eerste dat onderzoeksinstellingen de facto geïnteresseerd zijn in de artefacten van hun onderzoekers. Ten tweede dat de processen van ophalen en archiveren schaalbaar moeten zijn, het gaat om heel veel materiaal. De combinatie van beide perspectieven heeft geleid tot een prototype voor een fictief onderzoeksinstituut: myresearch.institute. “Voor een tiental onderzoekers, die geselecteerd werden omdat ze erg actief op het web zijn, werden de identiteiten die ze in de webportals gebruiken, verzameld. Die identiteiten worden gebruikt als sleutel voor de portal APIs om dagelijks te kijken of een onderzoeker iets nieuws heeft gedeponeerd. Bij GitHub kan dat al gauw om 50 tot 100 nieuwe bijdragen per onderzoeker per dag gaan. Metadata over die bijdragen wordt in een institutionele databank gestopt en de bijdragen zelf worden met web-archiveringstechnieken opgehaald en gearchiveerd”. Voor het ophalen van het materiaal heeft het ‘Scholarly Orphans Project’ een innovatieve procedure ontwikkeld, Memento Tracer. “Omdat steeds meer websites voor gebruikersinteractie van client-side JavaScript gebruikmaken, zijn ze moeilijk automatisch te archiveren. De enige techniek om dergelijke webpagina’s met hoge kwaliteit te archiveren, is om een gebruiker alle essentiële interacties te laten uitvoeren en de resultaten van die interacties weg te schrijven naar een webarchiveringsbestand. Perfect, maar niet schaalbaar. Bij de Memento-Tracer-aanpak doet een curator eenmalig een sessie van interacties met een bepaald type pagina, bijvoorbeeld een landingspagina van SlideShare. Een browser plugin legt deze interacties vast: wat is er achter de schermen gebeurd toen de curator aan het klikken was, welke JavaScript calls zijn aangeroepen. Dit levert een JSON-file met instructies op die als template wordt gebruikt voor een web-archivering crawler. Met deze methode kunnen alle pagina’s van hetzelfde type op industriële schaal en met hoge kwaliteit geharvest worden. Het is momenteel nog experimenteel, maar het concept is echt een doorbraak die we van tevoren niet verwacht hadden”.

‘Volg het online spoor van de onderzoeker’

Duurzaam archiveren

“Initieel waren enkel de eindresultaten van wetenschappelijk onderzoek (papers) beschikbaar op het web. In toenemende mate zijn ook artefacten die tijdens het wetenschappelijk proces gemaakt worden daar te vinden. Er zijn momenteel nog geen frameworks die dat alles archiveren”. Volgens van de Sompel is dit probleemdomein relevant voor DANS. “Het duurzaam bewaren van data is momenteel de missie van DANS, het archiveren van artefacten zou hier naadloos op aan kunnen sluiten. De onderzoeker geeft aan DANS een overzicht van zijn identiteiten bij de door hem gebruikte portals, DANS haalt alles op en archiveert het materiaal”. Dit kan zowel een dienst voor de onderzoeker zijn, als voor de instellingen. “De onderzoeker blijft doen wat hij of zij doet, maar al het materiaal wordt automatisch gearchiveerd. En DANS kan dit vervolgens terugleveren aan de instellingen, waardoor de instellingen een overzicht krijgen van al het materiaal wat hun onderzoekers gecreëerd hebben. Zo zorgen onderzoekers, instellingen en DANS samen voor persistentie van de wetenschappelijke record.”

Technologie en beleid

We vroegen van de Sompel wat hem aantrekt in zijn nieuwe functie bij DANS. De meerwaarde van DANS is in zijn ogen de combinatie van technologie en beleid: “DANS is een voortrekker op dit gebied. Technologisch kan er heel veel, maar het beleid moet de context bieden, het zorgt voor meer realiteitszin”.

myresearch.institute
tracer.mementoweb.org