De nieuwe online databank bij het VTi

Open data en opensourcesoftware

In dit artikel worden de overwegingen geschetst die bij het Vlaams Theater Instituut, steunpunt voor podiumkunsten, geleid hebben tot het opzetten van een grote vernieuwing van zijn IT. Eerst en vooral werden Microsoft Access en SQL Server vervangen door opensourcealternatieven en werd een nieuwe webinterface voor de gespecialiseerde metadata gebouwd, zowel voor de documentatiemedewerkers als voor gasten en voor online bezoekers. Vervolgens werd de inhoud van de databank ook principieel open gesteld voor hergebruik door anderen. Niet door een achterpoortje te openen voor Open Archives Initiative (OAI[1]) of een ander protocol, maar door de toepassing van Linked Open Data, een technische strategie uit het semantic web die het zinvol uitwisselen van data tussen computers vereenvoudigt. Tot slot besliste het VTi om ook de met Ruby on Rails ontwikkelde software als opensourceapplicatie beschikbaar te stellen en de mogelijkheden voor een breder draagvlak voor deze applicatie te onderzoeken.[2]

Dries Moreels (links)
Informatiebeheerder bij BAM

Tom Klaasen (rechts)
Zaakvoerder van 10 to 1

Aanleiding en context

Sinds het einde van de jaren tachtig is het VTi[3] een doorgedreven documentatie over de actuele podiumkunsten in Vlaanderen aan het opbouwen en van bij de start was er ook het bewustzijn dat databanken bij het beschrijven en beheren van die documentatie van onschatbare waarde zouden zijn. Microsoft Access was gedurende vele jaren het instrument om deze databank helemaal rond de specifieke context van de podiumkunsten-documentatie en de specifieke noden van de medewerkers te organiseren. In 2009 is deze databank totaal vernieuwd met opensourcecomponenten en up-to-date webtechnologie: data.vti.be. De Franstalige collega's van Contredanse[4] zijn met dezelfde technologie aan de slag gegaan en lanceren hun toepassing daarover begin 2011. Maar ook andere gespecialiseerde documentatiecentra kunnen met de in deze applicatie gebundelde technologieën hun bestanden openen.

De kern van de VTi-databank bestaat uit detailfiches over elke professionele dans- of theaterproductie sinds 1993-1994: beschrijvende basisinformatie, de producenten en organisatoren, de volledige cast en voor de periode 1998-2006 zelfs integrale voorstellingskalenders. Ook voor personen en organisaties worden er extra beschrijvende elementen opgeslagen (subsidies, contactinformatie, opleiding enz.). Omdat alle vermeldingen van personen en organisaties dynamische links zijn, kan een relatie tussen een persoon en een productie worden weergegeven in een cast van een productie, maar ook omgekeerd als een element op het cv van de betrokkenen. Dus: uit de cast van een opera doorklikken naar de lijst van alle opera's die bijvoorbeeld José Van Dam in Vlaanderen en Brussel gezongen heeft. Meer nog, uit de combinatie van de internationale speelreeksen met de historieken van producenten en co-producenten kan statistische informatie gepuurd worden over recente evoluties in internationale spreiding van theater of dans.

De klassieke titelbeschrijving van de documentatiestukken, boeken, tijdschriften enz. vormt de tweede belangrijke pijler in de databank. Ook daarbij wordt voor een geïntegreerde aanpak gekozen, waardoor referenties naar boeken en tijdschriften ook als een selectieve bibliografie van kunstenaars en kunstorganisaties functioneren. Je kan dus uit de titelbeschrijvingen van tijdschriftartikels in één beweging ook een lijst van artikels geschreven door bijvoorbeeld Josse De Pauw genereren, naast de lijsten van de producties waar-

aan hij meewerkte of de organisaties waaraan hij verbonden is of was.

Enkele van de commerciële automatiseringspakketten kunnen hetzelfde aanbieden door de combinatie van diverse modules voor bibliotheek-, museum- en archiefmanagement en het toevoegen van 'custom fields'. Toch is er bij de vernieuwing van deze databank gekozen voor software op maat. Daarbij speelden verschillende argumenten: een economische overweging van TCO (total cost of ownership) over vijf jaar, de specifieke inhoudelijke en logistieke noden van een documentatiecentrum als het VTi, het belang van actuele webtechnologie (semantic web) en doorgedreven meertaligheid.

Functionaliteit

De nieuwe webapplicatie heeft uiteraard als eerste functie om allerlei documenten te beschrijven, beschrijvingen te verbeteren, inhoudelijke verbanden aan te geven enzovoort. Wat dat betreft is de applicatie heel voorspelbaar. Maar om maximaal van het internet gebruik te kunnen maken, heeft alles in de databank een permanent en zinvolle URI gekregen (een zogenaamd 'Cool URI', www.w3.org/TR/cooluris/) die geen vreemde tekens bevat, niet meer wijzigt, in een oogopslag begrijpelijk is en in diverse zoekmachines daardoor optimaal werkt.

Als je op de homepage een zoekvraag over documenten rond ecologie lanceert, dan vind je allerlei resultaten, waaronder http://data.vti.be/articles/platform-duurzaamheid, een tijdschriftartikel uit het Nederlandse tijdschrift / periodicals/zichtlijnen, geschreven door /people/jeroen-de-leeuw. Het tijdschrift zelf zit in een archiefdoos, waarvan je de precieze plek kan oproepen: /warehouses/lav6-5000012 71. Er zijn natuurlijk nog meer items met het woord ecologie getagd: /tags/ecologie-2 en zo kom je snel bij /book_titles/performing-nature-explorations-in-ecology-and-the-arts uit, waarvan /people/gabriella-giannachi een van de auteurs is. Je ziet meteen waarom die naam een belletje deed rinkelen, want ze is ook de auteur van /book_titles/virtual-theatres-an-introduction, een topwerk over /tags/technologie in de podiumkunsten. Als je verder blijft surfen kom je misschien bij een ongepubliceerd document uit, bijvoorbeeld /ephemera/crew-vandaag-de-actualiteit-van-het-oneigentijdse-theater, waarin het vernieuwende werk van CREW_eric joris (/organisations/creweric-joris) wordt geanalyseerd. Om meer te weten over dit gezelschap kan je de lijst met hun producties oproepen /organisations/creweric-joris/production_by_organisations of een blik werpen op hun subsidiehistoriek (/organisations/creweric-joris/grants) of nakijken welke andere relevante documenten nog beschikbaar zijn (/organisations/creweric-joris/subject_of_documents). Uiteindelijk valt je oog op een korte video, /audio_video_titles/w-double-u-fragment, waarvan verschillende formaten beschikbaar zijn. Je klikt er een aan en de precieze vindplaats komt te voorschijn: /audio_video_media/opt20100020.

VTi Steunpunt Podiumkunsten (www.vti.be) - Foto Phile Deprez.

Op deze manier zijn alle dimensies en aspecten van de documentaire informatie precies te benoemen, zodat er op elk van die dimensies extra data bewaard kan worden. Het zijn aparte, maar gerelateerde digitale 'objecten' geworden. Zo kan je bijvoorbeeld registreren dat een bepaalde danser een solo had in een ballet, plus welk personage dat dan was in het verhaal. Aan alle objecten in de databank kunnen tags en memo's worden gehecht, waardoor iets nadien kan worden opgepikt als de informatie bijvoorbeeld onvolledig was bij het aanmaken ervan. Bovendien is het met zo'n relationele databank makkelijk om het magazijn te reorganiseren of bestellingen bij te houden: er zijn overal ankers voorzien, waaraan extra functionaliteit of nieuwe workflows naar behoeven kunnen worden ingebouwd. Maar je hoeft niet alle functionaliteit zelf te bouwen: zoekengines op het web kunnen de objecten in de databank nauwkeurig adresseren en daardoor veel beter vindbaar maken. Ook voor onderzoek en onderwijs is het aantrekkelijker om dit type resources op het web te integreren in een research blog of een syllabuswiki.

Het VTi heeft uitdrukkelijk gekozen om de databank maximaal zichtbaar en op het web beschikbaar te maken. Maar toch staat niet alles in de databank op het web gepubliceerd. Pas na het inloggen komen er meer data vrij te consulteren, ook data waarop nog niet alle kwaliteitscontroles zijn afgewerkt: het gaat dan om heel recente gegevens, maar ook om gegevens uit de periode 1950-1993 bijvoorbeeld. Ook voor het bewaren van de kleine privénotities moet je inloggen, zodat her en der achtergelaten commentaren als geheugensteuntje bij het raadplegen van de databank bewaard kunnen worden, een beetje zoals post-its in een boek kleven als je ermee werkt. Meer nog: alle ingelogde gebruikers kunnen in het scenario van het VTi tags of trefwoorden toevoegen waar zij het relevant vinden: deze worden wel publiek zichtbaar. Je hoeft daarvoor geen nieuw account en paswoord aan te maken. Dankzij OpenID kan je je bestaande accounts van Myspace, Flickr of van verschillende webmaildiensten gebruiken.

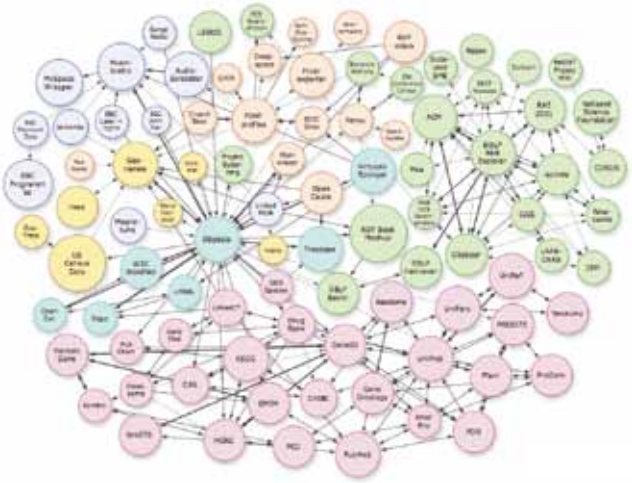
Linked Open Data

Linked Open Data is een wereldwijde beweging om de internethyperlink van het type 'http://' te gebruiken om databanken te koppelen die vroeger niet of alleen via specifieke protocols met elkaar gelinkt waren. Het is ook een concreet technologisch opzet om data met behulp van permanente URI's en RDF op het web te publiceren en uit te wisselen. Tim Berners-Lee heeft het idee in 2006 geformuleerd aan de hand van vier regels[5]:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Include links to other URIs. so that they can discover more things.

Deze regels zijn eigenlijk heel gelijkaardig aan de regels die gebruikt werden om het World Wide Web (WWW) te lanceren: toen kregen documenten vaste adressen en konden hyperlinks naar andere documenten toegevoegd worden als extra optie voor de gebruikers: omdat die regels relatief eenvoudig zijn, kon het WWW snel groeien en worden tot wat het nu is. Tim Berners-Lee probeert nu hetzelfde te bereiken voor data: als een organisatie de databank echt wil openen voor anderen, dan wijst deze aanpak een nieuwe weg. Het VTi is resoluut op die kar gesprongen: alle niet-privacygebonden data open en maximaal herbruikbaar.

Ondertussen zijn al een opmerkelijk aantal databanken op die manier open en gelinkt: Eurostat (statistieken over de EU lidstaten), DBpedia (een extract uit wikipedia), GeoNames (een globale geografische database), MusicBrainz (een databank met muziektracks en kunstenaars), Revyu (communitysite met besprekingen en kritieken over bijna alles), Project Gutenberg (literaire werken in fulltext), CIA World Factbook (landenfiches) enzovoort. De meeste hebben links met meerdere andere databanken, maar dat is niet noodzakelijk. Het VTi heeft al koppelingen met DBpedia gelegd, maar de automatische koppelingen leveren geen kant en klaar resultaat. Daarom is een traject gestart in het Europese netwerk van podiumkunstendocumentatiecentra, om hun databanken,waarin veel meer dezelfde namen voorkomen, via Linked Open Data met elkaar te verbinden.

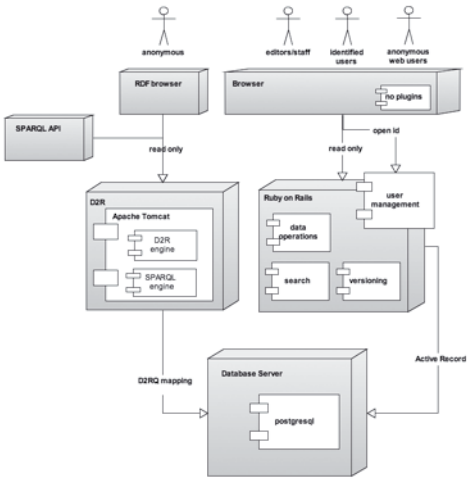


Linking Open Data cloud diagram
http://richard.cyganiak.de/2007/10/lod/

Architectuur

Het fundament van zo'n webapplicatie is een relationele databank. Bij het VTi en Contredanse is dat postgresql[6], maar het zou perfect ook een andere databank kunnen zijn. Daarop is een Ruby on Rails[7] webapplicatie geënt waarmee de databank kan worden aangevuld, aangepast enzovoort. Het is ook deze applicatie die ingelogde users extra mogelijkheden biedt (tagging, extra data, schrijfrechten, e.d.) en het is deze applicatie die als open source beschikbaar wordt gemaakt. Daarnaast is er ook een D2R server op de databank aangesloten, bestaande software[8] waarmee de data in de specifieke formaten voor het semantische web (RDF browsing, RDF dumps en SPARQL endpoint) wordt gepubliceerd.

Schematisch:



Om de realisatie van dit grote IT-project mogelijk te maken zijn verschillende technologieën tegen elkaar afgewogen. In de keuze voor het opensourcewebontwikkelingsplatform Ruby on Rails waren verschillende elementen belangrijk. Het is eerst en vooral een erg up-to-date technologie om interactieve websites te bouwen, waarrond een grote community van ontwikkelaars en webdesigners actief is. Componenten als tagging, versioning en login/authorisatie via OpenID[9] konden gewoon kant en klaar ingezet worden uit andere projecten. Er wordt heel streng onderscheid gemaakt tussen code die de vormgeving bepaalt en code die bijvoorbeeld interactie met de databank organiseert, zodat die code getest, beheerd en veranderd kan worden zonder wijzigingen (of foutmeldingen) in andere code te veroorzaken. Ook de meertaligheid was dankzij die opbouw (die MVC, 'Model-View-Controller' wordt genoemd) snel te realiseren.

Ruby on Rails is niet het enige MVC-gebaseerde webontwikkelingsplatform. In de meeste programmeertalen zijn er zulke platforms beschikbaar, bijvoorbeeld Struts of Cocoon in Java, CodeIgniter in PHP, Django of Zope in Python… Ruby on Rails was echter het eerste met het concept 'convention over configuration', een slimme vorm van automatisering van de softwareontwikkeling. Ruby on Rails neemt een heel groot aandeel van het aanmaken en aanpassen van webformulieren en webpagina's van de ontwikkelaars over. Net dit aspect maakte software op maat voor het VTi betaalbaar, omdat een OPAC al snel honderden types van pagina's telt.

Ruby on Rails heeft echter nog belangrijke voordelen op het vlak van de interactie met de databank. Alle informatie wordt via een gespecialiseerde component ('ActiveRecord') altijd veilig weggeschreven in de databank. Sommige webontwikkelingsplatformen en content management systemen denormaliseren (in de betekenis van E.F. Codd) relationele databanken om de informatie sneller op het web te kunnen tonen; langetermijn toegankelijkheid van de databank is dan niet zonder meer gegarandeerd. Ruby on Rails doet dat helemaal niet, waardoor de databank ook zonder de applicatie perfect gebruikt kan worden (bijvoorbeeld voor data mining). Ook als de applicatie aan vervanging toe is, is de onderliggende databank goed en begrijpelijk gestructureerd.

Dankzij de extra abstractielaag kan de data optimaal georganiseerd worden in relationele tabellen, gescheiden van alle code, die in de applicatie een plek heeft. Dat is een groot verschil met de klassieke IT-oplossingen voor documentaire informatie, waar zowel de applicatie als de onderliggende databanken georganiseerd zijn rond één standaard, in de bibliotheken meestal MARC21, waarop eventueel wel eens een lokale afwijking kan worden aangebracht, maar dat dan altijd ten koste van de conformiteit aan standaarden. Omdat bijvoorbeeld uitleningen of bestellingen niet in die descriptieve standaarden voorzien zijn, worden daarvoor dan uitbreidingen tegen die standaardmodellen aangebouwd met eigenaardige of sub-optimale databankstructuren tot gevolg. Er zijn voorbeelden hiervan te vinden in vele open source bibliotheek- en archiefsoftware en ook in de gesloten, commerciële software zal je zo'n kromme databankconstructies nog wel kunnen vinden. Meer nog, de applicatie die hier voor de documentatie van de podiumkunsten werd gebouwd, is dankzij de extra abstractielaag heel makkelijk over te dragen naar compleet andere werkterreinen: hout en bouw, fauna en flora…

Door de data onafhankelijk van een vooraf gekozen standaard op te slaan, wordt het bovendien makkelijker om die data nadien in meerdere standaarden weer te geven, zelfs in verschillende versies ervan. Dat is de taak van D2R, serversoftware ontwikkeld door C. Bizer aan de FU Berlin, waarin de informatie van een relationele databank wordt gepubliceerd als Resource Description Framework (RDF[10]), het dataformaat voor het semantic web. Als er in de toekomst nieuwe standaarden relevant worden of standaarden gewoon evolueren, kan die mapping telkens worden bijgestuurd en verfijnd. Dit is de filosofie van de gelaagde

metadata, zoals die in het project BOM-vl is ontwikkeld: het onderste laagje is de informatie in het lokale databankschema, daar bovenop komen laagjes van interoperabiliteit met andere bibliotheek- of archiefsoftware (in het geval van VTi via DCTERMS en MODS), laagjes voor interactie met online informatie over personen en organisaties (bij VTi via FOAF, omdat vele netwerk- en communitywebsites de links tussen de users als een ketting van 'friends of a friend' modelleren, waarvoor FOAF de standaard is).

Deze aanpak is uitvoerig gedocumenteerd, zowel in het recente BOMBOEK (Lannoo, 2010) als in een specifieke publicatie over dit metadatamodel (Universiteitsbibliotheek Gent, 2009). Ook in *Bibliotheek- & archiefgids* verscheen er in 2009 een samenvattend artikel[11] over deze aanpak. In het nu lopende IWT-project Archipel wordt dit metadatamodel gebruikt met het oog op langetermijn archivering, waarvoor opnieuw extra laagjes worden toegevoegd. (PREMIS, Provenance …)

Open Source

Tot slot: met de lancering van data.vti.be is het werk niet klaar en het is nog veel te vroeg om een streep te trekken en het project te evalueren. Het VTi heeft er immers voor gekozen om de nieuw ontwikkelde software en het grotere functionele geheel vrij te geven als opensourcesoftware, klaar voor andere gespecialiseerde collecties of doorgedreven documentatiecentra om ermee aan de slag te gaan. Een eerste stap was de keuze van Contredanse om de software over te nemen. Nu wordt er contact gezocht met anderen die de relevantie voor hun noden willen uittesten en ook met de applicatie aan de slag willen. Daartoe is de code gepubliceerd op Github als een nieuw, zelfstandig project onder de naam 'warburg'[12]. Het VTi werkt de volgende vijf jaar verder aan verfijning en uitbreiding van het systeem en het bedrijf 10 to1 is beschikbaar voor professionele dienstverlening. In de logica van opensourcesoftware is het echter perfect mogelijk om met andere ontwikkelaars aan de slag te gaan. Wie meer wil weten kan steeds bij de auteurs terecht.

Referentielijst

- *Rik Van de Walle e.a. (eds.), Bewaring en ontsluiting van multimediale data in Vlaanderen: Perspectieven op audiovisueel erfgoed in het digitale tijdperk, Lannoo Campus, 2010. (http://hdl.handle.net/1854/LU-1023867)*
- *Rik Van de Walle & Sylvia Van Peteghem (eds.), (Meta)datastandaarden voor digitale archieven, Universiteitsbibliotheek Gent, 2009. (hdl.handle.net/1854/LU-480734)*
- *Joris Janssens & Dries Moreels (eds.), Metamorfose in podiumland: een veldanalyse, VTi, 2007. (http://vti.be/nl/files/metamorfose-podiumland)*
- *Tom Evens & Dries Moreels (eds.), Access to Archives of Performing Arts Multimedia, VTi, 2009. (http://vti.be/nl/files/access-archives-performing-arts-multimedia)*

Noten

1. *http://www.openarchives.org*
2. *Zit artikel bouwt verder op een presentatie van Dries Moreels op Informatie aan Zee 2009 in Oostende, zie www.vvbad.be/node/4765.*
3. *http://www.vti.be*
4. *http://www.contredanse.org*
5. *http://www.w3.org/DesignIssues/LinkedData.html*
6. *http://www.postgresql.org*
7. *http://rubyonrails.org/*
8. *http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/*
9. *http://openid.net*
10. *http://www.w3.org/RDF*
11. *Van De Walle, Rik; Coppens, Sam en Mannens Eric, Een geslaagd semantisch metadatamodel voor langetermijnarchivering, in Bibliotheek- & archiefgids, (2009) 5; p. 17-21.*
12. *http://www.github.com/warburg/warburg*

SAMENVATTING

In dit artikel worden de overwegingen geschetst die bij het Vlaams Theater Instituut geleid hebben tot het opzetten van een grote vernieuwing van zijn IT (data.vti.be). Eerst en vooral werden commerciële databanktools vervangen door Postgresql (www.postgresql.org) en met Ruby on Rails (rubyonrails.org) werd een nieuwe webinterface gebouwd, zowel voor online presentatie als editing van de multidimensionele relationele metadata. Met behulp van D2R-server (http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/) werd de inhoud van de databank ook open gesteld voor hergebruik als Linked Open Data, een technische strategie uit het semantic web die het zinvol uitwisselen van data tussen computers vereenvoudigt. Tot slot besliste VTi om ook de ontwikkelde software als opensourceapplicatie beschikbaar te stellen via Github (www.github.com/warburg/warburg) en de mogelijkheden voor een breder draagvlak voor deze applicatie te onderzoeken.

ABSTRACT

This article discusses the considerations that led VTi (Flemish Theatre Institute) to a major overhaul of its IT (data.vti.be). First, commercial database tools were replaced with Postgresql (www.postgresql.org) and Ruby on Rails (rubyonrails.org) was used to build a new web interface, both for online editing and presentation of the multidimensional relational metadata. Using D2R Server (http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/), the database also opened up for reuse as Linked Open Data, a technical strategy from the semantic web to maximize the exchange of data between computers. VTi finally decided to licence the bespoke development as open source software and made it available through Github (www.github.com/warburg/warburg), in an effort to establish a broader user base for this application.