Koninklijke Bibliotheek

De Koninklijke Bibliotheek en Web 2.0: nieuwe gegevensarchitectuur maakt nieuwe concepten van dienstverlening mogelijk.

Auteurs:
Paul Doorenbosch, Koninklijke Bibliotheek
Theo van Veen, Koninklijke Bibliotheek

Thema:
Anticiperen op ontwikkelingen

Samenvatting

De Koninklijke Bibliotheek vernieuwt haar gegevensinfrastructuur. De basis daarvoor is een service georiënteerde architectuur met standaard protocollen en dataformaten. Deze architectuur moet het mogelijk maken om op eenvoudige en snelle wijze nieuwe diensten te realiseren met de data van de KB en die van anderen, om efficiënt gebruik te maken van services die op Internet voorhanden zijn en om het voor andere partijen eenvoudiger te maken diensten te ontwikkelen met KB-data. Uitgangspunt is dat alle data in de KB via één zoekactie snel beschikbaar zijn, dat (inter)nationale standaarden gebruikt worden en dat de communicatie webgebaseerd is.
De reorganisatie van de data-infrastructuur is een belangrijke stap om nieuwe vormen van webdienstverlening te kunnen ontwikkelen en faciliteren. Het ontwikkelen van webservices en het inzetten van door anderen beschikbaar gestelde webservices staan daarin centraal. Voorbeelden van dergelijke services zijn: zoeken, vertalen van zoektermen en resultaten; additionele informatie uit bronnen van derden ontvangen bij resultaten op het scherm; boeken, cd's e.d. lenen en kopen, relaties tussen data door de gebruiker laten vastleggen, social tagging. Veel van deze functionaliteit is ook nu al mogelijk, maar vraagt van de gebruiker om handmatig relevante functies te starten en zelf input te leveren. De mate van standaardisatie en openheid van de services bepaalt de mate waarin die processen (semi-)automatisch kunnen verlopen. Uitgangspunt blijft dat de gebruiker optimale regie krijgt en houdt over wat zich in zijn browser afspeelt om de resultaten die hij krijgt zo goed mogelijk bij zijn persoonlijke behoefte aan te laten sluiten.

Inleiding

Web 2.0[1] is al enige tijd hét concept dat Internet in haar ban heeft. Web 2.0 is geen revolutie, maar een stap verder in de steeds evoluerende internetwereld. Internet is in haar openbare jaren uitgegroeid tot het gigantische communicatienetwerk waarop iedereen zijn wetenschappelijke, hobbyistische of persoonlijke creativiteit kwijt kan. Het was tot voor kort echter meestal eenrichtingsverkeer: de aanbieder bepaalde. Web 2.0 kent vele definities en nuances. In een paar trefwoorden: onder eigen controle, interactie, sociaal, niet-exclusief. Een dergelijk democratisch concept is het Internet al sinds het aan de gemeenschap in handen is gegeven. Technisch was het misschien nog wat beperkt, het denken erover nog wat traditioneel en het gebruik nog wat teveel in handen van de voorlopers en de oude monopolisten. Internet had gewoon nog wat tijd nodig om te groeien en om mensen te laten wennen aan het idee van delen met en hergebruiken door anderen van eigen informatie.

Web 2.0 is het net van de gebruikers en niet langer alleen van de aanbieders. Bibliotheken spreken over Library 2.0. De daadwerkelijke omslag die je elders op Internet ziet gebeuren, is bij de bibliotheken nog lang niet gemaakt. Voor bibliotheken lijkt het omarmen van Web 2.0 meer een poging tegenwicht te bieden tegen het geweld van zoekmachines en softwaregiganten, de grote spelers op de informatiemarkt, dan het zoeken van eigen kracht. Leek het enige jaren geleden of de bibliotheken de positie van de uitgevers konden gaan herdefiniëren, nu lijken de aanbieders van zoekmachines de huidige bibliotheekfunctie als wegwijzer in de informatie te gaan overnemen. Google is de eerste zoekingang voor zowel de geïnteresseerde leek als de professionele gebruiker; in elk geval in de alfa- en gammawetenschappen en zelfs voor veel informatiespecialisten. Daarnaast kennen we ook nog de omvangrijke inspanningen die Google en Microsoft leveren om historische teksten digitaal te maken. Iedere monopolist is weliswaar een gevaar voor de vrije uitwisseling van informatie, maar moeten we ons ertegen verzetten? Nee, het Internet kent een bijna vanzelfsprekend groeiproces dat niet te stoppen is. Het heeft geen zin om dergelijke ontwikkelingen als bedreigingen te zien. Integendeel, we kunnen beter proberen er voordeel uit te halen, want nog nooit kreeg zo'n breed publiek makkelijke toegang tot informatie, waar ze vroeger niet eens het bestaan van wisten, laat staan dat ze er over konden beschikken. En dat was wat bibliotheken altijd voorstonden.

De grootschalige digitaliseringsprojecten die bibliotheken – al dan niet in samenwerking met IT-bedrijven - uitvoeren, houden de wetenschappelijke bibliotheken de komende tientallen jaren nog wel bezig. De look-and-feel van het gedrukte boek, de handschriften e.d. zal zeker nog geruime tijd een aantal bezoekers naar ons toetrekken. De nog steeds aanwassende stroom gedrukte publicaties moeten we toch minstens ergens in Nederland blijven bewaren. Als instelling bedoeld om informatie beschikbaar te stellen en als kennisautoriteit wordt de bibliotheek echter steeds marginaler. Waar wij het in moeten zoeken is dienstverlening: niet langer het object en de kennis centraal stellen, maar de gebruiker; gebruikers in groepen, in groepjes en individueel; mondige mensen op Internet die hun eigen keuzes bepalen en niet per se intermediairs nodig hebben; die steeds minder geloven in de persoon of de instelling als autoriteit. We moeten als openbare instellingen voor culturele en wetenschappelijke informatie opnieuw onze plaats vinden en daarbij beseffen, dat die plaats steeds zal en moet blijven veranderen.

Web 2.0: niet de oplossing voor dit vraagstuk, maar wel een concept waar binnen zich tal van nieuwe mogelijkheden aandienen.

Data-integratie

De KB heeft net als veel andere organisaties verschillende databases en catalogi die via verschillende websites worden aangeboden. Ze hebben vaak een eigen metadataformaat, eigen zoek- en retrievalsoftware en een op maat gemaakte website. Deze databases en catalogi zijn veelal niet integraal doorzoekbaar vanwege de verschillende dataformaten en de verschillende functionaliteit die specifiek is afgestemd op één toepassing.
Een achttal jaren geleden is de KB een traject gestart om daarin orde te brengen. Er werd een centrale XML-database gecreëerd, waarin metadatarecords werden opgeslagen. Weliswaar in een eigen KB-formaat, sterk geënt op het GGC-formaat, waar het merendeel van de records vandaan kwam, want een standaard XML-formaat was nog niet uitontwikkeld. In de loop der tijd zijn binnen deze XML-database steeds meer afwijkende formaten geïntroduceerd voor nieuwe diensten, en daarnaast werden ook – vaak uit praktisch oogpunt - toch weer nieuwe databases gebouwd, soms zelfs in een black-box-achtige structuur. Een nog niet voldoende ontwikkeld denken over uitwisselbaarheid en integratie was daar de oorzaak van. Projectdoelen hadden prioriteit boven het bouwen aan één gemeenschappelijke informatieruimte.
Wat binnen de KB speelt, speelt ook tussen de KB en andere organisaties: niet uitwisselbare data, hoe lang we ook al praten over interoperabiliteit. Metasearch is enige tijd gezien als dé mogelijkheid om verschillende databases in één keer af te zoeken en de data te combineren voor de gebruiker. Het resultaat is soms redelijk, maar vaak beperkt. Lange technische trajecten, beperkte performance, weinig selectiemogelijkheden. Zowel binnen organisaties als tussen organisaties speelt echter steeds meer de behoefte om diensten op verschillende manieren te kunnen integreren, bijvoorbeeld door de data uit de ene dienst te kunnen gebruiken als input voor een andere dienst. De gebruiker vindt zijn universele metasearch wel bij Google, integraal zoeken binnen gespecialiseerde gebieden samengesteld uit verschillende (deel)databases, vindt hij daar niet. Laat staan de mogelijkheid om data naar eigen inzicht te integreren in zijn eigen omgeving.
Om een flexibeler dienstverlening mogelijk te maken moest bij de KB een betere uitgangssituatie worden gecreëerd. Om data beter uitwisselbaar te maken en om grotere flexibiliteit te bieden in het aanbieden van nieuwe services is bij de KB besloten de gegevensinfrastructuur volledig te vernieuwen. Het uitgangspunt daarbij is dat zowel diensten binnen de KB als ook in de buitenwereld met een minimum aan a priori-kennis over de KB-infrastructuur daarvan gebruik kunnen maken. Het doel is dat met een minimum inspanning een maximum aan extra functionaliteit geboden kan worden, die de gebruiker naar eigen inzicht en onder eigen regie kan inrichten. De sleutelwoorden hierbij zijn standaardisatie, integratie en een service georiënteerde architectuur[2].

Standaarden

De eis dat er slechts een minimum aan a priori-kennis over de infrastructuur nodig zou moeten zijn, betekent vooral dat gebruik gemaakt moet worden van standaarden. Echter, waar officiële standaarden ontbreken of niet vrij algemeen aanvaard zijn, ontwikkelen we samen met een gemeenschap of desnoods intern standaarden die we op Internet publiceren en waarvoor we ons sterk zullen maken dat ze door anderen worden overgenomen.
De belangrijkste standaarden die de KB heeft geïmplementeerd zijn:

- Qualified Dublin Core[3]. Dit wordt gebruikt als het primaire formaat voor beschrijvende data. Hiermee wordt optimale uitwisseling met andere diensten bereikt. Ondanks de beperkingen van Dublin Core is er geen anders formaat dat generiek genoeg is om gebruikt te worden voor beschrijvingen in de diverse sectoren van cultuur en wetenschap. Dublin Core is bij de KB ook de standaard voor interne uitwisseling van beschrijvende data. Omdat Dublin Core niet altijd voldoende is voor specifieke functionaliteit is dit uitgebreid tot Dublin Core eXtended (zie verderop), waarbij de 'core', de standaard onveranderd blijft;
- SRU[4] (Search and Retrieval via URLs) voor zoeken en vinden. Deze standaard vervangt Z39.50 en is in tegenstelling tot Z39.50 gebaseerd op http. Dit maakt integratie met andere services eenvoudig. Ook intern gebruiken we SRU en geen native protocollen meer voor het zoeken in databases;
- OAI-PMH[5] (Open Archive Initiative Protocol for Metadata Harvesting) voor het vergaren van nieuwe data. Dit protocol wordt zowel intern als extern gebruikt om nieuwe of gewijzigde data van het ene systeem of systeemonderdeel over te halen naar het andere;
- MPEG21 DIDL[6] om de structuur van samengestelde objecten op te slaan en de locatie van de subobjecten vast te leggen.

Met deze standaarden zijn we in staat gegevens uit te wisselen met een minimum aan voorkennis, maar dit is nog niet voldoende voor een optimale dienstverlening. Ten eerste zijn er lokale uitbreidingen nodig op de standaarden. De toegepaste metadatavelden worden machineleesbaar gepubliceerd in een metadataregistry. Deze registry wordt bij voorkeur in internationaal verband opgezet. De metadataregistry van The European Library[7] lijkt daarvoor de meest aangewezen locatie.
Ten tweede moeten we aan de buitenwereld kenbaar maken wat we aan data en diensten kunnen aanbieden. Hiervoor worden de collecties beschreven volgens de NISO standaard voor collectiebeschrijvingen en bieden we een overzicht van beschikbare services. Voor het laagdrempelig integreren van beschikbare services wordt een nieuwe standaard voor servicebeschrijvingen ontwikkeld, omdat bestaande standaarden voor dit doel niet voldoen.

Persistentie van locaties wordt bereikt door gebruik te maken van een resolver. Deze resolver handelt alle locatiegebonden verzoeken af, waardoor wijzigingen in URL's slechts op één plaats hoeven te worden aangebracht, zonder dat de buitenwereld of andere interne systemen op de hoogte hoeven te zijn van die wijzigingen.

Multi-format metadata-opslag

Dublin Core heeft duidelijk nadelen t.o.v. formaten die specifiek voor een bepaald toepassingsgebied zijn ontwikkeld. Het heeft echter als voordeel dat het breed ondersteund wordt en onafhankelijk is van het toepassingsgebied. Om nu de voordelen van DC te behouden maar ook de mogelijkheid te hebben andere formaten te ondersteunen en applicatie-specifieke data te kunnen gebruiken, is besloten tot een database waarbij per record meerdere formaten mogelijk zijn en waarvan alleen een DCX (Dublin Core eXtended) -blok verplicht is. Dit DCX-blok bevat minimaal DC maar mag extra velden bevatten en wordt gebruikt voor indexering. Deze extensies worden in een metadataregistry gepubliceerd. Zoeken in de index levert default verkorte titels in Dublin Core. Hiermee wordt het mogelijk alle data integraal te doorzoeken en uniform te presenteren. Indien gewenst voor bepaalde functionaliteit kan echter gebruik gemaakt worden van andere beschikbare formaten. Er zijn hierbij geen beperkingen. Indien men voor een collectie een formaat nodig heeft dat niet standaard is dan kan dat toegevoegd worden, zolang er ook maar een DCX blok beschikbaar is. Uiteraard wordt niet geadviseerd om eigen formaten te gaan gebruiken, maar ook standaarden zijn eerst als niet-standaard begonnen. Met een weloverwogen gebruik van deze mogelijkheid houden we de voordelen van een standaard naast het voordeel van “geen beperkingen” opleggen.
Hoewel dit concept alle formaten mogelijk maakt is het vooral bedoeld voor specifieke toepassingen. Ten eerste voor het ondersteunen van andere standaard formaten zoals EAD, MARCXML, MODS etc. Een tweede reden voor het hebben van extra metadatablokken is het kunnen bewaren en op laten vragen van het originele record, indien dat geen DCX was. Een derde toepassing is het gebruik van een extra blok speciaal om specifieke indextermen te creëren, buiten de gepresenteerde metadata. Verder is het mogelijk om een blok met administratieve gegevens op te nemen.
Hoewel het mogelijk is data on-the-fly te converteren, is er voor gekozen, de data op te slaan in de formaten zoals die naar de client gestuurd worden. Hiermee wordt voorkomen dat complexe conversies de performance kunnen gaan beïnvloeden.

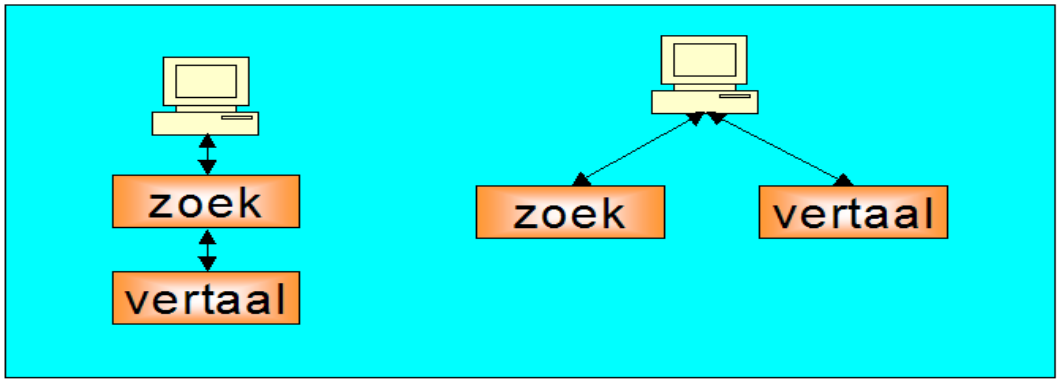
Metadataregistry

Essentieel voor het machine-leesbaar kunnen uitwisselen van metadata is het volgen van een standaard. Maar even essentieel voor verdere ontwikkeling is dat standaarden uitgebreid kunnen worden indien dat noodzakelijk is. Sommige standaarden zijn flexibel genoeg om lange tijd mee te gaan. Voor het metadatamodel is er een extra mechanisme nodig om uitbreidingen mogelijk te maken.
De filosofie achter het DCX concept is de volgende. Protocollen als SRU en OAI maken het mogelijk om verschillende dataformaten op te vragen. Nu zijn er vele formaten met Dublin Core plus iets extra’s. Om geen last te hebben van de wildgroei aan formaten die bijna hetzelfde zijn en omdat de meeste XML-applicaties de velden die ze niet kennen gewoon negeren is het handig om één naam te hebben voor al die verschillende DC varianten. Om nu de betekenis te kunnen achterhalen van onbekende velden en wildgroei aan nieuwe velden tegen te gaan maken we gebruik van een metadataregistry.
De metadataregistry is een overzicht van gebruikte metadatavelden met al hun karakteristieken. Hierin kunnen ook velden opgenomen worden die voorgesteld maar

nog niet geaccepteerd zijn. Indien een metadata-aanbieder de metadataregistry inspecteert alvorens nieuwe metadata te introduceren, wordt de kans op divergentie verkleind. Aanbieders van diensten kunnen de metadataregistry consulteren om eventueel de betekenis van nog onbekende metadatavelden te achterhalen. Om te zorgen dat de metadata van de KB zoveel mogelijk aansluiting vinden bij de metadata van vergelijkbare instellingen, liefst op Europees niveau, wordt hiervoor gebruik gemaakt van de metadataregistry van de European Library.

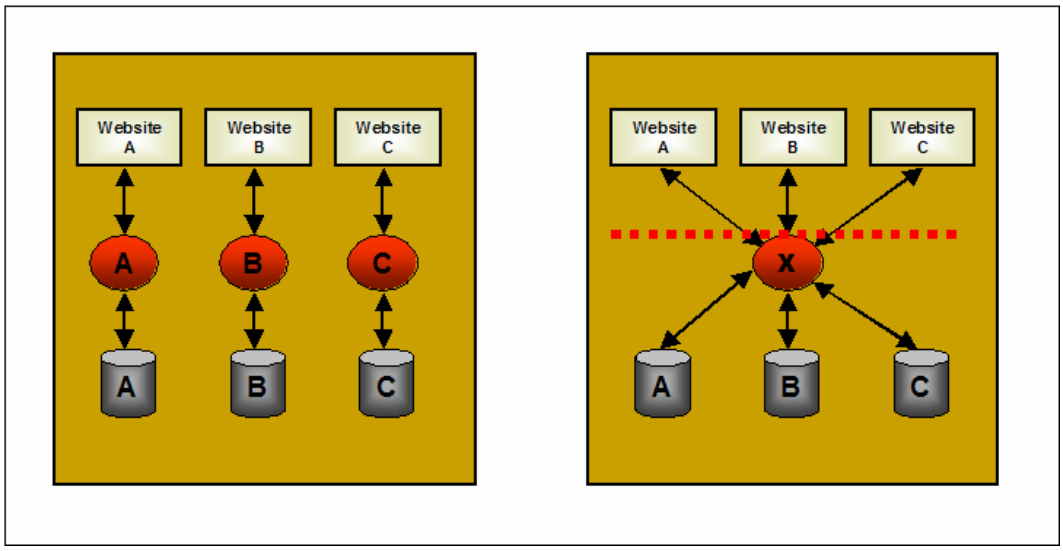
Het architectuurmodel

We gaan uit van gebruikersapplicaties die verschillende services benaderen. Deze applicaties kunnen op een server draaien of op het werkstation van de gebruiker. De huidige praktijk is veelal dat bij het benaderen van een dienst de integratiemogelijkheid met andere diensten bepaald is door de aanbieder van de dienst of de leverancier van de software. Dit is geschetst in het linker deel van



Figuur 1 Links voorbeeld van integratie die in de dienst is vastgelegd. Rechts is de integratie onder controle van de gebruikersapplicatie.

figuur 1 met als voorbeeld een zoekdienst en vertaaldienst. Rechts in figuur 1 zien we de situatie zoals we die graag willen hebben. De gebruikersapplicatie kan beide diensten benaderen en integratie vindt plaats in de gebruikersapplicatie, meestal op het werkstation. Deze integratie is onder controle van de gebruiker en de verschillende diensten hoeven er geen weet van te hebben dat hun output gebruikt wordt als input voor een andere dienst. Om dit mogelijk te maken is het nodig dat de client weet hoe de verschillende diensten benaderd moeten worden, d.w.z. hoe een verzoek geformuleerd wordt en hoe de output geïnterpreteerd moet worden. Het is dan ook wenselijk dat voor gelijksoortige diensten dit zoveel mogelijk hetzelfde is of aan dezelfde regels voldoet, bijvoorbeeld een bepaald protocol volgt.



Figuur 2 Links is de huidige praktijk weergegeven: elke website benadert een specifieke database via een eigen protocol. Rechts de situatie waarin alle webdiensten alle databases kunnen benaderen door hetzelfde protocol (X) te gebruiken.

In figuur 2 is weergegeven hoe de integratie bevorderd wordt als met verschillende diensten dezelfde taal kan worden gesproken. Nu zullen niet alle diensten via een standaard protocol benaderd kunnen worden. Om de potentiële integratie verder te vereenvoudigen is het wenselijk dat services zodanig beschreven worden dat een client m.b.v. een servicebeschrijving in staat is van een dergelijke service gebruik te maken.

Realisatie

Bij de implementatie van de nieuwe infrastructuur hebben we een generiek concept voor ogen gehad, maar we zijn gestart met de drie grootste collecties van de KB: de Algemene Catalogus, het e-Depot en Het Geheugen van Nederland. Streven is om in het voorjaar van 2007 alle collecties overgezet te hebben naar de nieuwe infrastructuur, zodat de oude definitief uitgezet kan worden.
Het proces dat er nu draait ziet er in grote trekken als volgt uit: centraal staat een publieksdatabase voor opslag van metadata in XML. Ieder record wordt vervaardigd in een werkomgeving (zoals catalogustitels in het GGC). Voor alle metadatabestanden waarvoor zo'n werkomgeving ontbreekt of niet meer adequaat is, realiseren wij een nieuwe werkdatabase met metadata-editor. Deze editor kan het te bewerken materiaal zowel uit de centrale opslag halen, als uit de werkdatabase, die qua structuur vrijwel identiek is aan de centrale database. In alle gevallen worden de records via een harvestingprotocol (OAI) opgehaald en geladen in de metadata-opslag. Deze records worden waar nodig geconverteerd naar Dublin Core, waarbij het originele dataformaat blijft bestaan. Per record worden de diverse formaten in een 'container' geplaatst en vervolgens geladen in de centrale database (publieksdatabase). Uiteraard bestaat er ook een mogelijkheid om batch-bestanden direct in de laaddirectory te plaatsen. De database wordt geïndexeerd, waarbij de Dublin Core-velden gedeeltelijk in de index worden opgeslagen om een snelle verkorte titelpresentatie te kunnen leveren. Voor de structuurbeschrijvingen volgens MPEG21 DIDL wordt hetzelfde proces gevolgd. Bij het opvragen worden de records uit de database gehaald en gepresenteerd. Daar waar het om samengestelde

objecten gaat wordt de structuurinformatie aangeboden. Vervolgens kan genavigeerd worden naar de daadwerkelijke objecten als die er zijn, of kan het resultaat gebruikt worden voor verdere acties. De objecten (afbeeldingen, full-text etc.) worden als losse files op een file-systeem opgeslagen. Om persistentie te bereiken loopt de toegang daartoe via een resolver. Het hele proces is in hoge mate modulair ingericht. In deze aanloopfase bewaren we op diverse plekken in het proces het dan gegenereerde materiaal om te voorkomen dat bij calamiteiten het proces van de grond af aan moet worden herstart. Op termijn zal daar een verdere vereenvoudiging in kunnen plaatsvinden.

De basisinfrastructuur die nu is opgeleverd, is gevuld met de drie collecties en zal verder uitgebreid worden met andere collecties. De gekozen principes zijn zo generiek dat bij het toevoegen van nieuwe collecties geen wezenlijke aanpassingen verwacht worden.
Uiteraard zijn we bereid onze keuzen en de hele uitwerking van de infrastructuur toe te lichten en bieden we de mogelijkheid, waar zinvol, om software te laten hergebruiken.
In de loop van het implementatietraject, dat zo'n anderhalf jaar heeft geduurd, zijn we tegen een hoop praktische problemen aangelopen. Wat echter misschien wel het meest lastige onderdeel is geweest, is de mapping van het GGC-formaat en de interne formaten naar Dublin Core. Hierin zal in de loop der tijd nog wel wat verfijning moeten worden aangebracht, maar de ontwikkelde conversiestylesheets zijn beschikbaar voor anderen.

Een laatste belangrijk onderwerp dat bij oplevering moet zijn geregeld, is het neerleggen van verantwoordelijkheden in de organisatie. Omdat de ontwikkelde infrastructuur gericht is op gebruikers, moet de eindverantwoordelijkheid voor doorontwikkeling ook primair een aangelegenheid zijn van de afdelingen die de gebruikersdiensten verzorgen, waarbij de IT-rol faciliterend is. Door de ver doorgevoerde scheiding tussen enerzijds data en functionaliteit en anderzijds techniek is dat ook haalbaar.

Wat hier is beschreven, is hoofdzakelijk het gereed maken van data voor integraal zoeken en presenteren, en de uitwisseling van data. Het is het rechttrekken van zaken die in het verleden uit elkaar zijn gaan lopen. Het is het bouwen van een gestandaardiseerde, stabiele, maar zeer flexibele data-repository. Daarmee zitten we nog steeds aan de aanbodzijde. Het uiteindelijke doel is een flexibele en op gebruikers toegesneden dienstverlening en een grote mate van zelfregie van die gebruikers over de dienstverlening.

Service-integratie

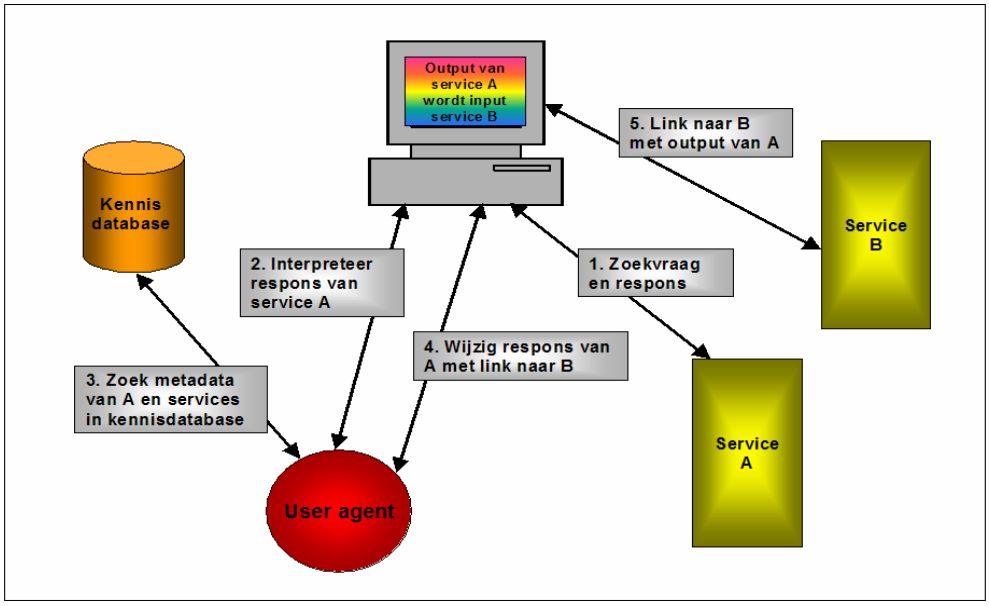
Data en diensten moeten met een minimum aan inspanning geïntegreerd kunnen worden met die van andere partijen. Met integratie wordt hier bedoeld dat de output van een webservice gebruikt kan worden als input voor andere webservices, liefst zonder enige wederzijdse afhankelijkheid en door de eindgebruiker zelf naar eigen inzicht in te richten. Een gebruiker wil bijvoorbeeld bij het doorzoeken van willekeurige bibliografische bestanden de mogelijkheid hebben om met een druk op de knop een gevonden boek in Amazon te bestellen. Een andere gebruiker wil

misschien de mogelijkheid hebben om de gesproken tekst van een in een database gevonden video in vertaalde ondertiteling om te zetten. Zo zijn er talrijke voorbeelden van services denkbaar die gebruikt kunnen worden voor de data die we op het web vinden. Nu zijn we meestal nog afhankelijk van de functionaliteit die een aanbieder van een dienst in de gebruikersinterface aanbiedt. Waar we naar toe willen is dat, als een gebruiker ergens een bruikbare service vindt, hij deze op een eenvoudige manier kan integreren met de zoekresultaten die verkregen zijn uit een willekeurige database. Om dit te realiseren is het naast standaardisatie van de toegang tot services ook nodig dat de services beschreven worden op een machine-leesbare manier. Door de gebruiker op een gebruikersvriendelijke manier in staat te stellen deze servicebeschrijvingen te manipuleren kan integratie door applicaties gerealiseerd worden zonder dat programmeervaardigheden van de gebruiker vereist worden.

Om het idee van integratie van services te verduidelijken gaan we uit van een situatie met de volgende componenten:

- een willekeurige webpagina met een zoekresultaat
- een stukje programmatuur (user agent) dat in staat is de webpagina te interpreteren en eventueel te modificeren. Dit kan een uitbreiding zijn van een portalapplicatie of een uitbreiding van de browser. De internetbrowser Firefox biedt faciliteiten voor dit soort uitbreidingen,
- services. Dit kan elke webapplicatie zijn die door een URL aangesproken wordt
- een kennisdatabase met beschrijvingen van relevante services. Deze servicebeschrijvingen bevatten o.a. informatie over hoe een service aangesproken wordt en welke criteria aanleiding zouden moeten geven deze service aan de gebruiker aan te bieden. Zo’n criterium kan zijn het voorkomen van een bepaald metadataveld in een zoekresultaat. Bijvoorbeeld het aanwezig zijn van een ISBN kan aanleiding zijn om een link naar Amazon aan te bieden.

Het scenario voor integratie is weergegeven in figuur 3 waarin de nummers de opeenvolgende stappen aangeven. De gebruiker komt op een webpagina van dienst A. De user agent interpreteert de webpagina en vindt daarin bepaalde gegevens, bijvoorbeeld een auteursnaam. De user agent controleert in de kennisdatabase welke services er op grond van het voorkomen van het veld auteursnaam aan de gebruiker moeten worden aangeboden en vindt service B. Vervolgens verandert de user agent de auteursnaam in een link naar de service B met de auteursnaam ingevuld op de juiste positie in deze link. De gebruiker heeft nu de mogelijkheid naar service B door te klikken met de auteursnaam automatisch in de URL ingevuld. De user agent heeft nu twee van elkaar onafhankelijke services geïntegreerd op het niveau van de presentatie, dus zonder dat er ingrepen in de services zelf nodig waren. Indien de gebruiker ook invloed heeft op de kennisdatabase, kan deze integratie geheel onder zijn controle gebracht worden.

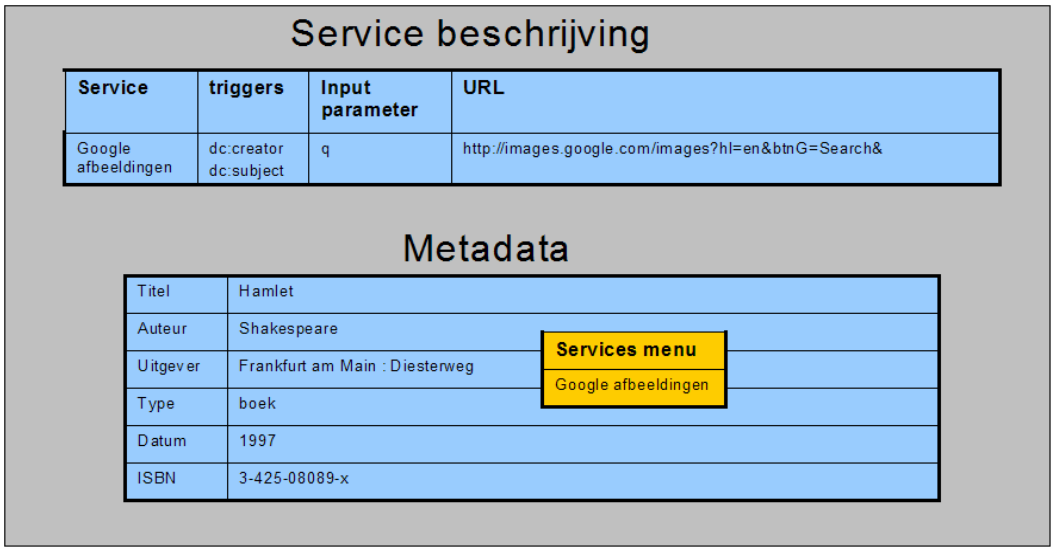


Figuur 3. Schematische weergave van de stappen om m.b.v. servicebeschrijvingen twee services te integreren.

We kunnen dit illustreren aan de hand van een werkende demo[8]. Hierbij wordt gebruik gemaakt van een portal die gebaseerd is op Ajax-technologie en via het SRU protocol gelijktijdig in meerdere databases kan zoeken. Ajax staat voor Asynchronous, Javascript en XML en wil zeggen dat vanuit een webpagina d.m.v. Javascript XML-pagina's opgevraagd kunnen worden bij verschillende servers. De resultaten kunnen verwerkt worden zodra deze ontvangen zijn zonder dat het scherm bevriest tijdens het wachten op de nog niet ontvangen resultaten. Het Ajax-concept leent zich uitstekend voor het integreren van services omdat het benaderen daarvan op de achtergrond kan gebeuren en er pas gebruikersinteractie nodig is als de repons van een service daartoe aanleiding geeft. In deze demo wordt het Ajax concept nu overigens nog alleen gebruikt voor het zoeken maar zal uitgebreid worden tot alle services die hiervoor geschikt zijn, d.w.z output in XML leveren.
Met deze demonstratieportal wordt nu gezocht in bibliografische en tekstbestanden. De zoekresultaten zijn metadata met Dublin Core in XML. Een XML-file met servicebeschrijvingen dient als kennisdatabase. Voor elke service wordt hierin aangegeven welke Dublin Core metadatavelden in de respons aanleiding moeten geven voor het aanbieden van deze service. Verder is het adres van de service en de bijbehorende URL-syntax erin vastgelegd. Een voorbeeld is de beschrijving van de bestelservice bij Amazon. In deze beschrijving is vastgelegd dat als een zoekresultaat een ISBN bevat er een link naar Amazon geboden moet worden. Ook is vastgelegd wat de URL is en waar de ISBN ingevuld moet worden.

Een ander voorbeeld is het zoeken van afbeeldingen in Google. Deze service kan bijvoorbeeld aangeboden worden op basis van het voorkomen van het auteursveld of een onderwerpsveld. In de beschrijving wordt dan aangegeven waar de auteursnaam of het onderwerpsveld in de Google URL ingevuld moet worden. Het concept wordt voor dit voorbeeld in sterk vereenvoudigde vorm geïllustreerd in figuur 4.
De demonstratieportal biedt de gebruiker de mogelijkheid een eigen file met servicebeschrijvingen te laden zodat de gebruiker vrij is in de keuze van services.

Het moge duidelijk zijn dat het concept service hier ruim genomen wordt en niet beperkt is tot bijvoorbeeld Webservices gebaseerd op SOAP[9].



Figuur 4 Schematische weergave van de relatie tussen servicebeschrijving en het presenteren van metadata. Omdat in dit voorbeeld dc:creator in de servicebeschrijving als trigger genoemd wordt, leidt het aanklikken van de auteur (is dc:creator) tot het presenteren van een services menu bij de auteur.

Het hierboven beschreven concept kan gezien worden als een vervolg op het OpenURL[10] -concept. OpenURL-servers bieden meestal een pagina met contextafhankelijke links aan op basis van een leveranciersspecifieke kennisdatabase, waarin is vastgelegd hoe en wanneer een link naar een specifieke service wordt geboden. In het hier geschetste concept wordt deze pagina overgeslagen en wordt vanuit elk metadataveld een directe dynamische en contextafhankelijke link aangeboden. Bovendien is de leveranciersspecifieke kennisdatabase vervangen door XML-servicebeschrijvingen die voor zowel personen als applicaties toegankelijk zijn.

Door het standaardiseren van de servicebeschrijvingen zijn deze niet gekoppeld aan een specifieke toepassing en wordt het mogelijk ze te publiceren en onderling uit te wisselen. Zodoende worden ze voor applicaties van derden bruikbaar. Instellingen kunnen hun services op een standaard manier publiceren, bijvoorbeeld via een link met een vaste filenaam. Met niet al te ingewikkelde applicaties kunnen deze servicebeschrijvingen geïnspecteerd worden en kunnen gebruikers interessante services toevoegen aan een persoonlijke database met servicebeschrijvingen. Deze database kan dan doorgegeven worden aan applicaties die dit concept geïmplementeerd hebben.

Een uitbreiding, die zorgt dat het concept niet beperkt blijft tot gebruik binnen een portal, is om webpagina's van zogenaamde "microformats" te voorzien.[11] Dit is een bekend concept bij OpenURL in de vorm van COinS (Context Object in Span).[12] Hierbij wordt in de HTML in een zogenaamde "span" alle relevante context informatie (bijv. auteur, ISBN etc.) vastgelegd. Een browser extensie (user agent) kan deze informatie in de HTML detecteren en op basis daarvan de pagina modificeren en

nieuwe links aanbieden. In sterk vereenvoudigde vorm kan men zich voorstellen dat een auteur in een webpagina weergegeven wordt als:

```
Dit werk is geschreven door <span title=”dc:creator”>Shakespeare</span>
```

Indien hier niets mee gebeurt, dan ziet de gebruiker gewoon de tekst:

```
Dit werk is geschreven door Shakespeare
```

Met behulp van een browser extensie kunnen dit soort "spans" opgezocht worden in de webpagina en op basis van de servicebeschrijvingen, in dit voorbeeld voor “Google afbeeldingen”, veranderd worden in een link naar “Google afbeeldingen” met “Shakespeare” als parameter:

```
Dit werk is geschreven door Shakespeare
```

Klikken op deze link levert dan de afbeeldingen van Shakespeare. Geavanceerder toepassing is mogelijk door een link niet alleen aan te bieden op basis van het voorkomen van een enkel metadataveld, maar op een combinatie ervan of zelfs als reactie op het ontbreken van resultaten. Uiteraard hoeft de toepassing zich niet te beperken tot het aanbieden van één service. Waar zinvol kan het leiden tot een aaneenschakeling van services. Kortom, door beschikbare services op de juiste wijze te beschrijven kunnen intelligente applicaties hiermee hun weg op het Internet vinden en datgene automatisch en eventueel op de achtergrond doen wat de gebruiker anders zelf zou moeten doen, of niet zou doen vanwege de hoeveelheid werk. De toegevoegde waarde voor de gebruiker kan dus enorm groot zijn.

Tot slot

In bibliotheken komt standaardisatie, uitwisselbaarheid en open content steeds meer in het beleid tot uiting. Dat geldt zowel nationaal als internationaal. The European Library, een samenwerking van nationale bibliotheken, gehuisvest in de KB, vervult daar op Europees niveau een belangrijke voortrekkersrol bij. Met een aantal projecten dat op stapel staat, wordt die rol verder geïntensiveerd. Bij archieven en bibliotheken moet het denken hierover nog verder ontwikkeld worden. Op internationaal niveau wordt met de komst van The European Digital Library – een Europees grootschalig programma, waarvan de coördinatie ook in de KB zal plaatsvinden – aan verdere ontwikkeling van de in dit artikel beschreven ideeën gewerkt. Uiteindelijk is het doel één Europese culturele en wetenschappelijke informatieruimte. De KB zal hier in samenwerking met nationale partijen en binnen het TEL/EDL-kader haar steentje aan bijdragen door de data optimaal uitwisselbaar te maken en door in wisselwerking met de gebruiker allerlei services te ontwikkelen die de gebruiker in zijn interactie met informatie kunnen ondersteunen.

Paul Doorenbosch, Theo van Veen

augustus 2006

Referenties

[1] Tim O'Reilly, *What is Web 2.0*,
http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] Over de uitgangspunten die gehanteerd zijn bij de vernieuwing van de gegevensarchitectuur verschenen artikelen in Dlib [Theo van Veen*, Renewing the Information Infrastructure of the Koninklijke Bibliotheek,*
http://www.dlib.org/dlib/march05/vanveen/03vanveen.html
en Ariadne [Theo van Veen, *Serving Services for web 2.0,*
http://www.ariadne.ac.uk/issue47/vanveen
Alle informatie over het data-architectuurproject in de KB is te vinden op:
http://research.kb.nl/vga

[3] DC/Qualified Dublin Core/Dublin Core (Dublin Core Initiative), http://dublincore.org

[4] SRU/Search and Retrieval via URL, http://www.loc.gov/standards/sru/

[5] OAI-PMH/The Open Archives Initiative Protocol for Metadata Harvesting,
http://www.openarchives.org/OAI/openarchivesprotocol.html

[6] MPEG21 DIDL: http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

[7] The European Library portal, http://www.theeuropeanlibrary.org/

[8] De demoportal is te vinden op: http://research.kb.nl/sruportal

[9] SOAP/Simple Object Access Protocol, http://en.wikipedia.org/wiki/SOAP

[10] The OpenURL Framework for Context-Sensitive Services,
http://www.niso.org/committees/committee_ax.html

[11] Microformats, http://en.wikipedia.org/wiki/Microformats

[12] OpenURL COinS (A Convention to Embed Bibliographic Metadata in HTML),
http://ocoins.info/

Over de auteurs

Paul Doorenbosch (*1948) is hoofd van de afdeling Product- en Dienstontwikkeling van de Koninklijke Bibliotheek (Research & Development).
Hij heeft Nederlandse Taal- en Letterkunde gestudeerd aan de Universiteit van Amsterdam en was daarna werkzaam als wetenschappelijk bibliograaf / eindredacteur bij het Bureau voor de Bibliografie van de Neerlandistiek van de KNAW. Van 1996 tot 2001 was hij afdelingshoofd Wetenschappelijke Informatie en Collectievorming (neerlandistiek, sociale wetenschappen en biomedische wetenschappen) bij het Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI).
Vanaf 2001 projectmanager bij de hoofdafdeling Research & Development van de Koninklijke Bibliotheek, verantwoordelijk voor het opzetten en uitvoeren van Het Geheugen van Nederland, nationaal digitaliseringsprogramma voor het cultureel erfgoed ten behoeve van een breed publiek en van het voortgezet onderwijs. Vanaf 2005 projectleider van de Vernieuwing van de Gegevensarchitectuur in de KB en sinds 2006 hoofd van de afdeling Product- en Dienstontwikkeling, waar dit project ook is ondergebracht.
Hij is tevens vice-voorzitter van het Dagelijks Bestuur van CATCH (Continuous Access to Cultural Heritage) NWO-programma voor ICT-kennisinstellingen en Cultureel Erfgoed).
Contact: paul.doorenbosch@kb.nl, Koninklijke Bibliotheek, Postbus 90407, 2509 LK, Den Haag, 070-3140161

Theo van Veen (*1953) is projectadviseur bij de afdeling Product- en Dienstontwikkeling van de Koninklijke Bibliotheek (Research & Development).
Hij heeft Natuurkunde gestudeerd aan de Technische Universiteit Delft. Na werkzaam te zijn geweest in de psychofysica en vervolgens in de procesindustrie is hij in 1988 bij de Universiteit Bibliotheek van de Rijksuniversiteit Utrecht als hoofd automatisering zich met bibliotheekautomatisering gaan bezighouden. Sinds 1998 is hij in dienst bij de Koninklijke Bibliotheek. Na bijgedragen te hebben aan het tot stand komen van de European Library is hij momenteel werkzaam als projectadviseur bij het project Vernieuwing van de Gegevens Architectuur van de KB. Binnen dit project houdt hij zich o.a. bezig met standaarden en ontwikkelingen op het gebied van Web 2.0.
Contact: theo.vanveen@kb.nl, Koninklijke Bibliotheek, Postbus 90407, 2509 LK Den Haag, 070-3140658