De bouwstenen van de digitale bibliotheek

DEN – Marco de Niet

Dit artikel is geschreven door Marco de Niet en gepubliceerd in De Digitale Bibliotheek. Red. Bart van der Meij en Kees Westerkamp. Rotterdam, Essentials/NVB, 2007], p. 67-85. Gepresenteerd tijdens het congres 'De digitale bibliotheek' op 13 juni 2007.

INLEIDING

Al ver voor de term 'digitale bibliotheek' in gebruik werd genomen, investeerden bibliotheken fors in automatisering. Bijna veertig jaar geleden nam het gecomputeriseerd catalogiseren van bibliotheekcollecties een aanvang, en nog voor de personal computer en het world wide web hun grote opmars maakten waren vele Nederlanders al vertrouwd met zoeken naar informatie via terminals in bibliotheken.

De focus bij die eerste automatiseringsgolf in bibliotheken lag bij het ondersteunen van hoofdtaken van de bibliotheek, zoals collectioneren, bewaren, vindbaar maken en uitlenen van objecten in de eigen collectie. Wel werden al vrij snel, dankzij bibliotheeknetwerken als Pica, de collecties van andere bibliotheken toegankelijk gemaakt voor de eigen gebruikers.
Pas met de komst van het web werd duidelijk dat automatisering niet per se dienend hoeft te zijn aan die traditionele taken. Hoewel het een gezond standpunt is om de techniek niet leidend te laten zijn, zien we ook dat technologische vernieuwingen ideeën aandragen voor nieuwe vormen van informatieverzorging en dienstverlening aan nieuwe doelgroepen. Met name de mogelijkheden om publicaties full text aan te kunnen bieden, naast of in plaats van bibliografische beschrijvingen, zorgen ervoor dat bibliotheken inmiddels anders aankijken tegen het toegang verschaffen tot hun collecties.

Het web wordt wel eens negatief omschreven als een 'technology soup', maar het heeft ook een mondiale centrifugale vernieuwingskracht in de hand gewerkt waar bibliotheken zeer van (kunnen) profiteren. Veel traditionele bibliotheektechnologieën zijn in de afgelopen jaren zo omgebouwd dat ze aansluiten op het web en feitelijk beter dan ooit tot hun recht komen. De kennis die bibliotheken hebben opgebouwd over databasebeheer, indexering en toegang bieden geeft ze een leidende rol bij het opbouwen van erfgoedbrede (internationale) zoeksystemen (i2010: Digital libraries, 2006).

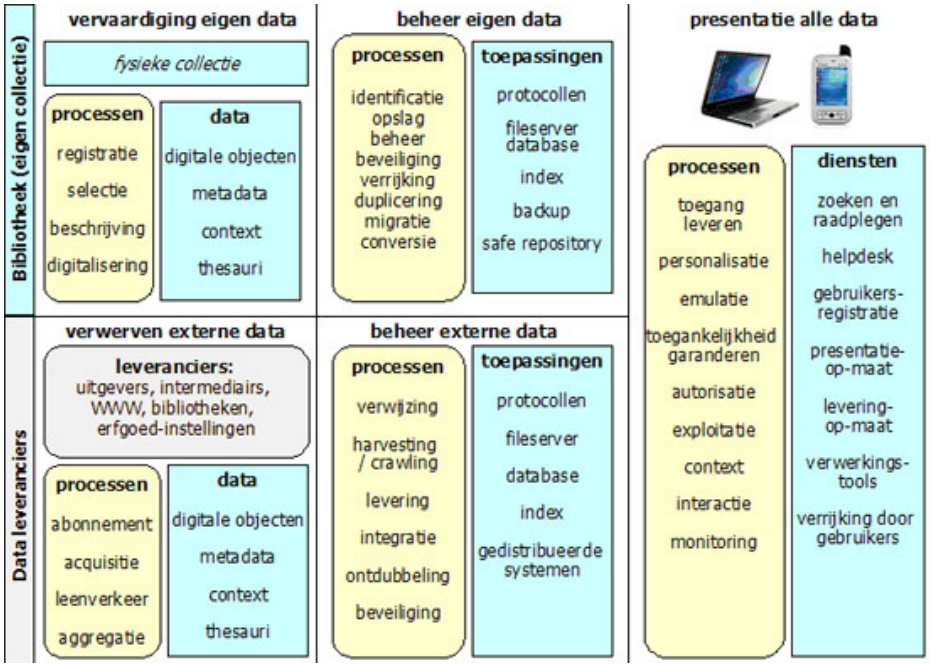
Op internationaal niveau wordt zwaar geïnvesteerd (zowel geld als menskracht) in de verdere ontwikkeling van het World Wide Web als platform voor uitwisseling van informatie en kennis. Ook de grote spelers in de ICT-wereld dragen bij aan zowel vernieuwing als standaardisatie van informatie- en kennisuitwisseling op basis van webtechnologie. Het is alleen al vanuit technologisch standpunt evident dat de verdere ontwikkeling van het web de blauwdruk biedt voor de technische infrastructuur van de bibliotheek van de toekomst. Daar komt bij dat het web de informatievaardigheden en -verwachtingen van de mondiale bevolking zo beïnvloedt, dat

bibliotheken het zich simpelweg niet kunnen permitteren een geheel eigen koers te varen, buiten het web om, omdat ze dan hun publieke functie in gevaar brengen.

Centraal bij het web, en dus ook bij de digitale bibliotheek, staat wereldwijde beschikbaarheid van informatie. Een belangrijk concept hierachter om die beschikbaarheid te realiseren wordt uitgedrukt met het begrip interoperabiliteit, de eigenschap van hard- en software om data op een gestandaardiseerde wijze uit te wisselen via netwerken. De internationale gemeenschap is vernetwerkt, en zeker voor kennisinstellingen en -dienstverleners is open uitwisseling van data een kernactiviteit geworden. Een informatiedienst van een bibliotheek mag geen 'gesloten doos' meer zijn, waar alleen een handjevol beheerders en gebruikers bij kunnen (bijvoorbeeld een op maat gemaakte database op een 'stand alone' PC in een leeszaal). Tijd- en plaatsonafhankelijke toegang, gebruiksvriendelijke zoek- en feedbackmogelijkheden en snelle levering van relevante informatie zijn een basisverwachting van de gebruikers van een bibliotheek geworden. Het vernieuwen van bestaande informatiediensten en het bouwen van nieuwe diensten dienen dan ook interoperabiliteit als uitgangspunt te nemen.

Hieronder worden enkele ontwikkelingen besproken die de interoperabiliteit van de digitale bibliotheek versterken. Onder 'digitale bibliotheek' wordt in dit artikel de bijdrage bedoeld van een individuele instelling aan het grotere geheel van informatievoorziening, op lokaal, nationaal of internationaal niveau. Die bijdrage kan gebaseerd zijn op eigen collecties, maar ook op het bij elkaar brengen van informatie die zich buiten de instelling bevindt, bijvoorbeeld bij uitgevers, andere bibliotheken of het web.

De ontwikkelingen worden geclusterd aan de hand van enkele belangrijke fasen uit de 'levenscyclus' van digitale data, namelijk vervaardiging, beheer en presentatie (ICT-register voor het cultureel erfgoed, 2007). In onderstaand schema zijn deze drie fasen in beeld gebracht. In de rechthoeken worden de concrete bouwstenen van de digitale bibliotheek genoemd. In de afgekante vlakken worden de processen benoemd die een bibliotheek kan uitvoeren in relatie tot die bouwstenen. Het schema is niet uitputtend bedoeld, en wordt ook niet uitputtend besproken, maar dient als een globaal overzicht van de mogelijkheden tot versterking van beschikbaarstelling en uitwisseling van informatie in het algemeen, en interoperabiliteit in het bijzonder.



VERVAARDIGING VAN DIGITALE DATA

Er zijn vier soorten data te onderscheiden waar bezoekers van een digitale bibliotheek mee te maken kunnen krijgen. Uiteraard de digitale objecten zelf (dit kunnen tekstuele documenten zijn, maar ook beeld- of audiovisuele bestanden of databases), beschrijvingen van die objecten (de metadata), contextuele informatie over de objecten (bijvoorbeeld in educatieve toepassingen) en een abstractie van kennis over de objecten in 'gecontroleerde vocabulaires' zoals thesauri en classificaties. In alle gevallen betreft het digitale informatie die de bibliotheek zelf bezit of vervaardigd heeft, maar zeker ook data van derden die elders is opgeslagen. Een digitale bibliotheek is niet simpelweg een digitale kopie van de eigen fysieke collectie, aangevuld met documenten die alleen in digitale vorm aangeschaft kunnen worden. Het toegankelijk maken van aan de eigen collectie gerelateerde informatie elders is ook een belangrijke functie van de digitale bibliotheek. Deze ontwikkeling wordt 'van collectie naar connectie' genoemd.

Semantische interoperabiliteit

Het vervaardigen van metadata en het onderhouden van thesauri is uiteraard niets nieuws voor bibliotheken. De meesten beschikken reeds lang over een ICT-infrastructuur om hun bezit digitaal te registeren, te ontsluiten en aan hun publiek aan te bieden. Dat kan variëren van een eenvoudige desktop database tot geavanceerde bibliotheeksystemen. In het afgelopen decennium is veel van deze bibliotheeksoftware door de leveranciers of ontwikkelaars voor het web geschikt gemaakt. Ontwikkelingen op dit gebied zitten dan ook niet zozeer aan de kant van de ‘harde’ ICT, maar bij de toepassingsmogelijkheden.

Om metadata uit te kunnen wisselen, zijn er afspraken nodig over catalogisering. De ervaring leert dat het bijna onmogelijk is om bibliotheken massaal over te laten stappen op een gemeenschappelijk metadatamodel, om die uitwisseling efficiënt te laten verlopen. De investeringen in kennis en infrastructuur die hiervoor gevraagd worden, blijken in de praktijk onoverkomelijk. Ook hebben de meeste bibliotheken moeite met het afstand doen van traditionele praktijken en eerdere investeringen. De oplossing is dat iedere bibliotheek kan blijven catalogiseren in het datamodel en software naar keuze, maar dat interoperabiliteit tot stand komt door gebruik te maken van een gemeenschappelijke tussenlaag, die bijvoorbeeld via dataconversie tot stand komt. Momenteel is het meest gebruikte datamodel voor die tussenlaag de Dublin Core Metadata Element Set, kortweg Dublin Core. Deze set bestaat in zijn meest simpele vorm uit 15 vaste velden, zoals Titel, Uitgever en Datum. Er zijn diverse mogelijkheden om deze set uit te breiden of te verfijnen. Deze aanpassingen dienen wel volgens van te voren vastgelegde regels toegepast te worden, zodat zoeksystemen wereldwijd ermee overweg kunnen (Dublin Core in samenwerkingsprojecten, 2006).

Dublin Core is een recht-toe-recht-aan datamodel, met een traditionele veldindeling die voor gebruikers vrij gemakkelijk te interpreteren is. Er zijn echter ook nieuwe datamodellen ontwikkeld die een nieuwe kijk geven op interoperabiliteit van metadata. Een van de belangrijkste ontwikkelingen op het web is de uitbreiding die het ‘semantisch web’ of web 3.0 wordt genoemd. Het web bestaat in zijn huidige vorm vooral uit digitale documenten die door de makers van websites aan elkaar gelinkt worden. Via zoekmachines zijn deze documenten vindbaar op woorden die in die documenten voorkomen (full text indexing). Maar in Google, Yahoo of MSN kun je niet aangeven of een zoekterm bedoeld is als persoonsnaam (‘cats’) of plaatsnaam (‘leiden’) en je kunt ook geen synoniemen uitsluiten (‘ontwikkeling’ als onderdeel van een fotografisch proces en niet als technische innovatie). Het semantisch web moet ervoor zorgen dat semantische verbanden tussen documenten wel en volledig geautomatiseerd gelegd

kunnen worden (‘geef mij alle documenten over de familie Cats voor zover woonachtig in de gemeente Leiden’). Thesauri en classificaties spelen hierbij een cruciale rol, maar ook de wijze waarop de metadata gestructureerd zijn, is van groot belang. In de erfgoedsector zijn inmiddels enkele datamodellen ontwikkeld om de ‘semantische ruimte’ om een object heen weer te geven.

In de museale wereld is het CIDOC Conceptual Reference Model (CIDOC-CRM) ontwikkeld (CIDOC, 2007). Dit model maakt het mogelijk om bijvoorbeeld een mummie en een opgezette eend in één zoekactie te vinden door de verwantschap tussen eigenschappen (beide objecten hebben een conserveringsbehandeling na overlijden ondergaan). Dergelijke semantische verbanden kunnen te complex of te weinig specifiek zijn om met behulp van een thesaurus te leggen, of ze kunnen niet gelegd worden omdat de objecten voorkomen in verschillende datasets die niet dezelfde thesaurus gebruiken. Momenteel wordt in diverse internationale projecten onderzocht of CIDOC-CRM de uitwisseling van data tussen instellingen uit verschillende erfgoedsectoren kan ondersteunen.

In de bibliotheeksector is al bijna tien jaar geleden een voorstel gedaan om op conceptueel niveau om te gaan met bibliografische beschrijvingen. IFLA publiceerde in 1998 de 'Functional Requirements for Bibliographic Records' (FRBR) (IFLANET, 2007). Hierin wordt uiteengezet hoe een publicatie op basis van conceptuele relaties beschreven kan worden. Dankzij FRBR zou het veel eenvoudiger worden om bijvoorbeeld vertalingen, bewerkingen, liedjes, toneelopvoeringen en/of films die teruggaan op dezelfde tekst bij elkaar te vinden. In Nederland is er nauwelijks aandacht voor deze ontwikkeling. Dat is jammer, omdat ze de vele miljoenen bibliografische beschrijvingen die de bibliotheken inmiddels gemaakt hebben een nieuwe toekomst bieden. Conceptuele modellen als FRBR doorbreken de 1-op-1 relatie tussen bibliografische beschrijving en full text-object, en geven de beschrijvingen een nieuwe functie, namelijk als verbinding tussen diverse objecten onderling en tussen objecten en contextuele informatie. Op deze wijze zouden bibliografische beschrijvingen een nuttige rol kunnen blijven spelen bij het vinden van full text documenten.

Modellering van digitale teksten

Bibliotheken kunnen bij de vervaardiging van metadata en thesauri terugvallen op een lange traditie. Wat het vervaardigen van de digitale objecten zelf betreft, kun je gerust spreken over een stille revolutie. Waren bibliotheken nog niet zo lang geleden collectionerende instellingen die alleen hun bibliografische beschrijvingen digitaal beschikbaar stelden, sinds enige jaren zijn ze belangrijke producenten van digitale ‘content’ geworden. Door het voorhanden hebben van grote, gedurende vele jaren verzamelde collecties vinden bibliotheken, vooral de bibliotheken met bijzondere collecties, het bijna vanzelfsprekend dat zij degenen zijn die ervoor moeten zorgen dat eerder vastgelegde kennis via internet beschikbaar blijft. Deze wil om historische kennis door te geven aan nieuwe generaties is niet alleen lovenswaardig maar ook broodnodig. Het is zeker realistisch te verwachten dat op internet een situatie zal ontstaan die vergelijkbaar is met de verhouding tussen de online publiekscatalogus en de kaartcatalogus: wat niet in de OPC vindbaar is, bestaat niet meer voor de gemiddelde bibliotheekbezoeker.

Enkele grote bibliotheken hebben digitaliseringstraten ingericht om hun fysieke collecties via nieuwe media aan te bieden en bouwen op die wijze kennis op over de mogelijkheden van digitalisering. Ook als je als kleinere bibliotheek de hulp inroept van een gespecialiseerd bedrijf, is minimale kennis over modellering van digitale teksten en gebruik van documentformaten

nodig, al was het maar om een kwaliteitscontrole uit te kunnen oefenen en het nut van het eindresultaat voor je gebruikers zeker te stellen.

Digitalisering is meer dan het één-op-één kopiëren van analoog naar digitaal. Er bestaan diverse mogelijkheden om het digitale eindresultaat te verrijken ter ondersteuning van flexibele zoek- en gebruiksmogelijkheden. Uit financiële overwegingen wordt een tekst vaak niet volledig machineleesbaar gemaakt (bijvoorbeeld door middel van OCR of data entry), maar alleen digitaal gereproduceerd en als plaatje aangeboden. Natuurlijk vergroot dit de mogelijkheid tot raadpleging van de tekst, maar dat gebruikers niet op de tekst zelf kunnen zoeken, zullen zij toch als een tekortkoming zien. Als een tekst volledig digitaal beschikbaar is, kan er in ieder geval op Google-achtige wijze in gezocht worden. De hoogste mate van zoeken in tekst en uitwisselen van passages wordt echter bereikt wanneer een tekst wordt voorzien van interne structurering, mark up genaamd. De internationale standaard hiervoor is de TEI, Text Encoding Initiative (TEI, 2007). TEI is ontwikkeld om teksten op een verantwoorde wijze te structureren en weer te kunnen analyseren, maar heeft inmiddels een breder toepassingsbereik gevonden. Omdat TEI volledig in XML wordt uitgedrukt, zijn de opgemaakte teksten geheel compatibel met het web.

XML is momenteel een van de meest gangbare formaten voor tekstbestanden op het web, tesamen met HTML en PDF (Boudrez, 2005). Hoewel het DOC-formaat van de teksteditor Microsoft Word een van de meest gebruikte tekstformaten in Nederland is, is het nooit een serieuze kandidaat geweest voor raadpleging via het web, omdat het te verknoopt is met de software van Microsoft.

In de formaten XML en PDF komen het internet en de praktijk van uitgevers samen. XML is een voor het web afgeleide van SGML, een opmaak- en markeertaal die al decennialang door uitgevers wordt gehanteerd voor het gereed maken van hun publicaties en databases. PDF is een formaat dat enerzijds de 'look-and-feel' van een set gedrukte pagina's kan aanhouden, en anderzijds goede zoek- en presenteermogelijkheden biedt voor digitale informatie. PDF is om die reden de standaard geworden voor het beschikbaar stellen van wetenschappelijke tijdschriftartikelen. Omdat PDF als formaat niet copyright-vrij is, en Microsoft's DOC niet platformonafhankelijk, is een nieuw formaat ontwikkeld dat wel geheel 'open' is, en dat nauw verwant is met de wijze waarop het web omgaat met tekstdocumenten: het Open Document format (ODF). Onlangs is dit formaat, dat op XML is gebaseerd, door ISO officieel tot internationale standaard benoemd. Dit formaat is in eerste instantie ontwikkeld voor het uitwisselen van kantoordocumenten (brieven, memo's, presentaties, spreadsheets), maar de ontwikkelaars stellen nadrukkelijk dat ODF ook geschikt is voor het verspreiden en toegankelijk houden van digitale rapporten en boeken. Juist door de garanties voor interoperabiliteit zou dit formaat heel belangrijk kunnen worden voor digitale bibliotheken (Information technology, 2006).

Bij thesauri zien we vergelijkbare ontwikkelingen als bij volledige teksten. Door thesauri uit te drukken in XML met een vast schema wordt de inzetbaarheid ervan vergroot, met name ten behoeve van de ontwikkeling van het semantisch web. In de bibliotheeksector zijn de meeste thesauri deductief, dat wil zeggen opgebouwd op basis van een specifieke collectie of groep collecties. Hierbij wordt een koppeling gemaakt tussen beschrijvingen van objecten in de catalogus en de thesaurus, zodat de objecten via gestandaardiseerde ingangen vindbaar zijn. Maar de kennis en intelligentie die informatieprofessionals in thesauri stoppen, zijn ook voor

andere toepassingen nuttig, zoals het bevragen van metadataverzamelingen elders, het doorzoeken van volledige teksten of het verknopen met andere thesauri. De ontwikkelingen rondom het gebruik van thesauri op het web zijn nog volop aan de gang. De twee belangrijkste instrumenten, beide opgezet vanuit het World Wide Web Consortium (W3C), zijn Web Ontology Language (OWL) en Simple Knowledge Organisation System (SKOS). OWL is een uitgebreide mark-up taal voor ontologieën in de breedste zin van het woord, SKOS biedt een simpeler model om de structuur van terminologiebronnen als thesauri in XML uit te drukken (W3C, 2007).

BEHEER VAN DIGITALE DATA

Uitwisseling met derden

De digitale bibliotheek van een instelling bestaat, het is al gezegd, niet alleen uit data die de instelling bezit of zelf heeft vervaardigd. Er zijn diverse mogelijkheden om data van derde partijen - het gaat ook hier om full text documenten, metadata, contextuele informatie en/of thesauri - op te nemen. De simpelste vorm is het opnemen van verwijzingen, bijvoorbeeld in de vorm van een set bookmarks die je voor je doelgroep onderhoudt. Als gebruikers van een digitale bibliotheek tegelijkertijd in de eigen data en de data van anderen moeten kunnen zoeken, zijn er diverse mogelijkheden om dit te realiseren. De zoekactie kan bijvoorbeeld via de interface van het zoeksysteem naar de aangewezen diensten geleid worden in een gedistribueerde zoekactie (meer hierover in de volgende paragraaf). Een andere mogelijkheid is de data binnenhalen en opnemen in je eigen zoeksystemen, bijvoorbeeld via crawlers of robots. Dit is de wijze waarop de grote zoekmachines op internet hun indexen vullen.
Een populair protocol voor het binnenhalen van metadata is het Open Archives Initiative Protocol for Metadata Harvesting (OAI). Ook OAI wordt volledig in XML uitgedrukt, en het protocol maakt gebruik van Dublin Core als gemeenschappelijk metadatamodel. OAI biedt ook de mogelijkheid de digitale documenten zelf binnen te halen die met de metadata worden beschreven (The Open Archives Initiative Protocol for Metadata Harvesting, 2007).

Opslag en toegang

Als een bibliotheek eenmaal beschikt over een basiscorpus digitale data, zelf aangemaakt of van derden binnengehaald, zijn bewerkingen nodig om de informatie beschikbaar te krijgen en te houden. Dit varieert van opslag op systemen waar informatiediensten mee kunnen communiceren tot contentmanagement ten behoeve van stroomlijning van de dienstverlening.

Digitale informatie opslaan lijkt simpel, maar het is een complex proces dat op basis van een zorgvuldig doordachte architectuur uitgevoerd dient te worden om stabiliteit en interoperabiliteit te garanderen. Onder de Nederlandse erfgoedinstellingen heeft het Instituut voor Beeld en Geluid in Hilversum momenteel de grootste opslagcapaciteit. Zij ontvangen wekelijks circa 1 Terabyte aan digitale content van de publieke omroepen, en ze hebben op het moment van schrijven een totale opslagcapaciteit van 1,8 Petabyte (1.800.000 Gigabyte). Deze infrastructuur wordt onder andere ingezet voor de portal omroep.nl en themakanalen van de omroepen op internet. Het instituut kan deze grootschalige en geavanceerde opslag (er wordt o.a. gebruik gemaakt van tape-robots) alleen onderhouden dankzij de expertise van het NOB en een flinke financiële injectie van de overheid. Nu zullen de meeste digitale bibliotheken niet op korte termijn met een dergelijke grootschaligheid geconfronteerd worden, maar de principes achter de architectuur van de opslag bij Beeld en Geluid zijn voor hen niet wezenlijk anders. Die principes zijn veilige opslag, bediening-op-maat en snelle toegang.

Veilige opslag vereist dat de data op meer dan één plek is opgeslagen. De kopieën kunnen binnen hetzelfde opslagsysteem aanwezig zijn, bijvoorbeeld bij servers die gebruik maken van RAID (Redundant array of independent disks), waarbij de data over diverse disks wordt verdeeld. De data kan ook op externe faciliteiten worden opgeslagen, zoals een backup-server of offline op dvd of cd-rom. Dit laatste heeft zeker niet de voorkeur, in verband met de onzekere levensduur van schrijfbare optische media en de arbeidsintensieve bewerking die nodig is om de data te migreren als de dragers onleesbaar dreigen te worden. Iedere instelling zou ook serieus moeten overwegen een kopie van de inhoud van de digitale bibliotheek elders onder te brengen, zodat ook bijvoorbeeld in het geval van brand of andere lokale rampen de data is veiliggesteld.

Bediening-op-maat wil zeggen dat de opslag zodanig geregeld is dat alleen de gedeeltes die nodig zijn opgevraagd kunnen worden. Dit is niet alleen een goede dienstverlening aan de klant, het is ook nuttig om overbelasting van het systeem te voorkomen. Dergelijke bediening-op-maat vereist dat informatie over de opbouw van de digitale objecten opgeslagen en beschikbaar is, bijvoorbeeld in de vorm van 'structurele metadata'. In het geval van een gedigitaliseerd boek bijvoorbeeld wordt in de structurele metadata bijgehouden uit hoeveel pagina's het boek bestaat, hoe de hoofdstukindeling is en waar de index of het register zit. Dé standaard voor structurele metadata in de bibliotheeksector is Metadata Encoding & Transmission Standard (METS), ontwikkeld door de Library of Congress, dat gebruik maakt van XML en rekening houdt met andere webconventies (METS, 2007).

Snelle toegang is een andere belangrijke eis om interactie met de gebruiker te garanderen. Een trage responstijd wordt als onacceptabel beschouwd en jaagt bezoekers weg. Het vergt wel aanzienlijke investeringen in krachtige netwerkomgevingen, bijvoorbeeld een Network Attached Storage (NAS) of een Storage Area Network (SAN) en hoogwaardige software (bijvoorbeeld databases en indexeersoftware) om volledig aan deze eis tegemoet te komen. Het zal dan ook niet in alle gevallen mogelijk zijn te garanderen dat alle data direct met één druk op de knop snel beschikbaar zijn. In dit opzicht wordt wel eens onderscheid gemaakt tussen hot, warm en cold storage. 'Hot storage' wil zeggen dat de informatie direct op een server staat die rechtstreeks is aangesloten op het netwerk. 'Warm storage' wil zeggen dat de data zijn opgeslagen op randapparatuur bij een server (bijvoorbeeld in een tape-robot). Bij het opvragen van een bestand moet eerst de goede tape geplaatst worden voor het bestand aan een gebruiker gepresenteerd kan worden. 'Cold storage' houdt in dat de data alleen op externe media zijn opgeslagen en alleen na interventie van een bibliotheekmedewerker door een bezoeker gelezen kunnen worden.

Duurzame toegankelijkheid

Als de toegang tot de digitale informatie volgens bovengenoemde criteria is gerealiseerd, is er nog geen garantie dat de data ook voor de lange termijn beschikbaar zullen blijven. Voor duurzame opslag en toegankelijkheid gelden aanvullende eisen aan hardware, software en documentformaten. Het is nog lang niet duidelijk hoe aan die eisen voldaan kan worden. Bibliotheken en archieven spelen wereldwijd een hoofdrol in het onderzoek naar digitale duurzaamheid.

Er is momenteel één officiële internationale standaard voor digitale duurzaamheid, de Reference Model for an Open Archival Information System (OAIS) (Verheul, 2006). Dit model beschrijft waaraan een digitaal archief ('repository') moet voldoen om de opgeslagen data voor de lange termijn te kunnen beheren. Zo geeft OAIS richtlijnen hoe een digitaal document ingelezen moet

worden en van aanvullende metadata moet worden voorzien, voor het daadwerkelijk gearchiveerd kan worden, en ook hoe documentatie rondom dit proces moet worden opgesteld. Enkele bibliotheken en bibliotheekorganisaties, waaronder OCLC en RLG, hebben het initiatief genomen om te komen tot een vaste set metadata die nodig is voor het duurzaam bewaren van digitale documenten, de zogeheten 'preservation metadata'. Het resultaat is PREMIS, een Data Dictionary for Preservation Metadata (PREMIS, 2007). Hierin is vastgelegd hoe informatie over bijvoorbeeld de software waar het document mee is vervaardigd geregistreerd moet worden, evenals het versienummer van het format waarin het document wordt bewaard. Dergelijke informatie is niet alleen nuttig voor eigen beheer, maar ook om toekomstige toegang tot de documenten vanuit andere informatiediensten te ondersteunen.

Omdat niet alles met metadata is op te lossen, wordt ook gekeken hoe documentformaten zelf een betere kans op een lang leven kunnen krijgen. Het grote probleem bij formaten is dat er steeds nieuwe versies verschijnen, met meer en meer functionaliteit, en dat software die een bepaald format kan inlezen niet altijd meer oude versies ervan kan verwerken. Om die reden is een afgeleid formaat van PDF gedefinieerd, dat inmiddels als ISO-standaard is erkend, en waarvoor Adobe afstand heeft gedaan van de auteursrechten. Het betreft hier PDF/A, waarbij de A staat voor Archiving. De functionaliteit van PDF is in deze versie dermate opgeschoond, dat het gebruik ervan voor langere tijd gegarandeerd is.

Voor formats geldt wel in het algemeen dat problemen zich opstapelen als de software waarmee oude versies nog wel gelezen kunnen worden, zelf niet meer op nieuwe generaties computers gebruikt kunnen worden. Er zijn dan twee mogelijkheden om de data toegankelijk te houden: de data migreren naar een nieuw format dat wel weer door nieuwe generaties hard- en software verwerkt kan worden - er zijn, ook in Nederland, diverse instellingen die al op flinke schaal oude databases migreren naar XML om zo de informatie beschikbaar te houden - of ervoor zorgen dat de oude software gesimuleerd kan worden op de nieuwe computers (emulatie) (Publicaties van Testbed Digitale Bewaring, 2007).

Identificatie

De wijze van opslag hangt nauw samen met de terugvindbaarheid van een document. Op het web worden documenten geïdentificeerd door de plek te benoemen waar ze zijn opgeslagen, de Uniform Resource Locator (URL). Dit is een kwetsbare wijze van identificatie: zodra je het bestand hernoemt of verplaatst naar een andere map, klopt de verwijzing niet meer, en is het document onvindbaar geworden. Het is dus nuttig om plaatsonafhankelijke identificatie van documenten te gebruiken. Er zijn al diverse oplossingen beschikbaar, maar deze worden alleen in professionele kringen toegepast. Omdat de gangbare webbrowsers deze nieuwe verwijzingen nog niet of nauwelijks ondersteunen, zijn ze nog geen gemeengoed op het web. De webgemeenschap zelf heeft de Uniform Resource Name (URN) ontworpen als basis voor geautomatiseerde informatieprocessen op het web (Uniform Resource Names, 2002). Uit de informatiesector is de Digital Object Identifier (DOI) afkomstig. (The DOI System, 2007). Voorbeelden hiervan zijn onder andere te vinden in het uitstekende online tijdschrift over digitale bibliotheken, D-LIB magazine (D-lib Magazine, 2007).

ZOEKEN EN PRESENTEREN

Als een bibliotheek ervoor heeft gezorgd dat haar digitale data goed (dat wil zeggen: volgens breed geaccepteerde schema's en standaarden) is gestructureerd, opgeslagen en gedocumenteerd, kan ze op een zeer flexibele wijze haar bezoekers bedienen. Was het in het

OPC-tijdperk gebruikelijk dat de informatiedienst centraal stond, en dat iedereen er maar mee moest leren werken zoals het was, in het webtijdperk verwachten gebruikers meer flexibiliteit in de dienstverlening. Ze willen zelf meer controle over het doorzoeken van de aanwezige data en de wijze waarop de resultaten gepresenteerd en verder verwerkt kunnen worden.

Presentaties-op-maat

Deze behoefte aan grotere controle over de informatiediensten komt niet alleen voor bij eigenwijze eindgebruikers die de bibliotheek niet als een autoriteit willen accepteren. Die controle kan ook noodzaak zijn voor mensen met een visuele handicap. Een belangrijke richtlijn voor de ondersteuning van visueel gehandicapten is het Web Accessibility Initiative (WAI) (Web Accessibility Initiative, 2007). Door je te houden aan de aanbevelingen die het WAI doet, kun je ervoor zorgen dat je informatiediensten door een zo groot mogelijke groep probleemloos geraadpleegd kunnen worden. In Nederland waakt DrempelsWeg over de toegankelijkheid van websites voor visueel gehandicapten.

Een stap verder gaan de gloednieuwe richtlijnen die zijn opgesteld voor de websites van de Nederlandse rijksoverheid. Met ingang van 1 september 2006 moeten al deze websites voldoen aan een set van 125 richtlijnen om de toegankelijkheid, duurzaamheid, uitwisselbaarheid en vindbaarheid te vergroten (Richtlijnen voor de toegankelijkheid, 2007). De richtlijnen betreffen onder andere het gebruik van links, paginastructuur en visuele elementen, zoals afbeeldingen, maar ook citaten.

De eerste richtlijn luidt: "Houd structuur en vormgeving zoveel mogelijk gescheiden: gebruik HTML of XHTML voor de structuur van de site en CSS voor de vormgeving ervan." Dit is op zich een verstandig advies, maar dan vooral met betrekking tot statische webpagina's. Als het gaat om dynamische pagina's, zoals bij de presentatie van records uit een catalogus, is het verstandiger XML als basis te gebruiken. Een van de grootste voordelen van XML ten opzichte van HTML is de absolute scheiding tussen inhoud en opmaak. HTML-codering betreft vooral opmaakelementen: hoe groot wordt de titel getoond, wat moet vet of cursief getoond worden, op welke positie staan afbeeldingen? Met XML is er veel meer flexibiliteit. Bij een tekst die met XML is gecodeerd, kun je aangeven wat de titel is, waar een nieuw hoofdstuk of een nieuwe paragraaf begint, maar ook welke woorden bijvoorbeeld een persoons- of plaatsnaam aanduiden. Afhankelijk van het medium waarmee een bezoeker de tekst wil raadplegen (PC, laptop, PDA, mobieltje), of de context waarin de data wordt geraadpleegd (bijvoorbeeld Blackboard-achtige toepassingen in het onderwijs of geïntegreerde catalogi), kunnen presentaties worden gegenereerd die passen bij dat medium of die context zonder dat de tekst hierop aangepast hoeft te worden. Die wisselende presentaties worden gerealiseerd met behulp van bijbehorende XML stylesheets of transformaties naar HTML. Een van de krachtigste toepassingen op dit gebied is RSS, de mogelijkheid om nieuwsberichten eenmalig te plaatsen en daarna op een veelheid aan media gepresenteerd te krijgen Vergelijkbaar met deze opzet is de presentatie van zoekresultaten afkomstig uit verschillende informatiediensten. Als de informatiediensten interoperabel zijn met elkaar, en data met gelijke structureringen kunnen uitwisselen, kan voor de gebruiker een voor het oog naadloze integratie tot stand worden gebracht, ook al blijven het achter de schermen gescheiden dataverzamelingen. Op deze wijze kunnen samenwerkingspartners in gezamenlijkheid, maar ook individueel, hun data naar believen aanbieden, presenteren en exploiteren.

Gedistribueerd zoeken

Er zijn diverse mogelijkheden om dataverzamelingen gelijktijdig doorzoekbaar te maken en in één presentatie te tonen. Als (meta)data na harvesting centraal wordt opgeslagen en geïndexeerd (zie paragraaf hierboven), kunnen diverse collecties in één gerichte zoekactie worden geraadpleegd. Dit is de methode die de grote zoekmachines op het web hanteren. Als er geen sprake is van harvesting, ligt de oplossing in het zenden van de zoekvraag ('query') naar de diverse dataverzamelingen. Dit kan bij volledige teksten eenvoudig met het basisprotocol van het web, het http-protocol. Hiervan maken 'meta-search engines' als Dogpile gebruik. Specifiek voor metadata bestond deze gedistribueerde zoekmogelijkheid al in het pre-web-tijdperk. Precies voor dit doel was immers het Z39.50-protocol ontwikkeld door de bibliotheekgemeenschap. Met behulp van een gemeenschappelijk metadataprofiel was het mogelijk om bibliografische beschrijvingen van de ene bibliotheek op te vragen en te bekijken in de OPC van een andere bibliotheek. De opvolger van Z39.50 die in webportals gebruikt wordt, is SRU, Search and Retrieve URL Service (SRU, 2007).

Een interessante variant op het gelijktijdig bevragen van verschillende informatiediensten is het OpenURL-concept (The OpenURL Framework, 2007). OpenURL's worden gebruikt voor 'context-gevoelige' diensten. Stel, je hebt een beschrijving van een interessante publicatie gevonden in een database, maar je kunt de volledige publicatie niet opvragen, omdat je niet daartoe gemachtigd bent. Via een OpenURL-zoekvraag kun je dan bekijken of diezelfde publicatie beschikbaar is in andere databases, waar je mogelijk wel toegang hebt tot de volledige publicatie. De conventie achter OpenURL is dat een publicatie vindbaar is met identificerende metadatavelden, bijvoorbeeld een ISBN.

INTERACTIEVE INFORMATIEDIENSTEN

De trends en technieken die hierboven beschreven zijn, gaan ervan uit dat de bibliotheek het aanbod heeft, en de eindgebruiker de vraag die beantwoord moet worden. De laatste ontwikkelingen op internet geven aan dat deze benadering wellicht op de schop moet. Onder de naam Web 2.0 worden webdiensten geschaard die tot stand komen door inbreng van wie maar wil. Een ieder met toegang tot het web kan bijdragen aan gemeenschappelijke online informatie of registratie van kennis. Een bekend voorbeeld is Wikipedia, een encyclopedie waaraan iedereen kan bijdragen. De snelheid waarmee Web 2.0-toepassingen opgepakt worden en groeien geeft aan dat dit een aanpak is die succesvol is en niet snel weer zal overwaaien. Het is nog onduidelijk wat dit voor consequenties heeft voor instellingen die informatiebemiddeling als professie hebben. Een reëel voorbeeld zal zijn dat de ene eindgebruiker van een bibliotheek een andere zal helpen, zonder dat er een professional tussen zit. Het ziet ernaar uit dat de digitale bibliotheek niet alleen interoperabiliteit achter de schermen moet realiseren om de eindgebruikers zo goed mogelijk te bedienen. Er moeten ook faciliteiten komen om de kennis van de eindgebruikers onderdeel te laten uitmaken van de digitale bibliotheek. Op deze manier werken we toe naar een wereldwijde digitale bibliotheek die niet alleen vóór iedereen is, maar vooral ook ván iedereen.

CONCLUSIES

In bovenstaande paragrafen is aan de hand van enkele fasen uit de levenscyclus van digitale data beschreven hoe de digitale bibliotheek (uit)gebouwd kan worden. Niet alle genoemde bouwstenen voor vervaardiging, beheer en presentatie van data worden al uitgebreid in de praktijk toegepast. Ook al kunnen ze beschouwd worden als kwaliteitsinstrumenten, de levensvatbaarheid ervan kan niet in alle gevallen gegarandeerd worden. Maar dat geldt ook voor

de beschreven bouwstenen die inmiddels wel al erkend zijn als standaard. Het succes of falen van een standaard hangt niet alleen af van de kwaliteit die de standaard kan garanderen, maar vooral ook van de bereidheid van instellingen zich aan de standaard te committeren en van de concurrentie van alternatieven. Standaarden zijn dan ook niet de essentie van de digitale bibliotheek, ze zijn middel tot een doel.

Wat is dat doel? Negatief geformuleerd zou je kunnen stellen dat het concept van de digitale bibliotheek zoals hier beschreven feitelijk draait om het veilig stellen van bibliothecaire tradities (inclusief werkgelegenheid) in een zich rap digitaliserende informatiemaatschappij. Bibliotheken waren voorheen vanzelfsprekende informatiecentra. Dat is in dit tijdperk van het web niet meer het geval. Ieder individu heeft de middelen binnen bereik om informatie te vinden, vervaardigen, bewerken, beheren, verspreiden en presenteren. Bibliotheken zullen zich veel meer dan vroeger moeten richten op het leveren van diensten-op-maat om hun bestaan als informatiedienstverlener te rechtvaardigen. Kwaliteit is daarbij een kernbegrip: bezoekers zullen bibliotheekdiensten gebruiken als ze ervan overtuigd zijn dat daar hun informatiebehoefte op een hoogwaardige wijze wordt opgelost. Dat waardeoordeel kan gebaseerd worden op de kwaliteit van de geboden informatie, maar ook bijvoorbeeld op de mate van efficiëntie en gebruiksgemak.

De bouwstenen die hierboven gepresenteerd zijn, hebben alle één ding gemeen: ze dragen bij aan een betere toegankelijkheid en uitwisselbaarheid van informatie. Het zijn deze twee principes die cruciaal zijn voor informatiedienstverlening-nieuwe-stijl, en daarmee het slagen van de digitale bibliotheek. Als een bibliotheek niet actief hierin investeert, zal haar functie gereduceerd worden tot uitleencentrum van boekengenres die in drukvorm populair blijven, of, als de collectie exclusief genoeg is, tot een oude-media-museum. Zonder digitale dienstverlening heeft de informatiefunctie van een bibliotheek geen toekomst. Wat die dienstverlening behelst moet door iedere bibliotheek in nauwe samenspraak met de gebruikersgroep worden vastgesteld, en met andere professionale partijen in de informatievoorziening. De besproken bouwstenen dienen bovenal dat doel: in gezamenlijkheid de dienstverlening voor toekomstige generaties informatiezoekers vormgeven.

LITERATUURVERWIJZINGEN

Boudrez, F. (2005). Standaarden voor digitale archiefdocumenten. Geraadpleegd op 9 maart 2007 op: www.expertisecentrumdavid.be/docs/eDAVID_standaarden.pdf

CIDOC, The CIDOC Conceptual Reference Model. Geraadpleegd op 9 maart 2007 op: hcidoc.ics.forth.gr/

D-lib Magazine. Geraadpleegd op 11 maart 2007 op: www.dlib.org

The DOI System. Geraadpleegd op 11 maart 2007 op: www.doi.org

Dublin Core in samenwerkingsprojecten en publieksgerichte ontsluiting. (2006). Geraadpleegd op 9 maart 2007 op: www.den.nl/docs/20050816173630/

i2010: Digital libraries. European Communities, 2006. ISBN 92-79-02332-2. Geraadpleegd op 9 maart 2007 op: ec.europa.eu/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf

ICT-register voor het cultureel erfgoed. Geraadpleegd op 9 maart 2007 op: www.den.nl/register

IFLANET Cataloguing Section FRBR Review Group. Geraadpleegd op 9 maart 2007 op: www.ifla.org/VII/s13/wgfrbr/index.htm
Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0. (2006) Geraadpleegd op 9 maart 2007 op: www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=43485
METS, Metadata Encoding & Transmission Standard. Geraadpleegd op 11 maart 2007 op: www.loc.gov/standards/mets/
The Open Archives Initiative Protocol for Metadata Harvesting. Geraadpleegd op 11 maart 2007 op: www.openarchives.org/OAI/openarchivesprotocol.html
The OpenURL Framework for Context-Sensitive Services. Geraadpleegd op 11 maart 2007 op: http://www.niso.org/committees/committee_ax.html
PREMIS, Preservation Metadata Maintenance Activity. Geraadpleegd op 11 maart 2007 op: www.loc.gov/standards/premis/
Publicaties van Testbed Digitale Bewaring. Geraadpleegd op 11 maart 2007 op: www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=146&categorie=2
Richtlijnen voor de toegankelijkheid en duurzaamheid van overheidswebsites. Geraadpleegd op 11 maart 2007 op: webrichtlijnen.overheid.nl/
SRU: Search/Retrieve via URL. Geraadpleegd op 11 maart 2007 op: www.loc.gov/standards/sru/index.html
TEI, Text Encoding Initiative. Geraadpleegd op 9 maart 2007 op: www.tei-c.org
Uniform Resource Names (URN) Namespace Definition Mechanisms (2002). Geraadpleegd op 11 maart 2007 op: www.ietf.org/rfc/rfc3406.txt
Verheul, I. Networking for Digital Preservation: Current Practice in 15 National Libraries. München: Saur, 2006.
W3C, World Wide Web Consortium. Geraadpleegd op 11 maart 2007 op: www.w3.org
Web Accessibility Initiative (WAI). Geraadpleegd op 11 maart 2007 op: www.w3.org/WAI/