

## Standaarden voor gegevensuitwisseling

T.M. VAN VEEN

### 1 Inleiding

Het is een trend in de informatiedienstverlening om gegevens uit verschillende bronnen te mengen en te hergebruiken in nieuwe diensten. Ook presenteren informatieaanbieders steeds meer de eigen data zodanig, dat anderen daarvan in hun diensten gebruik kunnen maken en zij deze gegevens, gemengd met die van anderen, weer kunnen aanbieden. Om dit mogelijk te maken, moet gebruik worden gemaakt van standaarden op het gebied van gegevensstructurering en gegevensuitwisseling. We hebben te maken met elkaar overlappende, aanvullende en opvolgende standaarden, waar soms een keuze uit gemaakt moet kunnen worden.

In dit artikel wordt een overzicht gegeven van standaarden die momenteel relevant kunnen zijn voor de informatie-infrastructuur van bibliotheken en vergelijkbare instellingen. Van elke standaard wordt het gebruik en/of de werking uitgelegd en wordt de bruikbaarheid of relevantie aangegeven. Voor een gedetailleerde beschrijving is een link toegevoegd naar de webpagina waar de standaard beschreven wordt.

Bij het adopteren van standaarden spelen ook andere aspecten dan kwaliteit mee, bijvoorbeeld de implementatiedrempel en adoptie door andere partijen. Denk hierbij aan wat er destijds is gebeurd met de videostandaarden VHS, Betamax en Video 2000.

De standaarden zijn als volgt onderverdeeld:

- Metadata-standaarden voor het specificeren van beschrijvingen van boeken, images en andere digitale objecten.
- Protocollen die beschrijven hoe gegevens uitgewisseld worden. Het

betreft hier zoeken, opvragen en vergaren (harvesten) van zowel metadata als andere data.

- Identificatie van (beschreven) objecten.
- Structurering van tekst of samengestelde objecten. In sommige gevallen wordt gesproken over structurele metadata.
- Authenticatie.
- Protocollen voor communicatie tussen applicaties.
- Thesauri/ontologieën.

Sommige standaarden kunnen gebaseerd zijn op andere standaarden. Zo maken veel van de standaarden die hier genoemd worden, gebruik van XML (eXtended Markup Language). XML is een manier om gegevens te structureren.

In XML wordt elk onderdeel van een verzameling datavelden voorafgegaan en afgesloten met de naam van zo'n onderdeel en deze naam wordt herkend aan < en >. Bijvoorbeeld een titelveld zou in XML er uit kunnen zien als:

```
<titel>Dit is een titel</titel>
```

Zo'n XML-onderdeel, tag genoemd, kan zelf ook weer opgedeeld zijn in tags. Een XML-document is op deze manier machineleesbaar en, in tegenstelling tot sommige andere standaarden om data te structureren, ook redelijk goed leesbaar voor mensen.

De standaarden voor gegevensuitwisseling zijn in dit artikel allemaal gebaseerd op het http-protocol. Dit is het protocol waarmee ook web-browsers zoals Internet Explorer met webservern communiceren.

Per standaard wordt ingegaan op de volgende punten:

- Wat betekent de afkorting?
- Waar is het voor/welk probleem wordt ermee opgelost?
- Welke standaard wordt ermee vervangen?
- Wat is de URL naar de site waar de standaard onderhouden of beschreven wordt?
- Wat is de relevantie en/of toekomstverwachting?

## 2 Metadata

Metadata dienen om objecten zoals boeken of video e.d. op een gestructureerde manier te beschrijven en te ontsluiten. Voorbeelden van metadata-velden zijn titel, auteur, onderwerp en de locatie van een object. Bibliotheekcatalogi bijvoorbeeld bestaan vooral uit metadata. Redenen om metadata te standaardiseren zijn de mogelijkheid tot uitwisseling van objectbeschrijvingen en het machineleesbaar maken ervan. Een voorbeeld van een toepassing is een portal die toegang geeft tot meerdere databases. Zonder een standaard metadatamodel moeten de metadata voor iedere database weer anders geconverteerd worden.

### 2.1 Dublin Core (DC)

Dublin Core is een eenvoudige standaard voor metadata-velden ten behoeve van het beschrijven van allerlei type objecten. Dublin (Ohio) is de plaats waar het initiatief voor deze standaard ontstaan is. Het doel van Dublin Core is om de wildgroei aan metadata-velden tegen te gaan en om in ieder geval een minimum set goed gedefinieerde velden te hebben voor het uitwisselen van metadata. DC vervangt niet een andere standaard en wordt vaak naast andere standaarden gebruikt. DC wordt onderhouden door de DCMI (Dublin Core Metadata Initiative) organisatie.

Naast de basisset van 15 elementen zoals title, creator, identifier enzovoorts, zijn er ook nog meer dan 40 verfijningen van die elementen. Het voordeel van DC is dat het niet afgestemd is op een specifiek domein en daarmee voor veel applicatiegebieden bruikbaar is. Het wordt dus ook gebruikt door overheidsinstellingen e.d. Dit in tegenstelling tot bijvoorbeeld MARC (zie 2.4) dat specifiek voor bibliografische beschrijvingen bestemd is en alleen in bibliotheken gebruikt wordt.

Omdat DC voor specifieke toepassingen meestal te beperkt is, worden hieraan vaak extra velden toegevoegd. In eerste instantie zijn dat toevoegingen die door de DCMI-organisatie ondersteund worden. Dit wordt dan *Qualified Dublin Core* genoemd. Is dat nog niet voldoende, dan worden ook velden uit andere sets van veldnamen gebruikt. Dit moet ergens worden vastgelegd. In de DC-wereld wordt hiervoor veelal een zogenaamd applicatieprofiel gebruikt. Dit is een lijst met gebruikte velden en hun karakteristieken zoals definitie, herhaalbaarheid en in welke dataset ze hun oorsprong vinden.

*Relevantie:* Dublin Core is zeer relevant als standaard formaat voor de infor-

matie-uitwisseling tussen verschillende omgevingen zoals bibliotheken, musea en archieven. Door onderscheid te maken tussen de DC-termen enerzijds en extensies hierop anderzijds kunnen metadata op een standaard manier met de buitenwereld gedeeld worden, terwijl men tegelijkertijd via extra termen specifieke lokale functionaliteit kan blijven bieden. Om met externe partijen informatie te kunnen uitwisselen over extra toegevoegde metadata-termen is het gebruik van een centrale database met de gebruikte metadata-velden (metadata registry) erg handig. Het kan ook van belang zijn voor de eigen organisatie om gebruikte metadata eenduidig vast te leggen.

*URL:* <http://dublincore.org/>

### 2.2 *ContextObjects in Spans (COinS)*

COinS is een conventie voor het plaatsen van bibliografische metadata in HTML op een zodanige manier dat er nieuwe, gebruikersspecifieke links gegenereerd kunnen worden. Door gebruik te maken van vrij verkrijgbare browseruitbreidingen wordt het mogelijk gemaakt met deze extra metadata een OpenURL (zie 3.4) te genereren en als link in de desbetreffende pagina zichtbaar te maken. COinS maakt gebruik van 'span' tags in een HTML document. Deze worden normaal gesproken gebruikt om een stuk tekst te markeren en van een eigen opmaak te voorzien. Een span heeft een aantal attributen, waaronder het attribuut 'title'. In COinS worden in het title-attribuut metadata toegevoegd die als OpenURL gecodeerd zijn. Een browserextensie kan deze specifieke span tags herkennen en veranderen.

*Relevantie:* Verwacht wordt dat dit mechanisme een van de belangrijkste mechanismen gaat worden om de gebruiker invloed te laten hebben op de functionaliteit die geboden kan worden via een webpagina bijvoorbeeld door automatisch links te laten genereren naar zelf gekozen diensten op basis van gegevens die via COIN S in een webpagina gecodeerd zijn. Indien het mechanisme grootschalige toepassing vindt, zal het gebruik veel algemener kunnen zijn dan alleen OpenURL.

*URL:* <http://ocoins.info/>

### 2.3 *Encoded Archival Description (EAD)*

EAD is een formaat om collecties te beschrijven. EAD wordt vooral binnen de archiefwereld gebruikt maar in toenemende mate ook in de bibliotheekwereld. Het verschil met collectiebeschrijvingen gebaseerd op Dublin Core is dat met EAD ook individuele objecten in één EAD-record beschreven

kunnen worden, terwijl met de gangbare Dublin Core formaten collecties alleen op collectieniveau worden beschreven.

*Relevantie:* EAD wordt volop gebruikt in de archiefwereld. De andere hier genoemde standaarden kunnen EAD niet vervangen. Voor bibliotheken heeft het gebruik van EAD alleen zin als individuele objecten niet in de catalogus (kunnen) worden opgenomen. Anderzijds zal door de wens om de gegevens van musea, bibliotheken en archieven steeds meer geïntegreerd toegankelijk te maken, passieve ondersteuning zinvol kunnen zijn, dat wil zeggen dat EAD records van andere instellingen wel gepresenteerd moeten kunnen worden.

*URL:* <http://www.loc.gov/ead/>

#### 2.4 *Machine Readable Cataloguing (MARC) en MARCXML*

MARC is een formaat om bibliografische gegevens vast te leggen. Velden (tags) worden geïdentificeerd met een nummer, en subvelden door een letter voorafgegaan door een '\$'. Veel landen hadden hun lokale varianten, maar tegenwoordig wordt veel gebruik gemaakt van MARC21 en UNIMARC. Hierin zijn weer zaken vastgelegd waar de MARC-standaard zich niet over uitlaat.

Indien MARC21 records in XML gerepresenteerd worden, dan spreekt men van MARCXML. MARC-tags en -subvelden worden dan omgezet naar XML-tags en XML-attributen. Dit maakt het mogelijk om XML-tools te gebruiken voor de presentatie en verwerking van MARC-records.

*Relevantie:* MARC is een nog niet weg te denken formaat voor bibliografische beschrijvingen. Maar MARC moet steeds meer terrein afstaan aan andere formaten zoals MODS en Dublin Core, en misschien dat MARC zijn langste tijd gehad heeft. Het gebruik van MARCXML maakt een overgang naar andere formaten (DC, MODS) makkelijker dan het gebruik van willekeurige eigen formaten, omdat er programma's beschikbaar zijn voor conversie van MARCXML naar DC en MODS.

*URL:* <http://www.loc.gov/standards/marcxml/>

#### 2.5 *Metadata Object Description Schema (MODS)*

MODS is een XML-standaard voor bibliografische metadata met name gericht op bibliotheektoepassingen. MODS is bedoeld om delen van MARC-records in XML weer te geven met tekstlabels in plaats van de numerieke MARC-velden. Het is aanzienlijk rijker dan Dublin Core, maar alleen ge-

richt op bibliografisch materiaal. MODS kan gebruikt worden als metadata formaat in SRU (zie 3.5) en als uitbreidingschema op METS (zie 5.2).

*Relevantie:* Het is nog niet duidelijk of de buitenwereld hier massaal op overgaat. Wel is het in portalsoftware raadzaam om in staat te zijn de MODS-records van andere dataproviders te kunnen lezen en verwerken op dezelfde manier als DC en MARCXML.

*URL:* <http://www.loc.gov/standards/mods/>

#### 2.6 ONline Information eXchange (ONIX)

ONIX is een standaard XML-schema voor uitgevers om langs elektronische weg informatie over boeken uit te wisselen met hun handelspartners, zoals bibliotheken, boekhandels en groothandels. Het formaat is zeer verschillend van de gangbare bibliografische formaten en vereist daarom conversies voor het gebruik door bibliotheken.

*Relevantie:* De verwachting is dat ONIX voor bibliotheken relevant gaat worden zodra bibliotheken de bibliografische metadata direct van de uitgevers krijgen aangeleverd en deze rechtstreeks in de catalogus kunnen invoeren.

*URL:* <http://www.editeur.org/onix.html>

### 3 Protocollen voor gegevensuitwisseling

Protocollen beschrijven hoe gegevens tussen twee partijen worden uitgewisseld. Het gaat hierbij zowel om het formaat van uitwisseling als om de beschrijving van de respons die moet volgen op een specifiek request. In dit artikel gaat het alleen om protocollen die via het web gebruikt worden. In veel gevallen is de request een URL met parameters en is de respons in XML. Het is ook mogelijk dat de request-parameters niet via de URL verstuurd worden, maar via een zogenaamde POST opdracht. De reden om gegevensuitwisseling te standaardiseren, is dat men verschillende diensten met eenzelfde applicatie wil kunnen benaderen zonder voor elke applicatie te moeten weten hoe die applicatie benaderd moet worden.

#### 3.1 Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH is een protocol voor het kunnen ophalen dan wel beschikbaar stellen van in XML gestructureerde records op basis van de datum waarop de records gewijzigd zijn. Het is dus geen zoekprotocol zoals bijvoorbeeld

SRU (zie 3.5) omdat er geen zoekopdrachten gegeven kunnen worden. Het wordt gebruikt om data, meestal metadata, uit verschillende bronnen te kunnen vergaren (harvesten), in een eigen systeem op te slaan en in dat systeem doorzoekbaar te maken. Het maakt het mogelijk om met één applicatie op een uniforme manier gegevens uit verschillende bronnen op te halen zonder voor iedere bron weer opnieuw programmatuur te hoeven ontwikkelen.

Het protocol bestaat uit zes commando's: Identify, ListSets, ListMetadataFormats, ListIdentifiers, ListRecords en Getrecord. De laatste drie commando's worden gebruikt voor het daadwerkelijk harvesten. De eerste drie worden gebruikt om – meestal via een userinterface – gegevens omtrent de bron en de beschikbare collecties (sets) en formaten te achterhalen.

*Relevantie:* OAI wordt al veel gebruikt voor het harvesten van gestructureerde data en er zijn mij geen alternatieven bekend.

*URL:* <http://www.openarchives.org/>

### 3.2 *Niso Circulation Interchange Protocol (NCIP)*

NCIP is een standaard voor uitleentransacties zoals het uitwisselen van uitleengegevens, gegevens betreffende leners, items en de eigenaar van de items. Het gaat om zoeken, update en notificatie. Het is vooral bedoeld om vanuit verschillende (bibliotheek)systemen transacties te kunnen doen in andere bibliotheeksystemen. Vooral voor interbibliothecair leenverkeer kan dit bruikbaar zijn.

*Relevantie:* NCIP wordt nog niet veel gebruikt. In Nederland verloopt het interbibliothecair leenverkeer via het NCC/IBL-systeem van OCLC Pica, waarbij geen gebruik wordt gemaakt van NCIP.

*URL:* [http://www.niso.org/committees/committee\\_at.html](http://www.niso.org/committees/committee_at.html)

### 3.3 *Opensearch*

Opensearch is een nieuwe, zeer eenvoudige standaard voor search and retrieval, waarbij de respons in RSS-formaat is (voor RSS zie 3.6). Opensearch kent (nog) geen echte query-taal en is gebaseerd op het zoeken op losse woorden. Opensearch is een eenvoudig alternatief voor SRU en daarmee een rivaal, omdat keyword search in veel gevallen afdoende is. Er is echter een tendens om Opensearch meer mogelijkheden te geven.

*Relevantie:* Opensearch biedt niet de functionaliteit die voor geavanceerde

search and retrieval nodig is, maar is niet te negeren door zijn eenvoud en door het feit dat grote spelers als Amazon het gebruiken. Waarschijnlijk zullen Opensearch en SRU naast elkaar gebruikt worden en zullen zowel portals als databases beide moeten kunnen ondersteunen. Voor complexe queries is Opensearch niet bruikbaar.

*URL:* <http://opensearch.ag.com/>

#### 3.4 *Open Uniform Resource Locator (OpenURL)*

OpenURL is ontwikkeld om met een generieke of gestandaardiseerde URL-syntax te kunnen linken naar resources zonder de exacte URL-syntax van die resources te kennen. Dit gebeurt via een speciale server, de zogenaamde OpenURL resolver, die de gestandaardiseerde URL's vertaalt naar de URL's van de diverse resources. Een voorbeeld van het gebruik is dat men in Picarta een artikel vindt en dan doorgelinkt wil worden naar de full-text bij de eigen instelling. Vanuit Picarta wordt dan een URL gegenereerd met het adres van de OpenURL server van de gebruiker en de standaard OpenURL syntax. De OpenURL server van de instelling van de gebruiker zorgt dan voor het aanbieden van de daadwerkelijke link naar het artikel. Een URL bestaat in deze situatie uit twee delen: de hostlocatie en de metadata. De hostlocatie is een OpenURL resolver en is afhankelijk van de gebruiker; de metadata komen overeen met wat in de bibliografische beschrijving aan de gebruiker gepresenteerd werd. De OpenURL resolver kan afhankelijk van de context andere links aan de gebruiker aanbieden bijvoorbeeld een link naar Amazon indien er een ISBN in de metadata voorkomt. OpenURL versie 0.1 werkt zoals boven beschreven. Versie 1.0 is een stuk complexer en biedt ook andere mogelijkheden om metadata over te dragen. Een veel gebruikte commerciële implementatie van OpenURL is het pakket SFX van ExLibris. Dit pakket bevat een database met gegevens over de URL syntax van veel bibliotheekcatalogi en databases met tijdschriftartikelen.

*Relevantie:* OpenURL wordt zeer veel gebruikt om de gebruiker linkmogelijkheden te bieden naar eigen diensten. Nu ook de mogelijkheid geboden wordt om vanuit Picarta en Google Scholar naar de OpenURL server van de instelling van de gebruiker te linken, is de relevantie vrij groot geworden. Veel gebruikers zoeken niet meer eerst via de eigen instelling maar via Google Scholar. Door in Google Scholar de naam van de eigen instelling of bibliotheek op te geven, biedt Google in de resultaatpagina een link aan naar de OpenURL server van die instelling. Zo worden gebruikers direct

vanuit het Google resultaatscherm naar de artikelen van de eigen instelling verwezen.

*URL:* [http://www.niso.org/committees/committee\\_ax.html](http://www.niso.org/committees/committee_ax.html)

### 3.5 *Search and Retrieve via URLs (SRU)*

SRU is een protocol voor zoeken en opvragen. In SRU verloopt een zoekvraag via een gestandaardiseerde URL en is de respons in XML conform een standaard schema. Het voordeel van een gestandaardiseerd zoekprotocol is dat met één interface verschillende bestanden kunnen worden doorzocht. SRU vervangt Z39.50. Dit is een standaard uit de tijd dat er nog geen World Wide Web was. De voordelen van SRU ten opzichte van Z39.50 zijn de lage implementatiedrempel en het feit dat SRU helemaal gebaseerd is op webstandaarden en daardoor makkelijk te integreren is met andere webapplicaties.

SRU kent een klein aantal parameters en maakt gebruik van CQL (Common Query Language) als zoektaal. De SOAP-versie (zie 7.1) van SRU heet Search/Retrieve Web Service (SRW). SRW heeft als voordeel dat er geen beperking is aan de grootte en structuur van de input-parameters. Het nadeel is de grotere complexiteit en de hogere implementatiedrempel.

*Relevantie:* Zeer relevant voor alle websites die een zoekinterface hebben. Voor bibliografische gegevens wordt dit waarschijnlijk het standaard zoekprotocol.

*URL:* <http://www.loc.gov/standards/sru/>

### 3.6 *Really Simple Syndication (RSS)*

RSS is een verzamelnaam voor een paar standaarden (bijv. RSS 1.0, RSS 2.0). De afkorting staat ook voor Rich Site Summary of RDF Site Summary. In dit rijtje past ook Atom. Het dient om een lijst items weer te geven in een eenvoudig XML-formaat met titel, beschrijving, link, tijdstip enzovoorts. RSS wordt gebruikt voor items die telkens vernieuwd worden, zoals bijvoorbeeld nieuwsitems, aanwinsten, films op televisie e.d. Door te klikken op een RSS-link krijgt men steeds de meeste recente items aangeboden. Zo'n link wordt meestal een RSS-feed genoemd, die op webpagina's wordt weergegeven als een oranje knopje met de tekst RSS of Atom. Speciale programma's, zogenaamde feedreaders kunnen regelmatig zo'n link checken om te zien of er een item bijgekomen is en dan de gebruiker waarschuwen. Men kan ook op een gepersonaliseerde webpagina de meeste recente links uit een RSS-feed tonen. Gebruikers kunnen via RSS op een ef-

ficiënte manier nieuwe informatie van diverse websites laten verzamelen en zich laten attenderen zonder die websites telkens te bezoeken. Zo is het vaak ook mogelijk om een zoekactie als RSS feed te behandelen en geat-tendeerd te worden op nieuwe items in een database.

Zie voor een uitgebreide behandeling van RSS het artikel van J. van der Harst in dit handboek (IV B 610 Really Simple Syndication?).

*Relevantie:* RSS is hot. Het belang komt voort uit de eenvoud van het concept en het feit dat het een pure noodzaak is geworden om informatie gefilterd aangeboden te krijgen. Het wordt wijdverbreid gebruikt in allerlei websites. Gebruikers willen steeds vaker de mogelijkheid hebben om geat-tendeerd te worden op veranderingen in bijvoorbeeld nieuwsitems in plaats van dat zij zelf telkens moeten kijken of er nog iets veranderd is.

*URL:* <http://blogs.law.harvard.edu/tech/rss>

#### 4 Identificatie van objecten

Om naar digitale objecten te verwijzen en die later te kunnen terugvinden, is het gebruik van een 'gewone' URL meestal niet voldoende. Vooral omdat de locatie (de URL) kan wijzigen of er kopieën op meerdere locaties kunnen zijn, is het noodzakelijk een persistente identificatie te hebben van het object los van een fysieke locatie. Uiteraard moeten er wel databases beschikbaar zijn waarmee op basis van zo'n identificatie een of meerdere fysieke locaties van de gezochte publicatie of het gezochte object gevonden kunnen worden. Op basis van het type identificatie zal men dus ook informatie moeten hebben over de locatie van die databases. Dat wil zeggen dat het type identificatie als zodanig herkenbaar moet zijn. In veel gevallen wordt dit gedaan door het type identificatie als prefix in de identifier op te nemen, bijvoorbeeld ISBN:1 2345678 voor een ISBN. Om aan te geven dat het om een gestandaardiseerde identifier gaat, wordt door sommige data-leveranciers hier ook weer de prefix URN voorgezet. URN staat voor Uniform Resource Name. Voor het ISBN voorbeeld zou dit er uitzien als URN:ISBN:1 2345678.

##### 4.1 Archival Resource Key (ARK)

ARK is een standaard voor het bieden van een persistente identificatie van objecten. De identificatie maakt onderscheid tussen de identificatie van het object en de identificatie van de organisatie die het object aanbiedt en

de services waarmee de beschrijving van het object gevonden kan worden. Hiermee wordt tegemoet gekomen aan het probleem dat als een locatie wijzigt een URL niet meer geldig is. Op basis van de inhoud van de ARK kan echter een vervangende ARK resolver gezocht worden. Het is niet de identificatie die persistentie biedt maar de organisatie of de service die door middel van de identificatie de locatie van het object kan leveren. De Digital Libraries of California beheert het register met organisaties die ARK's toewijzen.

*Relevantie:* Veelal zal gebruik gemaakt worden van DOI en NBN (zie verderop) en rekent men op de organisaties die een vertaalslag kunnen maken van DOI of NBN naar een fysieke locatie. ARK zal zijn toegevoegde waarde eerst moeten bewijzen, temeer omdat het (nog) niet vaak genoemd wordt.

*URL:* <http://www.cdlib.org/inside/diglib/ark/arkspec.pdf>

#### 4.2 Digital Object Identifier (DOI)

Een DOI identificeert een digitale publicatie. Een DOI bestaat uit een prefix en een suffix, bijvoorbeeld 10.1000/1234-5678. De suffix identificeert het object zelf en wordt toegekend door de uitgever. De prefix identificeert de uitgever van de identificatie en wordt toegekend door de DOI organisatie. Via een zogenaamde resolution service kan doorverwezen worden naar locaties waar de objecten zijn opgeslagen. Een resolution service is een service die met de DOI als input een URL naar een fysieke opslaglocatie kan leveren. Voor DOI is Crossref de registratie autoriteit, die ook de resolution service levert. De DOI vervult een vergelijkbare rol als NBN (zie 4.3) of ISBN. DOI's worden vooral voor tijdschriftartikelen gebruikt.

*Relevantie:* Zeer relevant omdat de meeste nieuwe elektronische tijdschriftartikelen hiermee geïdentificeerd worden. DOI's worden ook veel gebruikt als identificatie in referenties in artikelen en de resolution services hiervoor mogen dus niet meer verdwijnen.

*URL:* <http://www.doi.org/>

#### 4.3 National Bibliographic Number (NBN)

Een NBN bestaat uit een landcode, instituutcode en objectidentificatie. De objectidentificatie kan zelf ook weer een gestructureerde hiërarchische indeling hebben. Uit NBN's valt af te leiden bij welke organisatie of organisatieonderdeel objecten zijn opgeslagen; de feitelijke resolutie naar een

statische URL wordt aan deze organisatie overgelaten. Omdat ook de landcode in de NBN is opgenomen kan een resolution service eventueel ook verwijzen naar de resolution services van andere landen.

Omdat publicaties in elektronische depots meestal niet al bij hun ontstaan een NBN hebben, kan het wenselijk zijn om de feitelijke identificatie van het object te baseren op de oorspronkelijke identificatie (bijv. DOI) of een MD5 checksum (dit is een soort handtekening via een speciaal algoritme). Deze checksum wordt afgeleid uit het object zelf en is uniek voor het object.

*Relevantie:* NBN's zijn relevant voor persistente identificatie van objecten in nationale depotsystemen omdat steeds meer depotsystemen dit gaan gebruiken. Van de depotsystemen zelf mag verwacht worden dat ze niet zo maar verdwijnen.

*URL:* <http://ietfreport.isoc.org/all-ids/draft-hakala-nbn-oo.txt>

## 5 Structurering van complexe data-objecten

Digitale objecten kunnen afhankelijk van hun type een verschillende structuur hebben. Voor het presenteren van die objecten is daarom informatie nodig over de structuur. Een boek kan bestaan uit hoofdstukken en alinea's, een gedicht uit coupletten en multimedia-objecten kunnen uit tekst, video, audio, beeldmateriaal bestaan. Standardisatie van de beschrijving van de structuur van objecten is nodig om applicaties in staat te stellen een complex object te kunnen verwerken of te presenteren. Door te voldoen aan een standaard is bij die presentatie een minimum aan voorkennis vereist. De standaarden richten zich daarbij wel op een bepaald type object. Voor een gedicht is een andere structurering gewenst dan voor een multimedia-object. Bovendien ontstaan standaarden veelal in een specifiek toepassingsgebied. Zij zijn dan ook afgestemd op dat toepassingsgebied.

### 5.1 MPEG21 DIDL (*Digital Item Description Language*)

Dit is een standaard voor het beschrijven van de structuur van samengestelde objecten, zoals bijvoorbeeld een gedigitaliseerd boek met een afbeelding van de voorpagina, per pagina een afbeelding en apart de tekst als resultaat van OCR (Optical Character Recognition). Een samengesteld object wordt dan beschreven op een vergelijkbare manier als een directory structuur: een object is een item of een component die weer kan bestaan

uit items en componenten analoog aan directories en files. Per item kan aanvullende informatie zoals een beschrijving, link of type worden vastgelegd.

Door de structuur van een samengesteld object in een MPEG21 record vast te leggen, worden de links naar alle onderdelen gebundeld en kan het geheel op verschillende manieren gepresenteerd worden.

Het ontbreekt in MPEG21 echter nog aan een standaard voor semantische informatie om de inhoudelijke rol van objecten vast te leggen, bijvoorbeeld dat een afbeelding een afbeelding van een pagina uit een boek is of dat een afbeelding als thumbnail bedoeld is.

*Relevantie:* Door de toename van gedigitaliseerde objecten die uit meer dan één file bestaan, is het steeds meer nodig om van samengestelde objecten de structuurinformatie op te kunnen slaan en te kunnen uitwisselen. MPEG21 is generieker dan andere structureringsstandaarden zoals METS (zie 5.2) en wordt steeds meer gebruikt.

*URL:* <http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm>

### 5.2 Metadata Encoding & Transmission Standard (METS)

METS is een op XML gebaseerde standaard voor het verpakken van diverse soorten metadata in één pakket om dit uit te kunnen wisselen met o.a. depotsystemen. METS-objecten zijn onderverdeeld in zeven specifieke segmenten, te weten:

- een header;
- administratieve metadata;
- beschrijvende metadata;
- een lijst met files;
- de structuur;
- een sectie met links;
- een sectie met een beschrijving van het ‘gedrag’ van onderdelen.

METS wordt vaak genoemd in relatie met elektronische depotsystemen waarbij een groep files als één pakket worden aangeboden vergezeld van een METS-record, waarin de totale inhoud van zo’n pakket is vastgelegd. Het verschil met MPEG21 (zie 5.1) is dat METS specifieke onderdelen beschrijft, terwijl MPEG21 meer generiek is.

*Relevantie:* METS als structurering voor de opslag en uitwisseling van elek-

tronische publicaties moet afgewogen worden tegen MPEG21. De voorkeur gaat uit naar MPEG21 omdat dit meer generiek is dan METS. Voor het herkennen van METS-objecten die door de buitenwereld worden aangeboden, is het wel van belang dat METS eventueel passief ondersteund kan worden omdat het veel wordt gebruikt.

*URL:* <http://www.loc.gov/standards/mets/>

### 5.3 *Text Encoding Initiative (TEI)*

TEI is een standaard om tekstdocumenten in XML te structureren. Het geeft aan wat een tekstonderdeel voorstelt, bijvoorbeeld een titel, gedichtregel, alinea enzovoorts. Als grote teksten in TEI worden aangeboden, kan de structuur gebruikt worden om binnen de tekst te navigeren of om een document op een specifieke manier te presenteren. Ook is het mogelijk om heel specifiek naar onderdelen te zoeken. Als teksten niet bij het aanmaken al volgens TEI gestructureerd worden, moet de omzetting naar TEI meestal handmatig gedaan worden.

*Relevantie:* Voor documenten die volgens TEI gestructureerd zijn, is het wenselijk die te kunnen interpreteren en van de structuur gebruik te maken. Met andere woorden: passieve ondersteuning van TEI kan zinvol zijn. Het is de vraag of het actief volgens TEI structureren van gedigitaliseerde teksten waarvan de structuur nog niet machine-interpreteerbaar is, een toegevoegde waarde heeft. Voor eigen gedigitaliseerd materiaal zal het niet waarschijnlijk zijn dat oorspronkelijke complexe structuren automatisch naar TEI omgezet kunnen worden. In veel gevallen kan dan volstaan worden met meer generieke formaten zoals PDF of MPEG21. Men moet zich altijd afvragen of omzetting naar TEI een meerwaarde biedt ten opzichte van ontsluiting en slimme presentatie van ongestructureerde fulltext.

*URL:* <http://www.tei-c.org/>

## 6 **Authenticatie**

Authenticatie is het met een zekere mate van zekerheid vaststellen van de identiteit van een gebruiker op grond van bijvoorbeeld inloggegevens, een chipkaart of biometrische kenmerken. Redenen om dit te standaardiseren heeft vooral te maken met het feit dat er veel organisatorische federaties zijn waarbij men na geauthenticeerd te zijn, toegang kan krijgen tot systemen van de instellingen binnen zo'n federatie. Ook binnen een instelling

kan het wenselijk zijn dit te standaardiseren. In het beste geval hoeft men zich slechts één keer bekend te maken om toegang tot al die aangesloten systemen te krijgen (single sign-on) maar soms beperkt het zich tot het gebruik van slechts één wachtwoord.

Er zijn federaties die binnen die federatie een standaard hebben gerealiseerd, bijvoorbeeld Athens in de Britse universitaire wereld. Omdat een instelling onderdeel kan zijn van verschillende federaties, is het wenselijk dat men van dezelfde authenticatiesoftware gebruik kan maken en met dezelfde mechanismen te maken heeft.

#### 6.1 SHIBBOLETH

Shibboleth is geen echte standaard maar software die het mogelijk maakt geautoriseerde toegang te bieden tot afgeschermd webdiensten voor gebruikers die door hun eigen instelling geauthenticeerd zijn. Hierbij worden gegevens over de gebruiker door een dienst naar de instelling van de gebruiker gestuurd en geeft de eigen instelling de toegangspersmissie door aan de vragende dienst. Deze gegevens zijn zodanig ingericht dat de privacy van de gebruiker beschermd wordt. Dat wil zeggen dat de door de gebruiker benaderde dienst alleen weet waar hij geauthenticeerd is en welke rechten hij heeft, maar zonder privacy gevoelige gegevens te krijgen. De gegevens die voor dit doel tussen systemen worden uitgewisseld, zijn gestructureerd in XML. De structuur van deze XML voldoet aan een standaard en heet SAML (Security Assertion Markup Language)

Een voorbeeld van een mogelijke toepassing kan The European Library zijn indien toegang tot publicaties afhangt van lidmaatschap van een aangesloten bibliotheek zonder dat The European Library de individuele gebruikers hoeft te kennen.

*Relevantie:* Dit protocol wordt zeer relevant omdat het het enige protocol op dit gebied is. Het Engelse Athens gaat Shibolleth ondersteunen en ook het in Nederland door Surfnet geadviseerde A-select pakket gaat Shibolleth ondersteunen.

*URL:* <http://shibboleth.internet2.edu/>

## 7 Protocollen voor communicatie tussen applicaties

Er zijn veel diensten die uniek zijn en daarom niet volgens een bepaalde standaard benaderd hoeven te worden. Toch is er behoefte aan om de com-

municatie tussen twee applicaties of diensten zo veel als mogelijk te standaardiseren. Dit betreft de beschrijving van de input parameters, de output en de foutafhandeling.

#### 7.1 *Simple Object Access Protocol (SOAP) en Web Service Definition Language (WSDL)*

SOAP is een mechanisme om via het web diensten met elkaar te laten communiceren. Dit speelt zich meestal onzichtbaar voor de gebruiker af tussen webservers en database servers onderling. SOAP is een standaard voor gegevensuitwisseling, die gebruik maakt van het http-protocol. Met andere woorden: in de basis communiceren twee applicaties met elkaar, zoals een webbrowser en een webserver met elkaar communiceren. De gegevens die uitgewisseld worden, worden verpakt in XML-pakketten in een specifiek SOAP-formaat. Hoewel men twee applicaties met elkaar via het internet kan laten communiceren als ze elkaars ip-adres weten, biedt SOAP diverse extra faciliteiten om dit meer gestroomlijnd te doen met bijvoorbeeld een standaard foutafhandeling e.d.

Een SOAP-service wordt gedefinieerd door middel van Web Service Definition Language (WSDL). In feite is WSDL een aparte standaard, maar deze wordt in de praktijk bijna nooit los gezien van SOAP. Met WSDL wordt beschreven wat de input en outputparameters van een webservice zijn. Hiervan kunnen andere applicaties gebruik maken om zo'n service te kunnen benaderen.

Omdat webbrowsers (nog) geen SOAP ondersteunen, is het soms noodzakelijk een centrale component te bieden tussen de webbrowser van de gebruiker en webservices van derden. Voor het zelf aanbieden van services op basis van SOAP moet men zich terdege afvragen of er een toegevoegde waarde is voor het gebruik van SOAP ten opzichte van gebruik van URL's. SRW (zie SRU/SRW) is een voorbeeld waarbij gebruik gemaakt wordt van SOAP tussen de database en de zoekapplicatie.

*Relevantie:* SOAP is wereldwijd zo in opkomst dat het wenselijk is dat instellingen in staat zijn hiermee bouwstenen van de ICT-infrastructuur te realiseren. Indien men een service wil kunnen benaderen die alleen via SOAP toegankelijk is, zal men SOAP in ieder geval passief moeten ondersteunen. Het actief aanbieden van services via SOAP heeft niet altijd een meerwaarde en zou zich moeten beperken tot communicatie tussen applicaties. Zie ook REST.

*URL:* <http://www.w3.org/TR/soap/>

### 7.2 REST (*Representational State Transfer*)

REST betreft webservices gebaseerd op alleen http ('SOAP light'). De basis wordt gevormd door URL's met GET, POST, PUT en DELETE. SRU is een voorbeeld van een zogenaamde 'REST alike', dat wil zeggen dat het op REST lijkt maar niet echt REST is. Vooral nog is het geen officiële standaard, maar meer een classificatie van een concept. De reden om REST te noemen is vooral omdat men vaak een afweging moet maken tussen SOAP (XML als input en XML als output) en geen SOAP (URL parameters als input en XML alleen als output). Vaak wordt de term REST gebruikt om aan te geven dat geen SOAP, maar gewoon URL's gebruikt worden. Het voordeel van REST ten opzichte van SOAP is, dat men een service eenvoudig met een browser kan benaderen en dat men een link naar een REST-service in een HTML-pagina kan aanbieden.

Service die gewoon via een URL met parameters wordt aangesproken, wordt in het gewone spraakgebruik al vaak een REST-service genoemd. De in dit artikel besproken standaarden zoals SRU en OAI-PMH, zijn voorbeelden hiervan.

*Relevantie:* Waarschijnlijk wordt dit even relevant als SOAP, maar dan meer voor toepassingen die rechtstreeks met de browser communiceren, bijvoorbeeld een database record update.

*URL:* <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>

## 8 Thesauri/ontologieën

De hieronder genoemde standaarden hebben gemeen dat ze relaties op een gestructureerde manier weergeven of daarvan gebruik maken.

### 8.1 Resource Description Framework (RDF)

RDF is een mechanisme om web resources te beschrijven. Wordt veelal in verband gebracht met het semantische web. Het moet vooral gezien worden als een XML-taal om relaties tussen resources en eigenschappen van die resources te leggen. Een volledige uitleg van RDF valt buiten de scope van dit document. Kort gezegd komt het erop neer dat in XML alleen gestructureerd kan worden en dat RDF een aanvulling is om te groeperen en specifieke relaties te leggen.

*Relevantie:* De meningen over RDF zijn nogal verdeeld als het gaat om het

semantisch web. Het wordt echter veel gebruikt als onderdeel van andere standaarden en is als zodanig niet weg te denken.

*URL:* <http://www.w3.org/RDF/>

### 8.2 *SPARQL Protocol and RDF Query Language*

SPARQL is een W3C standaard zoektaal die vooral bedoeld is voor RDF metadata. Het vertoont enige gelijkens met SQL en wordt veel gebruikt voor het zoeken met gebruik van relaties tussen velden. Het gebruik van deze zoektaal moet afgewogen worden tegen CQL, de zoektaal die bij SRU gebruikt wordt.

*Relevantie:* Een probleem bij het gebruik van SPARQL is de performance en de complexiteit. Vooralsnog wordt SPARQL niet of nauwelijks gebruikt in bibliotheekomgevingen en meer in de wereld van ontologieën.

*URL:* <http://www.w3.org/TR/2004/WD-rdf-sparql-query-20041012/>

### 8.3 *Zthes*

Zthes is een standaard XML-formaat voor thesaurusrecords. Een record bestaat uit een hoofdterm en diverse termen om verbanden aan te geven, zoals 'related', 'narrower', 'broader', enzovoorts. Zthes wordt door verschillende instellingen gebruikt en kan een (tijdelijke) oplossing zijn voor veel losse bestandjes met thesauri in een niet-standaard vorm. Door Zthes te ondersteunen kunnen we in onze portals met hetzelfde algoritme zowel eigen thesauri als die van anderen gebruiken. Ook kunnen we de resultaten uit thesauri van de ene aanbieder gebruiken om zoekvragen te formuleren voor databases van een andere aanbieder zonder dat deze afhankelijk van elkaar hoeven te zijn. De Zthes structuur kan ook gebruikt worden voor het leggen van andere typen relaties zoals vertalingen van onderwerpstrefwoorden.

*Relevantie:* Het is niet duidelijk hoeveel Zthes gebruikt wordt. Het is alleen bekend in bibliotheekomgevingen. Waarschijnlijk zal OWL (zie hieronder) het gebruik van Zthes inhalen omdat het generieker is en ook in andere omgevingen meer en meer gebruikt gaat worden. Vooralsnog is het een eenvoudige manier om eigen thesauri te structureren.

*URL:* <http://zthes.z3950.org/>

### 8.4 *Web Ontology Language (OWL)*

Web Ontology Language gebaseerd op RDF. Het dient als taal voor het leggen van relaties tussen begrippen. Voor een term kan via OWL vastgelegd

worden hoe een term zich relateert tot andere termen. Voor thesauri is momenteel het Zthes profiel het meest gangbaar voor het leggen van relaties tussen termen. OWL is aanzienlijk complexer dan Zthes.

*Relevantie:* De relevantie van OWL hangt af van de mate waarin de buitenwereld OWL gaat gebruiken. OWL lijkt een 'overkill' ten opzichte van Zthes.

*URL:* <http://www.w3.org/TR/owl-features/>