WHITE PAPER

OPSLAG EN UITWISSELING VAN METADATA

VOORSTEL VOOR EEN ANDERE AANPAK

ERIK SIEGEL
V1.0 - AUGUSTUS 2002

0. INHOUDSOPGAVE

0.1. GEBRUIK EN COPYRIGHTS

1. INLEIDING

Metadata is overal: Iedereen die wel eens een boek in een bibliotheek aan de hand van de catalogus heeft opgezocht heeft met metadata gewerkt. Metadata zijn gegevens óver dingen. Metadata beschrijft, identificeert en/of classificeert dingen zodat we ze makkelijker terug kunnen vinden of eenvoudiger kunnen indelen.
In onze tegenwoordige informatiemaatschappij is metadata belangrijk: Zonder metadata wordt het vrijwel onmogelijk om in de bestaande brij van informatie iets terug te vinden. Probeer je maar voor te stellen hoe moeilijk het zou zijn iets op het Internet terug te vinden zonder een zoekmachine die je een ingang geeft op metadata over webpagina's.

Ik ben al enige jaren werkzaam in de educatieve sector, een branche waarin metadatering van materiaal steeds belangrijk wordt. Er zijn binnen deze sector diverse initiatieven om metadata te standaardiseren, waarvan IMS (zie `www.imsproject.org`) op dit moment (najaar 2002) de belangrijkste is.
Wil je deze metadatering gaan gebruiken dan stuit je echter op een groot aantal praktische problemen die wat mij betreft samengevat kunnen worden in: De standaarden zijn te log te monolithisch en te ingewikkeld. Het zijn "alles of niets" standaarden: Je gebruikt hun velden en waarden, zo niet dan is de standaard onbruikbaar. Voor sommige zaken (zoals bijvoorbeeld taxonomieën) zijn hele complexe constructies verzonnen waardoor ook de bijbehorende software erg ingewikkeld wordt.

Uit onvrede hiermee is binnen een project bij uitgeverij ThiemeMeulenhoff een andere aanpak bedacht. Deze kan gekarakteriseerd worden als licht, modulair en wijzigbaar. Alle gegevens passen er in en er is ruimte voor uitbreidingen zonder gelijk het hele systeem te verlaten. In plaats van een monolithische aanpak is er gekozen voor een aanpak die bestaat uit deelstandaarden. Sommige daarvan zullen centraal gestandaardiseerd moeten worden. Anderen kunnen echter heel goed bedrijfsafhankelijk zijn. Conversie van en naar andere standaarden (zoals IMS) is mogelijk.

Doel van deze white paper is om deze andere aanpak onder de aandacht van geïnteresseerden te brengen en zo hopelijk de discussie erover los te maken. Doelgroep is iedereen die met metadata werkt en te maken heeft met de internationale initiatieven tot standaardisatie ervan. Enige achtergrond op het gebied van metadata en XML (met name in het laatste hoofdstuk) is noodzakelijk.

Deze white paper begint met de problemen die er zijn met de huidige metadata aanpak (hoofdstuk 2) en beschrijft vervolgens het voorstel voor een alternatieve aanpak (hoofdstuk 3). Als laatste (hoofdstuk 4) wordt een implementatie ervan behandeld, waarbij uiteraard XML gebruikt wordt.
Bron voor dit verhaal zijn mijn ervaringen in de educatieve branche. Het lijkt me echter dat dit ook in andere sectoren moet spelen.

2. METADATA IN GEBRUIK

Een voorbeeld uit de educatieve hoek: Steeds vaker zijn educatieve teksten, opgaven en ander materiaal opgeslagen in informatiesystemen, op een zodanige manier dat er meerdere malen gebruik van kan worden gemaakt. Bijvoorbeeld: Dezelfde tekst kan in meerdere boeken worden opgenomen; Een docent kan een proefwerk uit een bank met opgaven samenstellen; Materiaal kan worden overgebracht naar de elektronische leeromgeving van een school zodat het online beschikbaar wordt.
Om bruikbaar te zijn moet al dit materiaal worden voorzien van metadata over bijvoorbeeld het onderwerp, het lesstofniveau, studiebelasting, etc. Deze metadata wordt gekoppeld aan het materiaal zelf opgeslagen.
Bij uitwisseling van het materiaal (bijvoorbeeld naar een elektronische leeromgeving toe) gaat de metadata (deels) mee. Omdat deze systemen veelal van verschillende leveranciers zijn, zijn er verschillende pogingen gaande de metadata uitwisseling te standaardiseren. De belangrijkste daarvan in de educatieve sector op dit moment (najaar 2002) is IMS (zie `www.imsproject.org`).

Hieronder volgt een samenvatting van de metadata gegevens volgens IMS:

- **Metametadata**: Ook de afspraken over metadata zijn aan wijzigingen onderhevig. Je zult dus moeten vermelden welke versie van de metadata je hier gebruikt.
- **Algemeen**: Zaken als:
  - Unieke identificatie
  - Type van het materiaal.
  - Titel of naam van het materiaal
  - Taal (of talen) waarin de inhoud van het materiaal is opgesteld
  - Sleutelwoorden
- **Logboek**: In een logboek hou je de versie en gebruikshistorie bij. Wanneer (en door wie) is hij gemaakt, wanneer gewijzigd, wanneer geconverteerd, etc.
- **Technisch**: Zaken als:
  - Formaat, waarvoor is deze bedoelt
  - Omvang
  - Fysieke locatie
  - Technische opmerkingen over het gebruik ervan
  - Platform waarvoor het bedoelt is
- **Educatief**: Allerlei zaken die iets zeggen over het educatieve gehalte en de educatieve context waarin je het gaat gebruiken. Zaken als:
  - Vak/vaardigheid waar het bij hoort
  - Soort van interactie met de leerling
  - Bedoeld voor het volgende
  - Moeilijkheidsgraad
- **Rechten**: Informatie over de rechten die geregeld moeten worden bij het gebruik van het materiaal.
- **Relaties**: Relaties tussen verschillende onderdelen, bijvoorbeeld:
  - Moet altijd/bij voorkeur voor/na
  - Hoort bij deze begeleidende tekst
  - Gebruikt deze andere resources (beeld, geluid)
- **Opmerkingen**
- **Classificatie**: Een vorm van classificatie van het materiaal in mogelijk meerdere classificatie systemen.

De mij bekende metadata standaarden, waaronder die van IMS, hebben een “alles in één” aanpak. Men probeert in één standaard alle mogelijke metadata voor een bepaald vakgebied wereldwijd vast te leggen. Tegen een dergelijke aanpak zijn de nodige bezwaren aan te tekenen:

- Monolithisch: Het geheel is volledig monolithisch opgezet. Men beschrijft in één standaard alle metadata die de samenstellende commissie bij elkaar wist te brengen. Een wijziging op de metadata betekent ook gelijk een wijziging van de gehele standaard.
- Metadata velden en hun toegestane inhoud blijken vaak erg probleem of domein specifiek te zijn. Toch zul je “jouw” metadata, die meestal net niet helemaal hetzelfde is, moeten afbeelden op velden en waarden in de standaard.
  Voor veel velden is dit geen probleem: Een titel is een titel en een auteur een auteur. Echter: Buiten deze standaard velden zijn er een groot aantal waarvan de betekenis en/of interpretatie minder eenduidig is. Ook zit er vaak metadata in systemen die slechts met veel fantasie op de in de standaard beschreven velden is af te beelden.
  Bijvoorbeeld: ThiemeMeulenhoff wilde graag de docentbelasting voor een stuk educatief materiaal vastleggen. Hiervoor is in de IMS standaard geen veld aanwezig.
- Als je gegevens hebt die niet op de standaard metadata velden af te beelden zijn, zul je de standaard moeten uitbreiden. Vaak is hier wel een mechanisme voor aanwezig. In de IMS standaarden bijvoorbeeld kun je met behulp van een `<extension>` element extra velden toevoegen. Er is geen mechanisme om zo’n uitbreiding tot (sub)standaard verheffen.
  Als je hier echter eenmaal aan begint ontbreekt verder iedere vorm van standaardisatie en controle. Wat in zo’n uitbreidingselement wordt geplaatst is geheel afhankelijk van de afspraken tussen de partijen die het gebruiken.
- Standaarden zoals deze zijn opgesteld door commissies die zijn samengesteld uit vertegenwoordigers van meerdere bedrijven. Pas na eindeloos vergaderen ontstaat er consensus. Wijzigingen op de standaard komen daardoor slechts zeer moeizaam tot stand. Het is een bijzonder log mechanisme.
- Omdat geprobeerd is om in de metadata standaarden alles te vangen zijn ze hierdoor op onderdelen bijzonder complex. Hiermee wordt ook de implementatie van de standaarden in software bijzonder ingewikkeld.
  Neem bijvoorbeeld de IMS afbeelding van taxonomieën. Een taxonomie is een indeling van materiaal in categorieën, bijvoorbeeld de SISO code zoals ze in de openbare bibliotheken wordt toegepast. IMS beeld zo’n taxonomie af op een nogal ingewikkelde structuur. Deze is noodzakelijk zo complex omdat er ruimte moet zijn voor zo veel mogelijk soorten van taxonomieën.
  In de software zul je deze ingewikkelde structuren moeten ondersteunen, terwijl je maar één of hooguit twee taxonomieën gebruikt.

Het implementeren van metadata volgens de internationale standaarden is hiermee een frustrerende bezigheid geworden: Natuurlijk zie je het belang om je aan standaarden te houden en je doet vreselijk je best om alle gegevens in te passen. Je blijft echter met velden zitten die niet passen zodat je rare trucs moet uithalen of velden oneigenlijk moet gaan gebruiken. Aan de andere kant zijn er velden die je moet vullen maar die je helemaal niet nodig hebt. Door de complexiteit kost de software eromheen ook nog eens erg veel geld en moeite. En uiteindelijk is het dan nog maar de vraag of je door al deze afwegingen en keuzes de gewenste compatibiliteit tussen systemen inderdaad bereikt.
Het zijn deze frustraties die me tijdens een project aangezet hebben te proberen een flexibeler systeem voor metadatering te verzinnen. Een niet monolithisch, minder log en vooral makkelijker aan te passen manier om metadata uit te wisselen en te gebruiken, zonder daarbij uit het oog te verliezen dat we het hebben over standaarden. Het volgende hoofdstuk beschrijft dit systeem.

3. EEN ANDERE AANPAK

In het voorgaande hoofdstuk zijn de problemen met de bestaande aanpak van metadatering uiteen gezet. Dit hoofdstuk beschrijft een andere aanpak die deze bezwaren grotendeels ondervangt.

Bij het ontwerpen hiervan is het volgende aan aandachtspunten meegenomen:

- Je zult goed moeten kunnen omgaan met wijzigingen van de metadata in de tijd. Bijvoorbeeld een classificatiesysteem voor onderwijskundig materiaal zal veranderen als het onderwijs van richting verandert. Of: Het is nog niet zo lang gelden dat een typisch metadata item als een adres er een postcode bij kreeg.
- Een wijziging in de metadata mag niet uitsluiten dat een oudere vorm ook nog voorkomt. Dit heeft natuurlijk alles te maken met de traagheid van de software ontwikkeling eromheen.

  Stel het classificatiesysteem voor iets wijzigt en we passen de metadata standaarden hierop aan. Bestaande software die rekening houdt met het “oude” systeem is echter niet zomaar aangepast. We zullen in de metadata de oude codering naast de nieuwe codering moeten kunnen laten bestaan, zonder dat er hierdoor conflicten ontstaan.
- Metadata gegevens hebben allerlei verschijningsvormen. Zo zal metadata over de auteur vaak uit een enkel veldje bestaan (naam), maar metadata over het gebruik van materiaal (log informatie) uit een hele lijst van data en gebeurtenissen.
- Metadata, zowel losstaand als in een document, moet als zodanig herkenbaar zijn. De gegevens moeten zijn omgeven door een duidelijk herkenbare “envelop”. Deze envelop maakt deel uit van de standaard.
- Als de metadata wordt ingelezen en verwerkt zullen er elementen kunnen voorkomen die niet worden herkend. Deze elementen moeten kunnen worden genegeerd zonder dat daarmee de rest van de metadata ongeldig wordt.

3.1. METADATA KLASSEN

Als alternatief systeem voor metadatering wordt hier, in analogie uiteraard met een klasse uit de object oriëntatie, het begrip *metadata klasse* geïntroduceerd.

Een metadata klasse is een bij elkaar behorend brok metadata dat eenduidig geïdentificeerd wordt met een naam en een versie. Iedere metadata klasse kent zijn eigen data structuur voor het opslaan van de daadwerkelijke metadata gegevens.

Een aantal voorbeelden:

- Een informatief bibliotheekboek in Nederland heeft een SISO code. Zoiets zouden we in een metadata klasse als volgt op kunnen slaan:

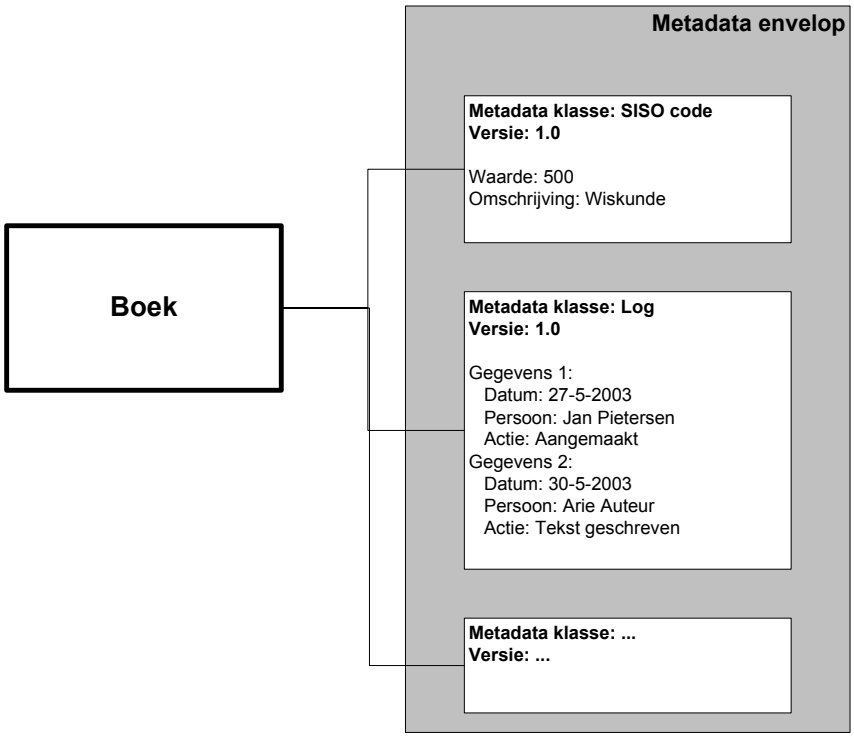
```
Metadata klasse:          SISO Code
Metadata klasse versie:   1.0
Bestaat uit een enkel decimaal nummer gevolgd door de
omschrijving. Bijvoorbeeld:
Waarde:          500
Omschrijving:    Wiskunde
```

- Als log informatie houden we bij wie een document heeft aangemaakt, gewijzigd, etc. Dit zouden we als volgt op kunnen slaan:

```
Metadata klasse:          Log
Metadata klasse versie:   1.0
Bestaat uit een lijst van structuren. Iedere structuur bevat een
datum, de naam van de uitvoerende en de actie die is uitgevoerd.
Bijvoorbeeld:
Gegevens 1:
      Datum:        27-5-2003
      Persoon:      Jan Pietersen
      Actie:        Aangemaakt
Gegevens 2:
      Datum:        30-5-2003
      Persoon:      Arie Auteur
      Actie:        Tekst geschreven
```

Aan iets wat we van metadata willen voorzien hangen we nu gegevens in de vorm van brokjes informatie behorend bij metadata klassen. Analoog aan de terminologie binnen object oriëntatie noemen we zo'n brokje informatie een *metadata object*. Een metadata object is altijd een instantie van een bepaalde metadata klasse.

Bijvoorbeeld de metadata objecten voor een boek zouden er als volgt uit kunnen zien:



Om al deze metadata objecten heen bevindt zich een gestandaardiseerde metadata envelop die het geheel bij elkaar houdt.

Een aanpak op deze manier heeft de nodige voordelen:

- Het systeem is flexibel met betrekking tot de datastructuren binnen een klasse. Er zijn geen principiële beperkingen wat je binnen een klasse zou kunnen opslaan.
- Het systeem breekt de metadata informatie op in kleinere onderdelen (componenten); Het is modulair.
- Uitbreidingen zijn mogelijk en wijzigen de basisstructuur niet.
- Als een metadata klasse wijzigt kunnen we deze wijzigingen eenvoudig opnemen: We maken een nieuwe metadata klasse (met in ieder geval een ander versie nummer) en voegen een object van deze klasse toe. Het oude object kan gewoon blijven bestaan (mits het nog valide is natuurlijk).

3.2. STANDAARDISATIE

Binnen een systeem als dit kun je heel goed gestandaardiseerde metadata mengen met eigen, bedrijfsafhankelijke, gegevens. Aan de standaardisatie kant zal er dan het nodige moeten gebeuren:

Ten eerste zal de envelop vastgelegd moeten worden: Hoe omring je metadata en maak je van losse metadata objecten één geheel. Wat staat er in/op zo'n envelop, hoe onderscheid je binnen de envelop de losse objecten van elkaar, etc.

Daarnaast zullen zo veel mogelijk metadata klassen gestandaardiseerd moeten worden. Waar mogelijk moeten hierover bindende afspraken gemaakt worden. Bijvoorbeeld metadata klassen voor:

- De "Dublin Core" metadata gegevens (zie `dublincore.org`).
- Boek taxonomieën zoals bijvoorbeeld de SISO code.
- Binnen branches: productcodes, bestelgegevens, etc.
- Log informatie.
- Een lijst van auteurs of medewerkers.

Om al deze klassen uit elkaar te houden zul je afspraken moeten maken over de naamgeving ervan. Dit is met name belangrijk als je eigen specifieke metadata klassen gaat toevoegen. De naam die je hiervoor kiest moet natuurlijk niet per ongeluk overeenkomen met de naam van een bestaande gestandaardiseerde klasse. Een systeem als de URI (Uniform Resource Identifier, zie `www.w3.org`) kan hier heel goed voor gebruikt worden.

4. AFBEELDING OP XML

Een systeem als hierboven omschreven is zeer geschikt om te worden gerepresenteerd met behulp van XML. Om met de deur in huis te vallen een voorbeeld van een metadata pakket zoals het er uit zou kunnen zien:

```
<env:metadata xmlns:env="http://www.mso.org/envelop" version="1.0">

      <env:object xmlns:siso="http://www.mso.org/siso"
         name="sisocode" version="1.0">
            <siso:code>500</siso:code>
            <siso:description>Wiskunde</siso:description>
      </env:object

      <env:object xmlns:log="http://www.mso.org/loginfo"
         name="loginfo" version="1.0">
            <log:record>
                  <log:date>2003-05-27</log:date>
                  <log:name>Jan Pietersen</log:name>
                  <log:action>Created</log:action>
            </log:record>
            <log:record>
                  <log:date>2003-05-30</log:date>
                  <log:name>Arie Auteur</log:name>
                  <log:action>Text written</log:action>
            </log:record>
      </env:object>

</env:metadata>
```

Binnen de afbeelding van metadata klassen op XML kan uitstekend gebruik worden gemaakt van het "namespace" mechanisme. Hierboven gebeurt dat drie keer:

- De envelop heeft zijn eigen namespace (`http://www.mso.org/envelop`).
- Voor iedere metadata klasse is ook een namespace vastgelegd:
  - `http://www.mso.org/siso` voor de SISO code
  - `http://www.mso.org/loginfo` voor de log informatie

Al deze namespaces zijn, zoals aan de naam te zien is, uitgegeven door een fictieve organisatie met als internet adres `www.mso.org` (Metadata Standards Organisation?). Bij iedere namespace hoort een XML schema dat de structuur van de inhoud verder vastlegt.

Om nog even wat verder te kijken: Rondom de metadata klassen en bijbehorende namespaces/schema's kun je een organisatie oprichten. Deze definieert de envelop en de belangrijkste klassen. Deelnemers aan het initiatief kunnen hun eigen klassen opsturen om zo centraal geregistreerd te worden. Er ontstaat een centrale databank voor metadata klassen die vanuit verschillende kanten en initiatieven gevoed wordt.