Quickscan naar mogelijkheden voor automatische metadatageneratie

Frank Benneker

Oktober 2006

Leerobjecten in de praktijk 6

samenwerken aan vernieuwing van het Hoger Onderwijs

DU Digitale Universiteit

Colofon

Quickscan naar mogelijkheden voor automatische metadatageneratie

Leerobjecten in de praktijk 6

Stichting Digitale Universiteit
Oudenoord 340, 3513 EX Utrecht
Postbus 182, 3500 AD Utrecht
Telefoon 030 - 238 8671
Fax 030 - 238 8673
e-mail buro@digiuni.nl

Auteur
Frank Benneker

Copyright

Datum
Oktober 2006

Distributie
De serie Leerobjecten in de praktijk wordt verspreid via de volgende erkende vakwebsites:

- www.du.nl/leerobjecten
- dspace.ou.nl
- www.sco-kohnstamminstituut.uva.nl
- www.hbo-kennisbank.nl.
- elearning.surf.nl

Inhoudsopgave

Inleiding

Voor ontwikkelaars van onderwijsmateriaal is het invullen van de metadata een daad van altruïsme: zelf zullen ze er geen baat bij hebben, alleen anderen hebben dat. Hun winst is er uitsluitend in gelegen dat, als de anderen ook braaf metadata invullen, zij zelf op enig moment weer van die inspanning kunnen profiteren. Onder dergelijke condities zal men slechts bereid zijn het allernoodzakelijkste te doen. De set van metadatacategorieën moet dus zo klein mogelijk zijn, waar mogelijk moeten gesloten keuzelijsten per categorie beschikbaar worden gesteld, en het instrument waarmee de metadata ingevuld kunnen worden, moet zo gebruikersvriendelijk mogelijk zijn. Die gebruikersvriendelijkheid richt zich niet alleen op bekende kenmerken als een aantrekkelijke user interface. Het invullen van metadata wordt ook veel aantrekkelijker als gegevens die in het kader van de definitie en aanvraag van een project zijn opgesteld, hergebruikt blijken te kunnen worden als metadata. Meer in het algemeen, wanneer het beschikbaar stellen van metadata geïntegreerd wordt in het gehele leerobjectontwikkelproces, zal dat de kwaliteit van de geleverde metadataset ten goede komen. Instrumenten voor het genereren van metadatabeschrijvingen zullen voorts ook geïntegreerd moeten worden in het leerobjectontwikkelproces waarin het ontwikkelde materiaal wordt opgeslagen en beheerd. Zo kan de integriteit van het materiaal en de ermee verbonden metadata beter gegarandeerd worden. Ook vergemakkelijkt dit de taak van het ontwikkelen van de zoekinstrumenten waarmee ontwikkelaars de learning object repository (LOR) zullen doorzoeken.

Er zijn een aantal manieren om grote hoeveelheden informatie te ontsluiten. Er zijn bijvoorbeeld de Google-achtige zoekmachines waar je ‘gewoon’ vrij een aantal woorden invoert en zoekt. Deze zoekmachines hebben echter ook een meer geavanceerde zoekoptie waarbij de trefwoorden specifieker ingevoerd worden (bijvoorbeeld alleen pagina’s in een bepaalde taal). Bij het uitsplitsen van de zoekopties kan er gekozen worden voor verschillende strategieën, bijvoorbeeld een wizard-achtige dialoog die de gebruiker stap voor stap leidt naar een kleinere set resultaten, er kan gekozen worden voor vrije tekstvelden of juist velden met een keuzelijst met een beperkt aantal opties enzovoorts. Zoekstrategieën vallen buiten de scope van deze quickscan maar er is wel een directe relatie met de metadata. De kwaliteit van de metadata is bepalend voor het te vinden resultaat. Ook Google is afhankelijk van de kwaliteit van de eigen metadata. Het investeren in een goed metadataproces en in de kwaliteit van de metadata is één van de belangrijkste voorwaarden voor een succesvolle LOR en het hergebruik van leerobjecten in het algemeen.

De DU en metadata voor leerobjecten

Making metadata go away: hiding everything but the benefits is de titel van een artikel van Duval en Hodgins (2004) dat mij uit het hart gegrepen is. Hoewel ik een relatieve nieuwkomer ben in de discussie over metadata voor leerobjecten (sinds 2000 ben ik vanuit mijn werk betrokken) lijkt de discussie soms eindeloos en niet gericht op concrete en bruikbare resultaten. Deze quickscan zal proberen een stapje op weg naar concrete oplossingen te zijn. We streven er niet naar om de discussie over te doen maar de DU te wijzen en te laten bouwen op goed werk dat door anderen is verricht.

Het metadateren van leerobjecten volgens de DU-richtlijn is een complex en tijdrovend proces. Bij het ontwikkelen van kleine hoeveelheden leerobjecten is de tijdsinvestering nog te overzien maar bij grotere hoeveelheden zal het metadateren steeds meer resources vergen.
Een normaal metadatarecord voor een leerobject, volgens de DU-richtlijn, kost al snel één uur om aan te maken. Stel dat we een bestaande collectie van 1500 leerobjecten willen beschrijven volgens de DU metadatarichtlijn dan is dit al gauw één manjaar werk. De conclusie is dus gerechtvaardigd dat het uitgebreid handmatig metadateren van grotere collecties leerobjecten een

moeilijk tot onmogelijk opschalingsproces is. Toch zijn metadata een noodzakelijk gegeven voor het vastleggen van essentiële kenmerken die het werken en vinden van leerobjecten mogelijk moet maken. Alternatieven voor het handmatige (menselijke) metadataproces zijn meer dan wenselijk. Een mogelijkheid is om de hoeveelheid metadata die door een persoon met de hand moet worden ingevuld sterk terug te brengen door gebruik te maken van “slimme” software die metadata automatisch kan generen.

Het doel van deze quickscan is in kaart te brengen of dergelijke “slimme” software bestaat en of deze bruikbaar is in de context van de Digitale Universiteit. Het is niet de bedoeling om deze software uitgebreid te testen en te komen tot een gedetailleerde vergelijking van deze producten. De quickscan is een korte verkenning en niet het eindpunt. Het brengt slechts de mogelijkheden in kaart die het metadateren van leerobjecten kunnen vereenvoudigen. De quickscan wil een bijdrage leveren aan een nauwere samenwerking tussen de diverse organisaties in Nederland die repositories voor onderwijsmaterialen inrichten en worstelen met het metadatawerkproces.

De DU metadatarichtlijn (Werkgroep DU-metadata richtlijn, 2004) die uitgaat van de Nederlandse versie van IEEE LOM is de basis van de metadata voor deze quickscan.

Een korte analyse

Waar gaat het om bij het automatisch generen van metadata voor leerobjecten? Deze paragraaf beschrijft enkele technieken en concepten die een rol spelen in dit proces.

Metadata extractie en automatische classificatie

Het is van belang om het onderscheid tussen metadata-extractie en het automatisch classificeren specifiek te benoemen. Het zijn twee verschillende invalshoeken voor het automatisch genereren van metadata.

- Extractie
  Op basis van algoritmen wordt informatie uit een document geëxtraheerd en ondergebracht in de relevante metadatavelden. Denk bijvoorbeeld aan herkenning van de taal en het toekennen van sleutelwoorden. Er bestaan goede en beschikbare oplossingen voor deze benadering. De toegepaste technologie is laagdrempelig en past gangbare en algemene concepten uit de computertechnologie toe.

- Classificatie
  De tool die automatisch classificeert beschikt over relevante domeinkennis. Deze domeinkennis wordt toegepast bij de analyse van het object, bijvoorbeeld een leerobject over het hart kan worden geclassificeerd met behulp van medische domeinkennis. De tool levert een domeinspecifieke classificatie. Belangrijk is de constatering dat dergelijke tools eerst een training moeten ondergaan met een set van documenten die een specifiek domein goed beschrijven. Er is sprake van een hoge investering in het begin van het proces. Winst is duidelijk te behalen bij grotere aantallen van te metadateren objecten. De toegepaste technologie komt voort uit onderzoekswerk uit de wereld van de kunstmatige intelligentie.

Wie metadateert?

Liddy (2005) beschrijft drie hoofdscenario’s voor het creëren van metadata:

1. metadata wordt aangemaakt door metadataspecialisten, een functie die men over het algemeen aantreft in bibliotheekorganisaties.
2. de metadata wordt aangemaakt door de eigenaar/ontwikkelaar van een leerobject.
3. de metadata wordt automatisch gegenereerd door software tools.

Methode (1) wordt als mogelijkheid ook genoemd door Hermans en De Vries (2006). Daarnaast onderscheiden Hermans en De Vries de situatie waarin géén metadata zijn voorgeschreven en waarin gebruikers vooral afgaan op later toegevoegde annotaties en ratings.
Een combinatie van de drie scenario's van Liddy is natuurlijk de voor hand liggende optie die in de praktijk wordt aangetroffen. De DU adviseert bijvoorbeeld in de metadatarichtlijn een combinatie van scenario (1) en (2). Liddy onderschrijft de aanname in de inleiding dat het handmatig metadateren kostbaar en arbeidsintensief is.[1]

Bruikbare metadata versus perfecte metadata
In een studie van het Telematica Instituut (Huijsen, Grootveld, Brussee, Setten & Porskamp, 2005) wordt op enkele plekken terecht opgemerkt dat men niet zonder meer blind kan varen op automaten om metadata te genereren[2]. Iedereen kent de voorbeelden om Google om de tuin te leiden. Bijvoorbeeld zoeken op de termen "raar kapsel" (d.d.10 januari 2006) levert nog steeds de homepage van de premier als eerste resultaat op. Een kritische kanttekening is op zijn plaats. Indien ons doel perfecte metadatageneratie is, die in elke situatie dezelfde metadata oplevert, stellen we een doel dat niet realiseerbaar is. De technieken waar wij naar op zoek zijn worden toegepast in een beperkt domein (het onderwijs) en worden toegepast op leerobjecten die door ontwerpers naar eer en geweten worden ontwikkeld. De context is duidelijk en de gebruikers van de metadatatools zijn bekend. In mijn ogen betekent dit dat de ontwerpvoorwaarden voor metadata-extractie- en -classificatietools eenvoudiger te realiseren zijn dan wanneer een generieke tool voor een willekeurig object in een onbekende context moet worden gebouwd. Blind varen op technologie is onverstandig maar we hoeven het ook niet moeilijker te maken dan het al is

De samenstelling van leerobjecten
Leerobjecten bestaan doorgaans niet uit één soort materiaal. Het zijn niet alleen teksten of alleen beelden. Een gemiddeld leerobject bestaat uit een combinatie van digitale materialen met elk hun eigenschappen en (on-)mogelijkheden voor het automatisch genereren van metadata.
Bij automatische generatie (extractie en classificatie) van metadata zijn de mogelijkheden die het bronmateriaal biedt van essentieel belang.

- Tekstgeoriënteerd materiaal
  Documenten die voornamelijk uit tekst bestaan bieden de beste mogelijkheden voor metadatageneratie. Ruime ervaring en expertise is opgedaan om dit te realiseren. Een belangrijke constatering is het feit dat de meeste algoritmen ontwikkeld zijn voor teksten in de Engelse taal. Voor het Nederlandse taaldomein bestaan er niet zoveel goede algoritmen. Naar de mogelijkheden voor andere taaldomeinen is geen onderzoek gedaan. Voor het Nederlandse taaldomein ligt hier een mooie onderzoekstaak met een doelgroep die graag gebruik wil maken van de resultaten.

- Multimedia materiaal
  Video-, beeld- en geluidfragmenten vormen een groeiend onderdeel van leerobjecten. Het automatisch genereren van metadata uit dergelijke bronmaterialen is niet zover ontwikkeld als bij teksten. Ook zijn de resultaten nog niet goed genoeg. Het werkproces wordt in meerdere stappen doorlopen, bijvoorbeeld spraak in video of geluid wordt omgezet in een tekstdocument[3]

[1] Liddy (2005): "However, manual metadata assignments, like cataloging is a labor-intensive and costly function, requiring special knowledge and training."
[2] Huijsen et al. (2005): "Dit verschil tussen de menselijke leesbare en aparte metadata kan niet alleen leiden tot pijnlijke onthullingen, het betekent helaas ook dat automatische metadata-extractie niet blind kan varen …"
[3] Zie bijvoorbeeld http:///www.blinkx.com.

(dit gaat al vrij aardig). Dit halffabricaat is de basis voor metadatageneratie. Ook wordt er gewerkt met algoritmen voor patroonherkenning. Helaas zijn deze technieken (nog) niet geschikt voor grootschalige toepassing. In de onderwijspraktijk zal in de meeste gevallen metadata aanwezig zijn. Het is bijna onwerkbaar als dergelijke objecten niet van de meest rudimentaire metadata worden voorzien. Een goed voorbeeld zijn de afbeeldingen die men maakt met een digitale fotocamera. Deze worden in de meeste gevallen automatisch met metadata beschreven (in het EXIF format[4]). Het omzetten van een metadatabeschrijving in de vorm x naar de vorm y (van EXIF naar LOM) is een relatief eenvoudige procedure. In het DU project Rechten Online is iets soortgelijks gedaan om materiaal vanuit de EML taal om te zetten naar het IMS QTI format. Dergelijke omzettingsalgoritmen moeten voor specifieke domeinen en doelen ontwikkeld worden.

- Samengesteld materiaal (complexe objecten)
  Leerobjecten zijn in veel gevallen opgebouwd uit meerdere elementen, bijvoorbeeld een tekst, een opdracht en enkele plaatjes (zie voor een overzicht van het hele scala aan mogelijkheden Schoonenboom, 2006). Elk onderdeel kan worden beschreven met behulp van metadata. Een voorbeeld van dit laatste is het complexe leerobject dat wordt besproken in Poortman en Sloep 2006. Er is duidelijk winst te behalen als de elementen binnen een dergelijke collectie (leerobjecten) relevante metadata zouden kunnen overerven (hergebruiken). In deze quickscan ben ik nog geen goede oplossing tegen gekomen. Verder onderzoek is nodig. Ook vanwege de mogelijkheden die hier liggen in combinatie met het operationaliseren van context- en profielinformatie.

Waar bevindt zich de metadata voor extractie?

Duval en Hodgins beschrijven in een aantal artikelen de mogelijkheden en concepten achter de extractie van metadata uit leerobjecten. Een vorm van zelfbeschrijving en deze vastleggen in de relevante velden van de LOM. Zij zien dit als een belangrijk onderzoeksdomein. In het paper *A LOM Research Agenda* wordt beschreven op welke wijze (semi-)automatische metadata te generen is.

- Extractie van metadata uit het leerobject
  Veel informatie die men wil opnemen in de metadata is aanwezig in het object zelf. Er zijn diverse technieken mogelijk om deze informatie uit het object te halen, bijvoorbeeld algoritmen voor taalherkenning en samenvatting, patroonherkenning bij afbeeldingen. Er bestaan HTML-scrapingtechnieken voor webpagina's en ActiveX-componenten om informatie uit MS-Officedocumenten te extraheren. In Officedocumenten bevinden zich meerdere metadata (auteur, datum, versie, enzovoorts). De metadata wordt opgebouwd uit de profielinformatie die in bijvoorbeeld MS Word wordt opgeslagen. Voorwaarde is wel dat de profielinformatie correct is.

- Overerven van metadata, gebruik maken van metadata van gerelateerde leerobjecten
  Leerobjecten maken vaak deel uit van een verzameling van objecten. Het ligt voor de hand om gemeenschappelijke metadata te laten overerven (hergebruik van metadata).

- Context- & profielinformatie
  Over de personen en cursussen (modulen) is in andere informatiesystemen in een instelling, bijvoorbeeld online studiegidsen, directory services en HRM-systemen, veel relevante informatie beschikbaar. Het operationaliseren van deze informatie in de metadata workflow zodat deze informatie automatisch in metadata wordt verwerkt, kan de nodige winst opleveren. Het AMG-

[4] Zie http://www.exif.org/

framework geeft enkele voorbeelden hoe een dergelijk netwerk van het aanleveren van metadata kan werken.

Behalve deze beschrijvende metadata introduceren Duval en Hodgins het concept van “social recommendation” voor het vastleggen van gegevens over leerobjecten die bij uitstek interessant[5] zijn voor mensen die op zoek zijn naar bruikbare leerobjecten. Deze vorm van metadata wordt na het gebruik toegevoegd en is een continu proces. Een zeer bekend voorbeeld is Amazon[6]. Deze “feedback”-metadata is bij uitstek niet automatisch te genereren omdat het persoonlijke interpretatie en waardering weergeeft. Onderdelen zijn wel automatisch aan te maken. Een repository kan dergelijke feedbackmechanismen ondersteunen, bijvoorbeeld door het inloggen van de persoon die de feedback geeft komt profielinformatie (naam, achtergrond enzovoorts) beschikbaar; deze kan automatisch worden ingevuld in een feedbackformulier. Ook procesinformatie als versie en datum is automatisch te genereren.

Het AMG-framework dat in Leuven wordt ontwikkeld is een concrete uitvoering van de voorstellen die door Duval en Hodgins zijn gedaan.

Ontwikkelingen in Nederland en België

Het automatisch genereren van metadata en classificeren van leerobjecten staat bij meerdere organisaties in Nederland en België op de agenda. Ook in andere landen wordt er veel onderzoek gedaan naar deze technologie. Voor de quickscan is besloten om een drietal initiatieven in Nederland en België kort te beschrijven, omdat daar mogelijkheden liggen voor de Digitale Universiteit voor samenwerking.

Het SURFnet onderzoek

In 2005 heeft SURFnet door Quo Vide een onderzoek laten verrichten naar categorisatiesoftware. Het doel van dit onderzoek was onder meer om de dienstverlening van de SURFnet zoekmachine te verbeteren met software die kwalitatief automatisch metadata kan genereren en bronnen kan classificeren. Dit onderzoek kenmerkt zich door een goede marktanalyse van beschikbare commerciële oplossingen. Wat men constateerde is dat commerciële oplossingen kostbaar tot zeer kostbaar zijn. Prijzen voor een jaarlijkse licentie lopen uiteen van 50.000 tot 1 miljoen euro. Dergelijke producten liggen buiten het bereik van de gemiddelde onderwijsinstelling.

In de eindrapportage wordt uitgebreid stilgestaan bij de diverse technologieën die worden toegepast en wordt de werkwijze van de diverse algoritmen kort beschreven. Het Telematica Instituut heeft in haar onderzoek in opdracht voor Kennisnet de meeste elementen voor haar domeinbeschrijving direct overgenomen uit dit onderzoek. Voor de Quickscan gaat het te ver om dieper in te gaan op deze zeer boeiende materie en wordt volstaan met enkele verwijzingen in de literatuurlijst naar de eindrapportages op de website van SURFnet. De eindconclusie van het onderzoek van Quo Vide is gebaseerd op een “proof-of-concept” waaraan drie leveranciers, die zijn geselecteerd uit de long list (bijlage 1), hebben deelgenomen.

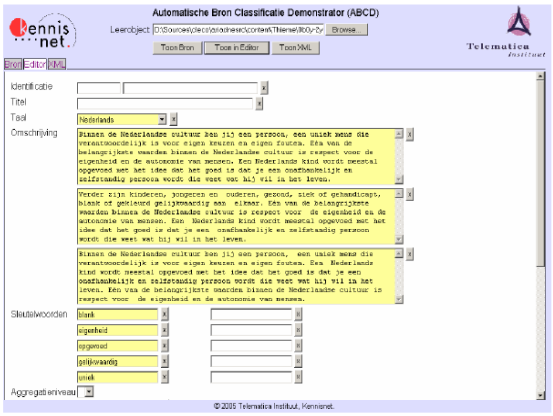
[5] Duval & Hodgins (2004): “we should be able to make use of information about how the learning object helped the user and organization to achieve the goal in affective and efficient way”.

[6] http://www.amazon.com

> Gezien de uitkomsten van de proof-of-concept is het niet mogelijk om een goed beeld te vormen van de technologische stand van automatische classificatie op dit moment. De proof-of-concept heeft helaas onvoldoende inzicht gegeven in de kwaliteit van automatische classificatie van diverse leveranciers. Wel is duidelijk geworden dat economisch gezien automatische classificatie op dit moment onvoldoende meerwaarde biedt voor de SURFnet FAST search engine. De kosten voor het inrichten, organiseren en onderhouden van de categorisatiesoftware wegen naar alle waarschijnlijkheid niet op tegen de baten.
>
> *Conclusie uit het onderzoek van Surfnet uitgevoerd door Quo Vide (dec 2005)*

De ABCD Demonstrator

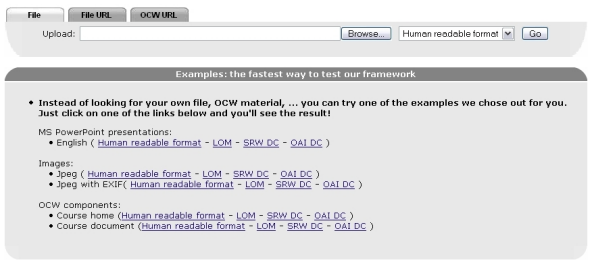
Het Telematica Instituut heeft in opdracht van Kennisnet een onderzoek uitgevoerd naar de mogelijkheden van automatische classificatie. De strekking van dit onderzoek wordt op een andere plek in deze quickscan besproken. Een belangrijk onderdeel van de opdracht van Kennisnet was het ontwikkelen van een demonstrator, de ABCD-demonstrator. Deze demonstrator laat zien dat mogelijk is om met beperkte middelen en tijd een bruikbare oplossing te ontwikkelen die het proces van metadateren kan ondersteunen. Een volledige automaat werd niet gerealiseerd maar de tool leverde aanvaardbare resultaten op basis van het suggestieconcept, dat wil zeggen er wordt een suggestie gedaan wat de metadata kan zijn voor een aantal van te voren vastgestelde metadatavelden. Metadata moet nog door een persoon geverifieerd en eventueel aangepast worden. De ABCD-demonstrator is helaas niet beschikbaar voor algemeen gebruik en navraag bij Kennisnet heeft nog geen antwoord opgeleverd op de vraag of Kennisnet deze tool beschikbaar stelt aan derden voor verdere ontwikkeling.



Figuur 13: Automatisch geëxtraheerde metadata in de demo metadata-editor.

Automatic Metadata Generation (AMG)

Een onderzoeksgroep van de Katholiek Universiteit Leuven is al enkele jaren zeer actief in het vakgebied van automatische metadatageneratie. Men heeft een conceptueel raamwerk ontwikkeld om diverse aspecten die een rol spelen bij het genereren van metadata een plaats te geven. Dit metadataraamwerk is de basis van een software-oplossing die in Leuven ontwikkeld is om daadwerkelijk automatisch metadata te genereren. De metadata wordt op basis van extractietechnieken uit de leerobjecten verkregen en tevens wordt informatie over de context waarin de leerobjecten zijn ontwikkeld en worden gebruikt, toegepast bij de generatie van een metadatarecord. Er wordt op zeer beperkte schaal gebruik gemaakt van algoritmen die eigenschappen zoals bijvoorbeeld de taal waarin het object is geschreven, herkennen. Er is in het



AMG-framework geen sprake van classificatie-algoritmen die wel worden toegepast bij de ABCD-demonstrator. De software is niet getest in deze quickscan. Wel is er een korte test gedaan met de online beschikbare demomodule via de website (http://ariadne.cs.kuleuven.ac.be/amg) van het AMG-project. Deze module biedt mogelijkheden om voor een beperkte categorie objecten metadata automatisch te generen in meerdere metadataschema's, zoals IEEE, LOM en Dublin Core. Men heeft onder andere de software ingezet om leerobjecten uit Blackboard te beschrijven met metadata en deze in een andere repository te ontsluiten. In het voorjaar van 2006 is een volledige herziene versie van de software verschenen. De software is beschikbaar voor derden en kan worden opgehaald via de website. De AMG-software is veelbelovend en verdient nadere bestudering in een vervolgproject.

Alternatieven voor metadatabeschrijvingen

Het invullen van de metadataset door ontwikkelaars voor het beschrijven van onderwijsmateriaal is niet de enige manier om op een zinvolle manier onderwijsmateriaal te ontsluiten. Het doel van de quickscan was niet om een volledig beeld te geven maar een korte momentopname. Het kort typeren van enkele alternatieven voor een metadatarecord (beschrijving) biedt wellicht enkele aanknopingspunten voor verder onderzoek naar andere opties om leerobjecten te zoeken en te vinden.

De Google benadering

Metadata in de vorm van LOMrecords is niet de enige wijze om leerobjecten vindbaar te maken. De technologie voor indexeringen en het vindbaar maken die wordt toegepast door zoekmachines zoals Google en Yahoo (bijvoorbeeld een Google zoekmachine voor leerobjecten) is niet onderzocht. In vervolgstudies is het raadzaam ook naar deze mogelijkheden van ontsluiten te kijken. Een interessant, toekomstig experiment is bijvoorbeeld een ontsluiting van LOREnetobjecten door scholar.google.com.

Social tagging als alternatieve benadering

Het klassieke traject van metadateren is een proces dat gekenmerkt wordt door een sterk a-priori formalisme. Alle mogelijke handelingen zijn beschreven en uitgewerkt met als doel een eenduidig en correct metadatarecord te genereren. In de praktijk wordt dit soms als beknellend en als een last ervaren. Nieuwe ontwikkelingen in het domein van social software bieden een alternatief voor dit strakke formalisme. Men creëert nog steeds metadata maar dit gebeurt op basis van persoonlijke intuïtie en inzichten. De filosofie achter dit sociaal netwerk is dat men de eigen objecten beschrijft en in het proces een consensus bereikt over de hanteren termen en begrippen. De consensus wordt bepaald door wat gangbare begrippen zijn in de alledaagse communicatie tussen mensen. Een afbeelding van een hond wordt 99 van de 100 keren gelabeld met een "tag" die iets te maken heeft met een hond of hondensoort. Metadata in een sociaal netwerk streeft niet naar 100% correctheid op basis van formele afspraken maar is een afspiegeling van wat de groep gebruikers een aanvaardbare beschrijving vindt. Een dergelijke verzameling van tags die door een netwerk van gebruikers ontstaat wordt een folksonomie genoemd, als alternatief voor de formele taxonomie die in principe wordt opgesteld door informatiespecialisten en bibliotheekorganisaties.

Dit is een ontwikkeling van de laatste jaren die zeer populair is bij de early adopters op het Internet. Meer onderzoek is nodig of een dergelijke manier van het beschrijven (metadateren) van objecten succesvol is bij het beschrijven, vinden en hergebruiken van leerobjecten.

Wie heeft er recht op de metadata?
In de discussie over metadata wordt vaak vergeten dat correcte metadata een commerciële waarde kan vertegenwoordigen. Metadata is de toegang tot de objecten en is bepalend voor het succes van een repository. Derhalve moeten er keuzes worden gemaakt met betrekking tot het eigendoms- en gebruiksrecht van de metadata. Wie heeft er toegang en wat mag men doen met de metadata? Een voorbeeld uit een andere praktijk zijn de bestanden met e-mailadressen en profielinformatie voor direct marketing. Een tweede voorbeeld is de AH-Bonuskaart. Het is belangrijk om bij het opzetten van een netwerk met leerobjecten niet alleen de rechten op het object maar ook de rechten op de metadata goed te regelen.

Sleutelartikelen & rapporten

Voor deze quickscan is gebruik gemaakt van diverse bronnen. Vier daarvan vormen de basis van deze quickscan. Deze bevatten interessante informatie voor verder onderzoek.

Huijsen et al. (2005). Automatische Classificatie Eindrapport, Telematica instituut.
Het onderzoek dat in opdracht van Kennisnet door het Telematica instituut is uitgevoerd richtte zich op het domein van automatische classificatie. Het onderzoek geeft een inzicht in de diverse mogelijkheden en problemen die de huidige stand van zaken van de techniek weerspiegelen. Het rapport kenmerkt zich door een grondige en klassieke benadering van de problematiek. Diverse technologieën en strategieën worden kort besproken. Men concludeert onder meer dat er voor het Nederlandse taaldomein weinig geschikte algoritmen beschikbaar zijn. In het onderzoek heeft men gebruikt gemaakt van een onderzoek dat SURFnet heet laten uitvoeren naar commerciële en open-source-oplossingen.

Naast de domeinanalyse was een onderdeel van het onderzoek om een demonstrator te ontwikkelen die op basis van het Kennisnet zoekprofiel bruikbaar is in het metadatawerkproces. De demonstrator maakt onder meer gebruik van elementen die ontwikkeld zijn binnen het AMG-framework uit Leuven. Dit is goed gelukt en de demonstrator biedt zeker aanknopingspunten voor verder onderzoek. Het verdient aanbeveling om met Kennisnet en SURF te overleggen of men hier niet gemeenschappelijk kan opereren. Zeer nuttig is de tabel waarin men de mogelijkheden per metadataveld bespreekt. Het Kennisnetprofiel komt goed overeen met de verplichte velden uit het DU-profiel.

Een kanttekening is mijn inziens de wat conservatieve benadering van de problematiek. Door het streven naar grondigheid en een klassieke manier van omgaan met metadata schuwt men het experiment en worden bij alternatieven vooral de zwakke kanten belicht en niet de mogelijkheden die alternatieven kunnen bieden.

Duval & Hodgins (2003). A LOM Research Agenda
Erik Duval en Wayne Hodgins zijn de veteranen van de ontwikkelingen van de LOM, de Learning Object Metadata. Een proces dat ergens medio de jaren ‘90 van de vorige eeuw begon en langzaam steeds meer toegepast wordt. De LOM is een zeer succesvolle afspraak om leerobjecten te metadateren. Duval en Hodgins zien de LOM niet als een doel maar als een voorzichtige eerste stap om serieus te werken aan hergebruik van onderwijsmaterialen. Een wellicht saaie, maar nuttige eerste stap. In dit research paper bespreken zij zestien onderwerpen die elk tot een onderzoeksopdracht kunnen leiden. Het doel is de volgende fase van hergebruik van onderwijsmateriaal te realiseren. Deze onderzoeksopdrachten lopen uiteen van het ontwikkelen van geschikte taxonomieën tot een businessmodel. Onderzoekopdracht 8 gaat specifiek in op het domein van automatische metadatageneratie. Ik ben van mening dat dit artikel richting kan geven

aan de discussie die wij moeten voeren met betrekking tot de implementatie van metadata voor leerobjecten. Het artikel staat vol interessante ideeën en concepten die schreeuwen om verdere uitwerking en discussie. Een kritiekpunt is dat men zich slechts richt op de early adopters en grote sprongen maakt door het land van leerobjecten.

Vandoolaeghe & Van Isterdael (2005). Metadata voor leerobjecten in een digitale leeromgeving
Dit artikel van Frederic Vandoolaeghe en Wim Van Isterdael beidt een uitstekende domeinbeschrijving van wat er speelt rondom het metadateren van leerobjecten. Veel discussies, losse eindjes en andere zaken die in metadataland spelen hebben hun weg gevonden in dit artikel. Op een heldere manier worden alle onderdelen besproken. Kortom een goede inleiding voor diegenen die meer willen weten over metadata voor leerobjecten en ook bruikbaar voor iedereen die alle ontwikkelingen en hun onderlinge verbanden en relaties in perspectief wil zien.

Cardinaels, Meire & Duval (2005). Automating Metadata Generation: The Simple Indexing Interface
Dit artikel biedt een goede beschrijving van de mogelijkheden van het AMG-framework zoals dat in Leuven is ontwikkeld. Interessant is ook de casusbeschrijving om met behulp van dit framework materiaal dat opgeslagen ligt in de Blackboardimplementatie in Leuven te beschrijven. Verder worden in het artikel kort de essentiële elementen en technieken die deel van het AMG-framework uitmaken beschreven. Dit artikel is de basis om de AMG-tools aan een verder onderzoek te onderwerpen.

Enkele aanbevelingen voor een vervolg

Een quickscan levert geen uitgewerkt plan op voor de toekomst; wel is het doel van deze quickscan enige handreikingen te geven voor nieuwe projecten en vervolgstappen.

Waar zijn we naar op zoek?
In een aantal studies wordt er een lans gebroken om de koninklijke weg te verlaten en de eis van volledigheid op te offeren voor het streven naar bruikbaarheid met een aanvaardbare inspanning. Het advies is om niet te streven naar perfectie maar naar een balans tussen het inzetten van softwaretechnieken en de inbreng van mensen voor het classificeren van leerobjecten

Volledige automatische metadata-extractie en -classificatie is op dit moment niet mogelijk, ondersteuning van de bestaande metadataprocessen wel. Deze ondersteuning is op de eerste plaats een optimalisatie van de workflow, bijvoorbeeld door slim gebruik te maken van sjablonen; op de tweede plaats door het inzetten van tools die losse velden van de verplichte DU-metadataset van de meest waarschijnlijke waarde voorziet.

Het AMG-werk uit Leuven verdient een nader onderzoek, bijvoorbeeld door het uitvoeren van tests met beschikbare DU-leerobjecten. Het AMG-framework is een goede kandidaat voor metadatageneratie-framework in DU-ontwikkelprojecten. Dit biedt mogelijkheden voor een netwerk van (web-)services die DU-instellingen kunnen toepassen in hun specifieke metadata-workflow.

Een voorzichtige stap voor een gemeenschappelijke aanpak
Het DU-project VAMP bouwt onder meer voort op de ervaringen van deze quicksan. In april 2006 is er een workshop georganiseerd waarin diverse organisaties uit Nederland (SURF, DU, Kennisnet, SURFnet) waren vertegenwoordigd om te komen tot een gemeenschappelijke aanpak voor het ontwikkelen van een toolset voor het automatisch generen van metadata. Ook de KU Leuven was vertegenwoordigd. Deze universiteit heeft zijn medewerking toegezegd om te onderzoeken of het AMG-framework kan worden ingezet als basis voor een dergelijke toolset.

Folksonomies

Metadata in de vorm van LOM-records is niet de enige wijze om leerobjecten vindbaar te maken. De technologie van zoekmachines (bijvoorbeeld een Google zoekmachine voor leerobjecten) is niet onderzocht. In vervolgstudies is het raadzaam ook naar deze mogelijkheden van ontsluiten te kijken. Het DU-project *Social Networking rond Leerobjecten* zal onderzoeken hoe het toepassen van folksonomies leidt tot goede resultaten bij het ontsluiten van leerobjecten.

Van wie is de metadata?

Onderzoek de keuzes die de DU kan maken met betrekking tot het eigendom en de gebruiksrechten van leerobjecten. Dergelijke metadata kunnen een economische waarde vertegenwoordigen. Een discussie over het eigendom en gebruiksrecht van metadata kan leiden tot een situatie die niet bijdraagt aan gebruik en hergebruik van leerobjecten. Het is van belang dat de DU een helder standpunt inneemt met betrekking tot haar eigen producten, over van wie de metadata is en welke eventuele beperkingen er op het gebruik liggen.

Literatuurlijst

Geciteerde literatuur

Cardinaels, K., Meire, M., & Duval, E. (2005). Automatic metadata generation: the simple indexing interface. *Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan* (pp. 548 – 556). New York: ACM Press.

Deken, J.J.E., & Wijland, M.W.P.J. van (2005, 18 augustus). *Resultaat Onderzoek naar Categorisatie Software voor SURFnet; Samenvatting – Rapport* (JD/2005/SURFNET/003). Noord-Scharwoude: Quo Vide. Beschikbaar op http://www.surfnet.nl/publicaties/surfworks2005/indi-2005-008-12.pdf.

Deken, J.J.E., & Wijland, M.W.P.J. van (2005, 20 december). *Resultaat Proof-of-concept Automatische classificatie SURFnet; publieke Samenvatting* (JD/2005/SURFNET/007). Alkmaar: Quo Vide. Beschikbaar op: http://www.surfnet.nl/publicaties/surfworks2005/indi-2005-012-30.pdf.

Duval, E., & Hodgins, W. (2004). Making metadata go away: “Hiding everything but the benefits”. In W. Jianzhong (Ed.), *DC-2004: Proceedings of the International Conference on Dublin Core and Metadata Applications* (pp. 29-35).

Duval, E., & Hodgins, W. (2003). A LOM Research Agenda. *Proceedings WWW2003, May 2003 Budapest.*

Hermans, H., & Vries, F. de, (2006). *Organisatiescenario’s voor het gebruik van leerobjecten* (Leerobjecten in de praktijk 2). Utrecht: Stichting Digitale Universiteit. Beschikbaar op http://www.du.nl/leerobjecten.

Huijsen, W., Grootveld, M., Brussee, R., Setten, M. van, & Porskamp, P. (2005, december). *Automatische Classificatie; eindrapport*. Enschede: Telematica Instituut.

Liddy, E. (2005). Metadata: A promising Solution. *Educause review May/June 2005*.

Poortman, S., & Sloep, P. (2006). *Onderwijsmodellen; de overdraagbaarheid van de didactische structuur van een complex leerobject; een case study*. (Leerobjecten in de praktijk 3). Utrecht: Stichting Digitale Universiteit. Beschikbaar op http://www.du.nl/leerobjecten.

Schoonenboom, J. (2006). *De omvang van leerobjecten* (Leerobjecten in de praktijk 4). Utrecht: Stichting Digitale Universiteit. Beschikbaar op http://www.du.nl/leerobjecten.

Vandoolaeghe, F., & Van Isterdael. W. (2005). *Metadata voor leerobjecten in een digitale leeromgeving*. Universiteit Antwerpen.

Werkgroep DU-metadata richtlijn (2004). *Werken met metadata in DU-projecten* (Deel 1. Handleiding en Deel 2: Bijlagen) (Versie 1.1, kenmerk ELO.DEL.2300/2301). Utrecht: Stichting Digitale Universiteit. Beschikbaar op http://www.du.nl/digiuni//download/temp/ELO.DEL.2300.werkenmetmetadatainDUprojectenhandleiding.pdf en http://www.du.nl/digiuni//download/temp/ELO.DEL.2301.werkenmetmetadatainDUprojectenbijlagen.pdf.

Yilmazel, O., Finneran, C.M., & Liddy, E.D. (2004). Metaextract: An NLP System to Automatically Assign Metadata. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries* (pp. 241-242).

Verdere literatuur

Cardinaels, K., Duval, E., en Olivié, H. (2002). Issues in Automatic Learning Object Indexation, In P. Barker & S. Rebelsky (Eds.), *Proceedings of ED-MEDIA World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 239-240).

Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging, 6*(4), 59-82.

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ercan Ozgencil, N., et al. (2002). Automatic Metadata Generation & Evaluation. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 2002, Tampere, Finland* (pp. 401-402). New York: ACM Press.

Metros, S.E. (2005, juli/augustus). Learning Objects: A rose by Any Other Name …. *EDUCAUSE Review, 40*(4), 12–13. Beschikbaar op http://www.educause.edu/apps/er/erm05/erm05410.asp.

Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. (2005). Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories. *Proceedings of EDMEDIA 2005, World Conference on Educational Multimedia, Hypermedia & Telecommunications, Montreal, Canada* (pp. 1407-1414). Chesapeake, VA: AACE. Beschikbaar op http://ariadne.cs.kuleuven.ac.be/amg/publicationsFiles/paperAMG2.doc.

Patton, M., Reynolds, D., Choudhury, G. S., & DiLauro, T. (2004, november). Toward a Metadata Generation Framework: A case study at Johns Hopkins University. *D-LIB magazine, 10*(11).

Weibel, S.L. (2005, juli/augustus). Border Crossings: Reflections on a Decade of Metadata Consensus Building. *D-LIB magazine, 11*(7/8).

Links

Automatic metadata generation. Website. Beschikbaar op http://ariadne.cs.kuleuven.ac.be/amg.

Bijlage 1: Lijst van leveranciers die in opdracht van Surfnet zijn onderzocht

Naast een literatuurstudie heeft Quo Vide een Request For Information (RFI) uitgestuurd naar de volgende leveranciers:

1. Teragram Categorizer
2. Inxight Categorizer
3. Wordmap
4. Nstein
5. Autonomy
6. ClearForest
7. Convera
8. Verity K2
9. Triple Hop Technologies
10. Engenium
11. Data Harmony
12. Entrieva (voorheen Semio)
13. SmartLogik
14. Temis
15. Kofax - Mohomine Classifier
16. Recommind
17. yellow brix
18. Entopia
19. Collexis
20. Irion
21. Interwoven

Bron: Deken en Wijland (2005).

Bijlage 2: Diverse open source softwaretools en algoritmen voor metadatageneratie

Een belangrijke set van open source softwaretools wordt geleverd door de Universiteit van Californië, de set tools onder de naam IVia zijn te vinden op de website: http://ivia.ucr.edu/. Het AMG framework maakt onder meer gebruik van deze tools. De metadata-extractietool is in staat metadata te generen voor de volgende velden (een duidelijke beperking is wel de Engelse taal):

- Titles
- Descriptions
- Keyphrases
- INFOMINE Categories (requires model)
- Language
- Media Type (i.e. MIME Type)
- Creator, contributor, and publisher
- Library of Congress Classification (requires model)
- LCSH
- LCC outlines

In de ABCD-demonstrator worden naast onderdelen van het AMG framework diverse open source onderdelen toegepast waaronder:

Voor taalherkenning wordt vaak het NGram algoritme gebruikt. Een open source Java-implementatie van dit algoritme is te vinden op: http://sourceforge.net/projects/ngramj.

Voor het genereren van een samenvatting gebruikt de ABCD-demonstrator het programma van Mark Watson, KBTextmaster, http://www.markwatson.com/opensource/. Men heeft dit programma aangepast om het te kunnen toepassen binnen de wensen van Kennisnet.

Een tool voor de extractie van metadata die niet besproken is, is MetaExtract van Syracuse University (Elizabeth Liddy). Deze is tool is gebaseerd op principes uit het domein van natuaral language processing (NLP). Een artikel van Yilmazel en anderen over MetaExtract (Yilmazel, Finneran & Liddy, 2004) geeft een korte beschrijving van de mogelijkheden van deze oplossing. Het is echter niet gelukt om de nodige aanvullende informatie over MetaExtract te vinden.