

# **METADATA FÖR SPRÅKRESURSER**

**En Application Profile inom området kulturarv**

**Sarah Larsson  
Tomas Lundén**

Examensarbete (20 poäng) för magisterexamen i Biblioteks- och informationsvetenskap vid Lunds universitet.

Handledare: Birgitta Olander och Colm Doyle

BIVIL:s skriftserie 2003:19

ISSN 1401-2375

© Lunds universitet. Biblioteks- och informationsvetenskap 2003

## Title

Metadata for Language Resources: an Application Profile for the Cultural Heritage domain

## Abstract

The EU-funded project *European Cultural Heritage Online* (ECHO) aims to integrate resources from various domains within the field of cultural heritage, and make them available on the Internet. To accomplish this integration it is necessary to describe the various resources in a way that will achieve cross-domain interoperability. This description of digital resources is commonly referred to as metadata. The most well known and used metadata standard is the *Dublin Core Metadata Element Set*, a general standard developed for the purpose of describing all types of digital resources. Since the mid-1990's a large number of specialized standards have emerged in different communities, for example *IMDI*, designed for detailed description of language resources.

This thesis aims to examine the possibility of combining a general and a specialized metadata standard to achieve cross-domain interoperability and at the same time provide a sufficiently detailed description of language resources for researchers within the field of linguistics. A combination of two (or more) metadata standards in a compound schema is called an *application profile*. The purpose of an application profile is to provide a tailor-made schema for a specific context. The present study examines the semantic and structural problems concerning the creation of an application profile for language resources, which combines elements from the metadata standards Dublin Core and IMDI.

Tack!

Vi vill tacka våra handledare: Birgitta Olander för ovärderlig hjälp, ett aldrig sinande tålamod och för att hon inte tappade tron på oss när vi bytt inriktning på uppsatsen för femtonde gången. Colm Doyle för hans sakkunskaper och knuffar i rätt riktning när vi *trodde* att vi visste vad metadata var.

Vi vill även tacka professor Sven Strömqvist och Marcus Uneson vid Institutionen för lingvistik, Lunds universitet.

## INNEHÅLLSFÖRTECKNING

<b>1. INLEDNING .....</b>	<b>4</b>
1.1 SYFTE .....	8
1.2 PROBLEMSTÄLLNING .....	8
1.3 FRÅGESTÄLLNINGAR .....	8
1.4 METOD .....	9
1.5 AVGRÄNSNINGAR .....	9
<b>2. TEORI OCH LITTERATURGENOMGÅNG .....</b>	<b>12</b>
2.1 KUNSKAPSORGANISATION OCH KATALOGISERING .....	12
2.2 METADATA – DEFINITIONER OCH ANVÄNDNINGSSOMRÅDEN .....	14
2.2.1 Olika typer av metadata .....	17
2.2.2 Metadata – standard, schema eller format? .....	19
2.3 DOKUMENT, RESURSER OCH DOCUMENT-LIKE-OBJECTS (DLO) .....	20
2.4 VAD ÄR EN APPLICATION PROFILE? .....	22
2.4.1 Namespaces .....	22
2.4.2 Application profiles .....	24
2.4.3 Några exempel på application profiles .....	28
2.5 KULTURARV (CULTURAL HERITAGE) .....	29
2.5.1 ECHO-projektet .....	30
<b>3. METADATASTANDARDERNA DUBLIN CORE OCH IMDI .....</b>	<b>31</b>
3.1 DUBLIN CORE METADATA INITIATIVE (DCMI) .....	31
3.1.1 Dublin Core Metadata Element Set Version 1.1 .....	32
3.1.2 Dublin Core Qualifiers .....	33
3.2 IMDI (ISLE METADATA INITIATIVE) .....	37
3.2.1 Språkresurser – definition och konstitution .....	38
3.2.2 IMDI Metadata Elements for Session Descriptions Version 2.5 .....	40
3.2.3 IMDI:s ordlistor .....	47
3.2.4 Förändringar och versioner av IMDI .....	47
<b>4. BESKRIVNING AV EN SPRÅKRESURS I DUBLIN CORE OCH IMDI .....</b>	<b>49</b>
4.1 BESKRIVNING AV EN SPRÅKRESURS I DUBLIN CORE .....	49
4.2 BESKRIVNING AV EN SPRÅKRESURS I IMDI .....	51
4.3 KOMMENTAR TILL BESKRIVNINGARNA .....	55
<b>5. APPLICATION PROFILE FÖR BESKRIVNING AV SPRÅKRESURSER .....</b>	<b>56</b>
5.1 MÅLGRUPP OCH SYFTE .....	56
5.2 REDOVISNING AV APPLICATION PROFILE FÖR BESKRIVNING AV SPRÅKRESURSER .....	56
5.3 BESKRIVNING AV EN SPRÅKRESURS MED VÅR APPLICATION PROFILE .....	60
5.4 KOMMENTAR TILL BESKRIVNINGEN .....	61
<b>6. ANALYS OCH DISKUSSION .....</b>	<b>63</b>
6.1 ALLMÄN DISKUSSION OM PROFILEN .....	63
6.2 ELEMENTEN I PROFILEN .....	64
6.3 PROBLEM RÖRANDE SEMANTIK .....	65
6.4 PROBLEM RÖRANDE STRUKTUR .....	67
6.5 HUR STRUKTURPROBLEM KAN PÅVERKA SEMANTIKEN .....	69
6.6 INTEROPERABILITET KONTRA ÄMNESSPECIFICITET .....	70
<b>7. SLUTSATSER .....</b>	<b>73</b>
<b>FÖRKORTNINGAR .....</b>	<b>76</b>
<b>KÄLL- OCH LITTERATURFÖRTECKNING .....</b>	<b>78</b>

# 1. Inledning

*World Wide Web* är en mycket stor samling av söktjänster, hemsidor, textdokument, videofiler, ljudfiler och andra typer av resurser. Exakt hur stor webben är beror på hur man mäter, men att mängden material är överväldigande råder det inga tvivel om. Antalet hemsidor (definierade som domänadresser) på webben har uppmätts till följande av Internetkonsultföretaget Netcraft:

juni 1993: 130 st.

april 2000: 14 322 950 st.

januari 2003: 35 424 956 st.<sup>1</sup>

Det finns andra sätt att försöka bestämma hur stor webben egentligen är. Antalet individuella hemsidor och dokument av olika slag är avsevärt större än ovanstående siffror, eftersom olika hemsidor kan ha samma domänadress. (Exempel: [www.metadata.se](http://www.metadata.se) är en domänadress och en hemsida, [www.metadata.se/sarahs\\_sida](http://www.metadata.se/sarahs_sida) är en annan hemsida som delar domänadress med den förstnämnda. I Netcrafts undersökning räknas dessa två sidor som en sida, hos andra skulle de räknas som två olika.) Vad gäller antal dokument på webben, så uppgick de enligt en undersökning i januari 2000 till ungefär en miljard.<sup>2</sup>

Det är idag välkänt för alla som någon gång gjort en sökning via en sökmotor på Internet, vilka enorma kvantiteter söksvar som oftast blir resultatet. Inte sällan förefaller också svaren irrelevanta för det man verkligen letade efter. Internet eller World Wide Web har ingen organiserad katalog, där dokument och resurser har ordnats enligt givna regler. En enkel belysning av problemet kan se ut som följer: Wayne Jones skriver att han gör en sökning på namnet "Clifford A. Lynch" på sökmotorn Google, den 7 september 2000 (Clifford A. Lynch är chef för *Coalition of Networked Information*, i USA). Han får då 2 270 träffar. Vissa av träffarna är dokument *av* Lynch, vissa är dokument *om* honom, andra tycks vara om eller av

---

<sup>1</sup> Siffrorna från 1993 mättes av Matthew Gray vid Massachusetts Institute of Technology, de från 2000 av företaget Netcraft. Redovisat i Gill, Tony, 2000. "Metadata and the World Wide Web 2000". Den senaste undersökningen har vi själva hämtat från Netcrafts hemsida.

URL: <http://www.netcraft.com/survey>

<sup>2</sup> Undersökningen gjordes av Inktomi, ett företag som konstruerar sökmotorer, tillsammans med NEC Research Institute. Se Gill 2000.

någon annan med samma namn, ytterligare andra träffar rör en Clifford J. Lynch, och många träffar innehåller endast namnet ”Clifford” eller ”Lynch”.<sup>3</sup> Problemet här är alltså att det inte finns någon möjlighet för den som söker att bestämma sökningen till ett visst fält, som till exempel titel, författare, ämnesord, på det sätt man kan göra i en bibliotekskatalog. Ett sätt att lösa problemet skulle vara att göra en beskrivning av dokumenten på webben, så att det blev möjligt att på ett effektivare sätt söka, identifiera och återfinna relevanta dokument. En sådan beskrivning kallas även metadata.

Begreppet metadata används idag främst angående beskrivning av resurser inom digitala nätverk och kanske i synnerhet World Wide Web. Men begreppet har existerat redan före uppkomsten av webben, i själva verket sedan mitten av 1970-talet. Ursprungligen användes termen framför allt inom områden som datavetenskap, systemvetenskap, databashantering och informationssystem.<sup>4</sup> Med framväxten av WWW har dock termen metadata blivit alltmer intressant och relevant för biblioteksvärlden (samt arkiv och museer) och för humaniora och samhällsvetenskaperna i allmänhet. Under 1990-talet formligen exploderade detta fält och ett stort antal metadatastandarder har sett dagens ljus de senaste tio åren.

Den förmodligen mest utbredda metadatastandarden för elektroniska resurser är *Dublin Core Metadata Element Set* (ofta bara kallad Dublin Core), men det finns ett antal andra standarder och scheman för metadata, ofta utformade specifikt för ett speciellt ämnesområde eller domän. Ett exempel är IMDI (*ISLE Metadata Initiative*), som är en standard tänkt att användas för språkresurser inom lingvistik. Det finns dessutom standarder som kan sägas vara mediespecifika, som MPEG-7 (utvecklad av *Moving Pictures Expert Group*). Den har skapats för att beskriva multimediala resurser och göra dem sökbara på en mängd olika sätt.

Under de senaste åren har det börjat växa fram projekt världen över för att bevara och tillgängliggöra *kulturarv* (resurser inom konst, arkitektur, historia m.m.) digitalt, i form av olika webbaserade databaser. Det innebär att det finns behov för väl fungerande metadata-scheman, som på ett tillfredsställande sätt kan beskriva de mycket varierande typer

---

<sup>3</sup> Jones, Wayne, 2002. ”Preface: Meting Out Data”, i Jones, W., Ahronheim, J.R. & Crawford, J., 2002. *Cataloging the Web: Metadata, AACR, and MARC 21*. Lanham, Maryland & London: The Scarecrow Press, s. v.  
<sup>4</sup> Se t.ex. Borgman, Christine L., 2000. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, Mass. & London: The MIT Press, s. 68. Gilliland-Swetland, Anne J., 2000. ”Setting the Stage”, *Introduction to Metadata: Pathways to Digital Information*.

av resurser som finns. Metadata behövs för att resurserna effektivt ska kunna sökas och återfinnas i den närmast oändliga mängd dokument och resurser som utgör World Wide Web.

Eftersom många av dessa projekt inom området kulturarv syftar till att bringa samman olika ämnesområden i en och samma databas eller portal, ställer det krav på att resurserna kan beskrivas på ett standardiserat sätt för att effektivt kunna sökas och återfinnas. Ur denna aspekt kan Dublin Core förefalla vara lösningen på alla problem. Standarden, som utarbetades i en första version 1995, har som ett av sina syften att kunna beskriva alla typer av elektroniska resurser. Dublin Core utformades av bibliotekarier, vilket kanske förklarar denna lovvärda ambition. Och Dublin Core har många fördelar, framför allt enkelheten. Den kan utan alltför stor ansträngning förstås och användas av såväl katalogisatörer som producenter av resurser inom olika ämnesområden och på olika nivåer. Enkelheten kombineras också med flexibilitet och en viss komplexitet (möjligheten att uttrycka så kallad kvalificerad Dublin Core). Standarden används i många projekt runtom i världen och finns i skrivande stund översatt till 25 språk. Den är det närmaste man har kommit till en generell metadatastandard för resursbeskrivning på webben, så allt borde därmed vara frid och fröjd i den digitala informationsvärlden. Icke desto mindre är det tydligt att Dublin Core har begränsningar och dessa gäller till exempel för resurser som av olika anledningar kräver en djupgående beskrivning. Det kan handla om forskningsinformation i ett visst ämne eller beskrivning av resurser såsom ljud-, bild-, video- och multimediafiler. Här kan det finnas användare som inte endast vill hitta fram till resursen, utan sedan vill söka inom resursen, till exempel på segment av en videofilm eller inspelningen av en röst. För dessa typer finns förstås, som ovan nämnts, ofta mycket specifika metadatastandarder utformade.

Vi återkommer därmed till idén om digitala kulturarvsdatabaser med allmän tillgänglighet på webben. De resurser som samlas i dylika projekt kan både vara av varierande ämnesinnehåll och bestå av olika medier eller format, med olika grader av komplexitet. De kan komma från skilda domäner och institutioner, bibliotek, museer, arkiv, universitet m.m., alla med sina egna sätt att hantera resurser. I detta kan Dublin Core vara av nytta, eftersom ett annat av dess syften är att underlätta *interoperabilitet* mellan standarder och domäner på webben. Med interoperabilitet menas möjligheten för domäner (t.ex. olika ämnesdatabaser eller resurssamlingar) att samverka i ett digitalt nätverk (genom metadata, men också med hjälp av tekniska lösningar). Det är ovanligt att projekt av detta slag vill använda sig av en enda ”ren” metadatastandard, till exempel Dublin Core som den är konstruerad. Ofta anses en standard

inte kunna täcka de varierande behov som projektet har. Man kan vilja ha en något djupare beskrivning än vad Dublin Core kan erbjuda. Ett tillvägagångssätt är att utgå ifrån Dublin Cores uppsättning av beskrivningselement och göra lokala tillägg inom dessa ramar, *extensions*, som man anser nödvändiga för projektets egna behov. Man kan också skapa en *application profile*, som enkelt kan beskrivas som en tillämpning där man tar beskrivningselement från olika existerande standarder och sätter samman dem i ett eget metadataschema. I en *application profile* kan man även skapa egna element om man finner det nödvändigt.

Institutionen för lingvistik vid Lunds universitet deltar när detta skrivs i ett EU-finansierat projekt kallat *European Cultural Heritage Online* (ECHO). Målet i ett första stadium är att skapa metadata för språkresurser inom lingvistisk forskning och göra resurserna tillgängliga på webben. Därefter skall andra ämnesområden tillkomma. ECHO-projektets mål är en utgångspunkt för syftet med vår uppsats.

Det finns, som vi redan nämnt, en metadatastandard kallad IMDI som är utvecklad för att ingående kunna beskriva olika typer av språkresurser (vilka kan vara videoinspelningar, bandinspelningar, textdokument m.m.). Men problemet som redan skisserats är att om dessa resurser skall kunna integreras med andra ämnesområden krävs att det finns möjlighet för interoperabilitet mellan områdena. Det är dessutom så att eftersom dessa resurser i första hand är skapade av och ämnade för forskare, ställs krav på en mer djupgående och ämnesspecifik beskrivning än vad som är möjligt att göra med Dublin Core, som är en generell metadatastandard. En möjlighet för ECHO-projektet skulle vara att skapa en *application profile* för beskrivning av språkresurser, som för interoperabilitetens skull baseras på Dublin Core, men för vissa nödvändiga ämnesspecifika egenskaper använder element från IMDI. Detta är vad vi vill undersöka i föreliggande uppsats.

Eftersom den övervägande delen av litteraturen i ämnet metadata är skriven på engelska bör vi ta upp hur vi använder engelska och svenska begrepp. Det centrala begreppet *application profile* har inte någon svensk översättning och vi har valt att inte försöka översätta det själva. Vi kommer att använda det engelska uttrycket, men använder omväxlande uttrycket ”profil” för att undvika alltför mycket upprepningar. Inte heller termerna *namespace* och *qualifier* förefaller ha någon svensk översättning, såvitt vi kunnat finna. Vi håller oss även där med de



engelska begreppen. I övrigt försöker vi använda svenska termer. Där det svenska ordet kan förefalla tveksamt återger vi det engelska uttrycket direkt efter inom parentes.

Vi vill här också ge grundläggande definitioner för två begrepp som är viktiga för förståelsen av uppsatsens frågeställningar. Det är termerna *semantik* och *struktur*, som de förstås i relation till ämnet metadata. Semantik innebär meningen eller innebörden av de individuella metadataelementen och deras olika delar. Struktur betyder den hierarki eller struktur i vilken metadataelementen är sammansatta. Påpekas bör att vi i denna uppsats menar metadatastandardens inneboende struktur, så som den har konstruerats. Struktur kan också innebära vilken teknisk syntax metadatan sätts in i, som t.ex. HTML eller RDF. Denna tekniska betydelse av termen struktur är det alltså inte fråga om här.

## 1.1 Syfte

Syftet med vårt arbete är att undersöka om man genom att utgå från en generell och en ämnesspecifik metadatastandard kan underlätta interoperabilitet mellan olika ämnens resurser och samtidigt tillhandahålla en ämnesspecifik beskrivning. Det handlar om att försöka integrera bredd och djup i ett och samma beskrivningsschema.

## 1.2 Problemställning

För att nå detta syfte har vi formulerat följande problem:

Går det att skissera en application profile, baserad på dels en generell standard (Dublin Core), dels en domänspecifik standard (IMDI) för språkresurser, som kan fungera som vägledning för utveckling av ämneskompatibla beskrivningsscheman? Kan man identifiera några generella problem vid utvecklandet av en sådan application profile?

## 1.3 Frågeställningar

- Vilken typ av beskrivningselement i en application profile för språkresurser kan hämtas från en generell metadatastandard (för att uppnå interoperabilitet)?
- Vilken typ av beskrivningselement i en application profile för språkresurser måste hämtas från en ämnesspecifik metadatastandard för att nå tillräckligt djup i beskrivningen?

- Hur stor betydelse har skillnader i struktur och semantik mellan de olika standarderna för möjligheterna att utveckla en application profile som är både generell och ämnesspecifik?

## 1.4 Metod

Med Dublin Core som grund ämnar vi skapa en application profile som kan användas för att beskriva språkresurser. Profilen kommer att innehålla *namespaces* (för närmare definition se 2.4.1) från Dublin Core och IMDI. Vi använder en mappning mellan standarderna i vårt arbete. En mappning innebär en uppställning som visar relationer och likheter mellan två eller flera metadatascheman. Se bilaga 1 där en mappning mellan IMDI och Dublin Core redovisas.<sup>5</sup> Genom att göra en mappning från en komplex standard till en mer generell kan man få en utgångspunkt för att underlätta interoperabilitet. Mappningen är dock främst ett hjälpmedel och vi gör framför allt egna bedömningar. Utifrån dessa bedömningar ska vi identifiera de IMDI-element vi betraktar som generella och som därmed kan uttryckas i Dublin Core. Det menar vi kan ge oss en grund för att besvara vår första frågeställning rörande interoperabiliteten.

För att besvara vår andra frågeställning om att nå djup i ämnesbeskrivningen kommer vi att undersöka vilka av IMDI-elementen som är nödvändiga för att ge en tillräckligt ämnesspecifik beskrivning av språkresurser.

Den tredje frågeställningen rörande skillnaderna i struktur och semantik mellan standarderna skall vi försöka besvara genom en analys av standarderna och skillnaderna dem emellan. Vi skall visa på dessa problem genom belysande exempel. Bland annat ämnar vi beskriva en språkresurs med respektive Dublin Core, IMDI och vår application profile, för att kunna jämföra vilken typ av information som vinnrespektive förloras i de olika fallen.

## 1.5 Avgränsningar

---

<sup>5</sup> Denna mappning är utförd av ISLE-initiativet och är gjord utifrån en äldre version av IMDI, Version 2.2. På grund av detta överensstämmer inte alla element med den version vi arbetar med i uppsatsen, Version 2.5. Om de olika versionerna av IMDI, se avsnitt 3.2.4. För att se en stor mappningstabell mellan många olika standarder hänvisas till Getty research institute:  
URL: [http://www.getty.edu/research/institute/standards/intrometadata/3\\_crosswalks/index.html](http://www.getty.edu/research/institute/standards/intrometadata/3_crosswalks/index.html)

Vi har valt att inrikta oss på ett av de ämnesområden som ingår i ECHO-projektet (se 2.5.1), nämligen språkresurser. Dessutom kommer vi att avgränsa oss inom metadatastandarden IMDI. Standarden består av tre olika metadatascheman för att beskriva olika typer av språkresurser. Dessa tre scheman är *Session Descriptions*, *Lexicon Descriptions* och *Catalogue Descriptions*. Vi kommer i uppsatsen att endast fokusera på *Session Descriptions*, eftersom det är det största schemat inom ISLE:s metadatainitiativ. Vi har också gjort en avgränsning inom *Session Descriptions*. Här har vi valt att inte ta med den del av schemat som beskriver s.k. annotationer, då denna beskriver något som ligger utanför den egentliga resursen. I *Session Descriptions* finns även något som kallas *Sub-schema*. Denna del tar vi inte heller med i vårt arbete, för att begränsa det redan tillräckligt komplexa metadataschemat. (En närmare genomgång av IMDI kommer i avsnitt 3.2.)

Det finns ytterligare några avgränsningar: en är att vi har valt att utgå ifrån endast Dublin Core och IMDI i valen av beskrivningselement och namespaces till vår profil. Man utgår inte alltid från en dylik begränsning, utan undersöker kanske fler standarder innan man bestämmer vilken eller vilka namespaces man vill använda sig av. En annan avgränsning är frågan om gränssnitt, det vill säga hur olika sökfält kan synas för användaren. Det är något som också kan påverka användandet av metadata, men i denna uppsats kommer vi att bortse från detta perspektiv. Ytterligare en avgränsning rör tekniken. För att kunna göra en application profile användbar och interoperabel i en digital miljö krävs också ett tekniskt ramverk. Denna tekniska aspekt kommer vi endast att nämna kort och inte ta hänsyn till i uppsatsen. Stuart Weibel, som är chef för Dublin Core Metadata Initiative och verksam vid OCLC, menar att man kan tala om tre olika sorters interoperabilitet:

A resource community is characterized by common semantic, structural and syntactic conventions for the exchange of resource descriptions. Standards used for semantic interoperability include Dublin Core, Anglo-American Cataloging Rules (AACR2), TEI and FGDC. Structural interoperability will be based on the Resource Description Framework (RDF). The W3C Extensible Markup Language (XML) will form the basis for syntactic interoperability.<sup>6</sup>

---

<sup>6</sup> Citerad hos Lazinger, Susan S., 2001. *Digital Preservation and Metadata: History, Theory, Practice*. Englewood, CO: Libraries Unlimited, s. 144.

I uppsatsen kommer endast den semantiska interoperabiliteten att diskuteras. Det är den del av interoperabiliteten som har att göra med metadatastandardernas element och deras innebörd. Den strukturella och syntaktiska interoperabiliteten handlar om hur metadata uttrycks, i vilket tekniskt ”språk” den infogas.

## 2. Teori och litteraturgenomgång

### 2.1 Kunskapsorganisation och katalogisering

Kunskapsorganisation kan definieras som konsten att organisera kunskapen så att den kan återfinnas.<sup>7</sup> Man katalogiserar för att fastslå ett dokumentens existens, dess identitet, var dokumentet finns och eventuellt om det finns tillgängligt.<sup>8</sup> Alltså för att strukturera och återfinna dokumentet i en samling. Istället för att använda termen katalogisering talar man ibland om bibliografisk kontroll eller ”bibliographic management”. En som har betytt mycket för utvecklandet av den moderna tidens katalogiseringsteorier är Charles A. Cutter som 1876 kom ut med sina katalogiseringsregler under titeln *Rules for a printed dictionary catalogue*. Här beskriver Cutter vilka ändamålen är med en katalog och hur man uppnår dessa.<sup>9</sup> Under ändamålen ligger bland annat att:

- Hjälpa en person att finna ett visst dokument om dess titel, upphov eller ämne är känt.
- Visa vad biblioteket förvaltar inom ett givet ämne, eller totalt bestånd.<sup>10</sup>

För att katalogisera en resurs måste man analysera dess form och innehåll och undersöka vilka uppgifter som behöver beskrivas för att man ska kunna återfinna den vid senare sökning. Vid en sådan analys är det av yttersta vikt att ta hänsyn till var och hur informationen skall användas och vad användarna har för behov. Då man vill kunna presentera alla dokument i en katalog på ett likartat, jämförbart och förutsägbart sätt krävs fasta regler för hur beskrivningen skall gå till. Dessa regler bör vara utformade med avseende på förutsägbarhet, logik och standardisering.<sup>11</sup> Standardiseringen gäller även för de data man lägger in i katalogen, för detta använder man sig ofta av någon slags kontrollerad vokabulär. De regler som bestämmer hur en katalog skall vara utformad kallas (katalog)format. Så här skriver Sten Hedberg, f.d. förste bibliotekarie vid Uppsala universitetsbibliotek, i ämnet:

---

<sup>7</sup> Benito, Miguel, 2001. *Kunskapsorganisation: en introduktion till katalogisering, klassifikation och indexering*. Borås: Tarancos bokförlag.

<sup>8</sup> Strunck, Kirsten, Lund, Haakon, Thorlund Jepsen, Erik, 1998. *Katalogiseringsteori*. København: Danmarks Biblioteksskole.

<sup>9</sup> Benito 2001.

<sup>10</sup> Rowley, Jennifer & Farrow, John, 2000. *Organizing knowledge: An introduction to Managing Access to Information*. Aldershot: Gower Publishing Ltd.

<sup>11</sup> Björkhem, Miriam & Lindholm, Jessica, 2000. *Metadata för det digitala biblioteket: objektbeskrivning av elektroniska resurser*. Magisteruppsats i Biblioteks- och informationsvetenskap, Lunds universitet.

En katalog består av **poster** (eng. **records**). - En post i katalogen representerar en **individ**. Varje post består av ett **urval av den information** som totalt existerar om en individ. Vilket urvalet är, bestäms av katalogens **syfte**. I en och samma katalog måste samma urval tillämpas i alla poster. Den minsta enheten av sådana uppgifter kallas **(informations)element**. [...] Om man katalogiserar dokument (och/eller resurser) skapar man information om individer som själva består av information. Som begrepp för sådan **överordning** används förledet "**meta-**", och "**information**" är lika med "**data**". En **katalogpost som avser en informationsresurs** kan då kallas **metadata**, information om denna katalogpost i sin tur **meta-metadata**. Även om sålunda **metadata** kan avse all kataloginformation om information har termen en pregnant betydelse: **kataloginformation om elektroniska resurser**.<sup>12</sup>

På senare år har de dokument som skall beskrivas förändrats drastiskt när det gäller medium och format. Miriam Björkhem och Jessica Lindholm skriver i sin uppsats *Metadata för det digitala biblioteket* att en riktigare term än dokumentbeskrivning kanske skulle vara "objektbeskrivning".<sup>13</sup> Elektroniska dokument har den egenheten att innehållet kan finnas tillgängligt för en användare trots att det inte finns fysiskt på plats. Detta gör att den som ansvarar för samlingen, t.ex. biblioteket, inte äger eller har kontroll över dokumentet, vilket även påverkar katalogiseringen. De resurser som finns tillgängliga på webben, utanför bibliotekens kontrollerade resurser (t.ex. databaser) är oftast katalogiserade (beskrivna) av upphovsmannen eller inte katalogiserade alls. Biblioteken kan välja att katalogisera ett eget urval av dessa typer av resurser, men de elektroniska dokumentens föränderliga natur gör att det finns en risk att katalogposten efter en tid inte överensstämmer med dokumentet. Alla de nya typer av resurser som uppkommit har ställt nya krav på reglerna för hur man skall katalogisera dem. Många dokument är även blandningar av olika dokumenttyper, till exempel resurser som innehåller både ljud, bild och text. Allt detta har bidragit till framväxten av nya format och standarder för dokumentbeskrivning.

---

<sup>12</sup> Hedberg, Sten, [odat.], "Metadata - kataloginformation på internet".

<sup>13</sup> Björkhem & Lindholm 2000.

## 2.2 Metadata – definitioner och användningsområden

Metadata har definierats på en mängd olika sätt. Den vanligaste och kortaste definitionen är ”data om data”. Den förekommer allt som oftast som inledning hos skribenter i ämnet.<sup>14</sup>

Hudgins, Agnew & Brown ger en mer fullödlig definition:

The term ”metadata” commonly refers to any data that aids in the identification, description and location of networked electronic resources. A primary function of metadata is resource discovery: metadata increases the odds that a user will be able to retrieve appropriate information and assess its usefulness and availability.<sup>15</sup>

Det finns ett stort antal andra definitioner som skulle kunna diskuteras, men ovanstående fungerar väl för vårt syfte. Definitionen pekar på ett par saker som är centrala även inom traditionell katalogisering, nämligen att genom beskrivningen av dokumentet kunna identifiera och (åter)finna det. Detta ligger nära Cutters ändamål med en katalog. I själva verket är, såsom Hedberg beskriver ovan i 2.1, bibliotekens gamla katalogkort en form av metadata. Skillnaden gentemot katalogkortens tid är att det idag inte endast handlar om böcker och annat tryckt material, utan även en mängd andra medier (många av dem elektroniska) och att det allt mindre handlar om att visa bestånd i ett visst biblioteks samling, utan snarare att möjliggöra direkt åtkomst till dokumentet via nätverksbaserade lösningar.

Är alltså metadata och katalogisering endast olika beteckningar för samma fenomen? Vi vill hävda att det beror på vilket perspektiv man anlägger. Marita Fagerlind och Gunilla Gisselqvist skriver i uppsatsen *Metadata enligt Dublin Core* att katalogisering är en term som ”hör hemma speciellt i bibliotekskretsar och innefattar tryckt eller åtminstone fysiskt material, medan metadata är resursbeskrivning av elektroniska objekt, företrädesvis på Internet”.<sup>16</sup> Vi menar att denna strikta uppdelning inte helt håller streck, framför allt i det att formuleringen implicerar att bibliotek inte skulle syssla med metadata och elektroniska resurser, vilket de bevisligen gör, om än i varierande utsträckning. Inte heller kan man hävda att katalogisering

---

<sup>14</sup> Se t.ex. Borgman 2000, s. 68. Lazinger 2001, s. 139. Schwartz, Candy, 2001. *Sorting Out the Web: Approaches to Subject Access*. Westport, Conn. & London: Ablex Publishing, s. 9.

<sup>15</sup> Hudgins, J., Agnew, G. & Brown, E., 1999. *Getting Mileage out of Metadata: Applications for the Library*. Chicago: American Library Association, s. 1.

<sup>16</sup> Fagerlind, Marita & Gisselqvist, Gunilla, 1999. *Metadata enligt Dublin Core: tillämpningar och konsekvenser i de svenska kvalitetssöktjänsterna SAFARI, Svenska miljönätet och Svesök*. Magisteruppsats i Biblioteks- och informationsvetenskap, Lunds universitet, s. 12.

idag endast innefattar fysiskt material. Rebecca Guenther, specialist på MARC och nätverksutveckling vid Library of Congress, skriver i artikeln ”MARC 21 as a Metadata Standard” om de revisioner av MARC-formatet som inleddes i början av 1990-talet, för att just kunna omfatta Internetresurser, vilket bland annat fick som resultat att man lade till fält 856, där man kan lägga in en URL, det vill säga en webbadress, för direkt länk till resursen.<sup>17</sup> Det finns även de som hävdar att metadata är (eller åtminstone bör vara) synonymt med katalogisering. I en intressant och något polemisk artikel skriver Michael Gorman, som är chef för Library Services vid California State University, att elektroniskt material helst bör katalogiseras enligt standarder som redan används för annat material, det vill säga MARC-formatet. Själva termen metadata tycks han mena är ett modebegrepp: ”My belief is that ’metadata,’ as presently conceived, will evolve toward standardization of elements and content and will be indistinguishable from real cataloguing in a relatively short time.”<sup>18</sup> Jennifer A. Younger, chef för University Libraries vid Notre Dame University, lägger fram en framtida utvecklingsbild som skiljer sig något från Gormans, men som också innebär att åtskillnaden mellan metadata och katalogisering upplöses. Hon föreslår att man byter beteckning på katalogisatörer (”cataloging librarian”) till ”metadata librarian”. Katalogisering av traditionella fysiska dokument skulle då bli en underkategori (och bara en av arbetsuppgifterna) under metadataarbete i en digital nätverksmiljö med olika metadatastandarder i samverkan.<sup>19</sup>

Denna diskussion till trots (och oaktat att de gamla katalogkortet också är ett slags metadata), är det onekligen så att metadata som begrepp och metadatastandarder i praktiken idag används specifikt för elektroniska dokument och resurser, vilket Hudgins, Agnews & Browns definition ovan också betonar. Men det är snarare andra aspekter av metadata än vilka dokument- och resurstyper som beskrivs, som skiljer den från bibliotekens katalogisering. Clifford A. Lynch, chef för *Coalition of Networked Information*, vill poängtera ett bredare perspektiv på metadata, som han menar att biblioteksvärlden ofta inte är medveten om. Han hävdar att diskussionen inom biblioteksdomänen förfäktar en tanke om att metadata kan ge en gudomligt given, ideal beskrivning av ett informationsobjekt, ”a sort of Platonic ideal of

---

<sup>17</sup> Guenther, Rebecca, 2002. ”MARC 21 as a Metadata Standard: A Practical and Strategic Look at Current Practices and Future Opportunities”, i Jones et al, s. 41ff.

<sup>18</sup> Gorman, Michael, 2002. ”Metadata: Hype and Glory”, i Jones et al, s. 181.

<sup>19</sup> Younger, Jennifer A., 2002. ”Metadata and Libraries: What’s It All About?”, i Jones et al, s. 8f.



description”.<sup>20</sup> Detta är en missuppfattning, enligt Lynch. Dels därför att metadata innebär mer än bara beskrivning och klassificering, dels för att det är meningslöst att tala om metadata utan att specificera vad man vill använda den till, i vilken kontext den ska fungera.

Metadata becomes interesting and useful when we *employ* it to construct information retrieval and management systems (with catalogs being only one rather specialized example), and when we *apply and exploit* it to make information more accessible or more manageable. The context of use is everything, and only when we talk about the contexts of use can we really talk about metadata in an informed and meaningful way.<sup>21</sup>

Anne J. Gilliland-Swetland, Associate Professor i Information Studies vid University of California, framhåller också att metadata måste relateras till användningsområdet och exemplifierar:

An Internet resource provider might use *metadata* to refer to information being encoded into HTML metatags for the purposes of making a Web site easier to find. Individuals digitizing images might think of metadata as the information they enter into the header field for the digital file to record information about the image, the imaging process, and image rights. A social science data archivist might use the term to refer to the systems and research documentation necessary to run and interpret a magnetic tape containing raw research data. An electronic records archivist might use the term to refer to all the contextual, processing, and use information needed to identify and document the scope, authenticity, and integrity of an active or archival record in an electronic recordkeeping system. [---] In all of these diverse interpretations, metadata not only identifies and describes an information object; it also documents how that object behaves, its function and use, its relationship to other information objects, and how it should be managed.<sup>22</sup>

Huruvida biblioteksvärlden verkligen är omedveten om de olika användningsområdena för metadata vill vi låta vara osagt. Men vi ska se närmare på hur olika typer av metadata har definierats.

---

<sup>20</sup> Lynch, Clifford A., 2002. ”Future Developments in Metadata and Their Role in Access to Networked Information”, i Jones et al, s. 183.

<sup>21</sup> Lynch 2002, s. 183.

<sup>22</sup> Gilliland-Swetland 2000.

### 2.2.1 Olika typer av metadata

Metadata har uppdelats i varierande typer och användningsområden. Lorcan Dempsey och Rachel Heery delar in metadata i tre kategorier ("bands"), beroende på hur strukturellt och semantiskt avancerad metadatan är.

**Band 1:** metadata som är fulltextindex, t.ex. sökmotorers fulltextindexering av hemsidor på webben. Ex.: AltaVista, Lycos. Det är automatiskt genererad metadata.

**Band 2:** metadata som är enkel i struktur och inte domänspecifik (dvs är generell). Ex.: Dublin Core, IAFA/WHOIS++. Kan vara både automatiskt och manuellt genererad.

**Band 3:** metadata som är domänspecifik och med komplex struktur. Ex.: MARC, *Encoded Archival Description* (EAD). (Här skulle, enligt vår mening, även IMDI kunna placeras.) Kräver manuell generering, ofta med expertkunskaper.<sup>23</sup>

Andra har kategoriserat metadatastandarder utifrån primära användningsområden eller funktioner. Roy Tennant, chef för Web & Services Design vid California Digital Library, identifierar tre sådana områden:

**1) Beskrivande metadata:** t.ex. titel, författare (skapare), ämnesord, dvs element för att söka och återfinna resurser.

**2) Strukturmetadata:** beskriver hur en resurs är strukturerad, t.ex. om en resurs innehåller flera filer.

**3) Administrativ metadata:** t.ex. hur filen är producerad, vem som innehar rättigheter etc.<sup>24</sup>

Gilliland-Swetland presenterar en bredare och i vår mening mer väldefinierad indelning, som till viss del överlappar Tennants:

**1) Administrativ:** metadata som används för att hantera och administrera resurser. Ex.: rättighetsinnehav, villkor för åtkomst av resursen, placering (det kan vara en digital reproduktion av ett fysiskt konstverk), etc.

---

<sup>23</sup> Dempsey, Lorcan & Heery, Rachel, 1998. "Metadata: a current view of practice and issues", *Journal of Documentation*, 54 (2), s. 156. Lorcan Dempsey är f.d. chef för *UK Office for Library and Information Networking* (UKOLN), numera vicepresident för OCLC. Rachel Heery är biträdande chef för forskning och utveckling vid UKOLN.

<sup>24</sup> Citerad hos Lazinger 2001, s. 142. Roy Tennant, "Digital Libraries: 21st-Century Cataloging", *Library Journal* (April 15, 1999).

**2) Beskrivande:** för att beskriva och identifiera resursen. Ex.: indexering, kataloginformation, ev hjälpmedel för att finna resursen etc.

**3) Information om bevarande ("preservation"):** Ex.: dokumentation om resursens fysiska tillstånd, vilka åtgärder som ev. gjorts för att bevara en resurs, såväl fysiskt som digitalt, m.m. Denna typ av metadata är speciellt viktig för museer och arkiv.

**4) Teknisk:** Ex.: hårdvaru- och mjukvaruinformation, filformat, data om säkerhetsrutiner (t.ex. lösenord) etc.

**5) Användning:** användningsnivå och typ av användning av resursen. Ex.: utställningsinformation (för museiföremål och konstverk), information om återanvändande av innehållet i andra versioner av resursen etc.<sup>25</sup>

Christine L. Borgman, professor i Information Studies vid University of California, har utifrån detta gjort ytterligare differentieringar. Hon lägger fram sex användningsområden för metadata, som dock i mångt och mycket överensstämmer med Gilliland-Swetlands. De fyra första kategorierna ovan är i princip identiska hos Borgman. Hon har sedan ersatt *användning* med två egna kategorier, som hon kallar "intellectual access" och "intellectual organization".<sup>26</sup> "Intellectual access" innebär metadata som beskriver resursens innehåll, det vill säga ämnesindexering. Borgman menar alltså att denna aspekt är så viktig att den kräver en egen kategori. I de ovan citerade indelningarna återfinns ämnesord i den beskrivande typen av metadata. Att Borgman trycker på denna aspekt beror på att hon identifierar funktioner för metadata för användning i digitala bibliotek. Sålunda är det ett exempel på hur användningskontexten påverkar vilka aspekter av metadata som lyfts fram. Onekligen är indexering och beskrivning av resursens innehåll viktig när det gäller att söka och återfinna en resurs i ett visst ämne. "Intellectual organization" innebär hos Borgman nödvändigheten av standardisering av hur innehåll i olika metadataelement uttrycks. Detta för att kunna säkerställa ett identiskt användande och därmed kunna organisera en digital "samling" enligt standardiserade kriterier.<sup>27</sup>

---

<sup>25</sup> Gilliland-Swetland 2000.

<sup>26</sup> Borgman 2000, s. 75.

<sup>27</sup> Ibid., s. 75f.

Att det finns så många olika modeller och indelningar av metadata kan man säga är ett tecken på att, som Borgman uttrycker det, ”the communities have not yet agreed on a common model, and that metadata is a fast-moving area of research and practice”.<sup>28</sup>

### 2.2.2 Metadata – standard, schema eller format?

I litteraturen kring metadata används olika begrepp för ett regelsystem såsom till exempel Dublin Core Metadata Element Set. Omväxlande används termerna *standard*, *schema* (på engelska *scheme*) och *format*. Såvitt vi kan se brukas termerna i princip som synonymer. Det förefaller inte vara vanligt att man försöker definiera termerna eller differentiera mellan dem. Ett försök att definiera *schema* och *format* (begreppet *standard* uppmärksammar de inte) gör dock Stefan Sjölund och Elon Wismén i uppsatsen *Dublin Core: ett schema för metadata*. De framhåller att litteraturen i stort sett behandlar begreppen synonymt, men menar att en distinktion är möjlig. Sjölungs & Wisméns definition: ”Vi reserverar ordet ’schema’ för det utrymme som medges för val av metadata, medan vi låter ordet ’format’ stå för det sätt på vilket metadata uttrycks.”<sup>29</sup> Därefter kopplar de schema till semantik, det vill säga innebörden i själva beskrivningselementen och format till syntax, alltså på vilket sätt elementen uttrycks. För vår del har vi svårt att hålla med om Sjölungs & Wisméns definition. För det första kan vi som sagt i litteraturen inte se något stöd för distinktionen mellan schema och format. Än mer problematiskt är att likställa format med syntax, enligt vår mening. Syntax innebär visserligen hur metadatan uttrycks, till exempel i HTML, XML eller RDF (mer om detta kommer i avsnitt 2.4.2), men termen metadataformat används inte på detta sätt, utan man brukar då hellre tala om just syntax.

Vi har inte som ambition att på något vis slutgiltigt definiera dessa begrepp. Men vi vill påpeka att det kan vara dags att börja göra detta i högre grad. Framför allt gäller det begreppet *standard*, vilket vi anser används oklart. Standard bör vara förbehållet ett regelsystem som av en internationell organisation som till exempel *International Organisation of Standardization* (ISO) eller en motsvarande organisation inom en viss domän, erkänns och officiellt antas som standard. För närvarande tycks begreppet snarare antas och användas av de organisationer och projekt som konstruerar uppsättningar av metadataelement. Det riskerar att urholka begreppet. Vi kommer ändå att använda begreppet *standard* i vår uppsats, något vi strax återkommer till.

---

<sup>28</sup> Ibid., s. 70.

<sup>29</sup> Sjölund, Stefan & Wismén, Elon, 1999. *Dublin Core: ett schema för metadata*. Magisteruppsats i Biblioteks- och informationsvetenskap, Högskolan i Borås, s. 17.

Vad gäller termerna format och schema har vi som tidigare påpekats svårt att se någon distinktion mellan dem. Vi väljer att i uppsatsen inte använda termen format, främst för att det inte ska förväxlas med format i betydelsen filformat för en elektronisk resurs. Vi kommer sålunda använda termerna *standard* och *schema* på följande vis: *standard* brukar vi för uppsättningen metadataelement så som den är konstruerad, i teorin. Termen *schema* låter vi stå för en uppsättning metadataelement som den används (eller är tänkt att användas) i praktiken. Ofta sammanfaller båda dessa betydelser. Dublin Core Metadata Element Set är både en standard och ett schema. Men det kan hända att en lokal tillämpning av Dublin Core gör inskränkningar eller ändringar i dess regler. Denna tillämpning är då ett metadataschema, men inte en standard, enligt vår definition. Vi menar att vi har fog för denna definition inte minst genom framväxten av application profiles, som har sin grund just i distinktionen mellan teori och praktik inom metadatafältet. Mer om detta tar vi upp i avsnitt 2.4.

### 2.3 Dokument, resurser och Document-Like-Objects (DLO)

Dokumentet är den fysiska representationen av ett verk. För inte särskilt länge sedan omfattade begreppet dokument till allra största del tryckta medier såsom böcker och tidskrifter. Då det på senare år uppkommit en mängd nya materialtyper har begreppet utökats till att innefatta även bildmedier, auditiva medier, elektroniska dokument, multimedier o.s.v. Strunck, Lund & Thorlund Jepsen skriver:

Et bestemt dokument repræsenterer alle de fysiske objekter, som har de samme karakteristika i form af intellektuelt indhold og fysisk form. Når et værk realiseres, kan den resulterende repræsentation af værket lagres fysisk på papir, lydbånd, videobånd, lærred, gips, hard disk o.s.v. Denne fysiske lagring udgør en fiksering af værket i form af et dokument. Dokumentet kan eksistere i ét eksemplar, når der er tale om f.eks. en forfatters originale manuskript, en lydoptagelse i et historisk arkiv, et maleri etc. I andre tilfælde kopieres eller reproduceres dokumentet således at det kan publiceres og gøres til genstand for salg osv. [---] (O)m et dokument kopieres i et eller flere eksemplarer, er der dog stadig tale om det samme dokument. [---] Når en produktionsproces indebærer en ændring i den fysiske form, kan vi sige, at det resulterende produkt er et nyt dokument.<sup>30</sup>

---

<sup>30</sup> Strunck et al 1998. Strunck och Thorlund Jepsen är ass. fagleder vid Danmarks Biblioteksskole, Lund är lektor vid Danmarks Biblioteksskole.

En relativt ny form av dokument är de elektroniska. Dessa kan inte relateras till något bestämt exemplar i ett visst bibliotek och är lättare att ändra och flytta. Detta har gjort att det har uppstått ett behov av nya termer. *Informationsresurs*, *elektronisk resurs* eller enbart *resurs* har blivit accepterade som samlande beteckning för ett dokument eller en informationstjänst som gjorts tillgänglig i maskinläsbar form.<sup>31</sup>

På Dublin Cores hemsida ger man dock termen *resurs* en mycket bredare definition:

A resource is anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), and a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources.<sup>32</sup>

Dublin Core utvecklades i första hand för beskrivning och återvinning av enkla HTML-dokument, s.k. *document-like objects* (DLO:s). Detta slogs fast på Dublin Cores första workshop 1995. Dock utarbetade man då aldrig en klar definition av vad som räknas som DLO. Det finns många som har försökt förklara vad termen innebär och det varierar ganska mycket mellan olika definitioner. Carl Lagoze, forskare i Information Science vid Cornell University, skriver såhär om document-like objects:

The essence of a DLO is simplicity in structure and lifecycle; the DLO abstraction does not address issues such as compound sub-parts (e.g., chapters, sections) nor complex inter-relationships with other resources, physical or digital. The image of stand-alone objects described by static one-stop catalog records is perhaps better suited to shelves of books than to the Web -- few Web pages are stand-alone items, especially resources such as databases and video streams. On the other hand, the DLO is useful as a simple metaphor for characterizing the variety of Web resources that form the corpus for so-called cross-domain resource discovery. Treating a cross-section of resources as uniformly simple is a useful fiction that makes it possible to: 1) make simple statements about them with uniform structure, and 2) use these statements to search across the resources in a simple and uniform manner.<sup>33</sup>

---

<sup>31</sup> Hedberg [odat.].

<sup>32</sup> Dublin Core Metadata Initiative. URL: <http://www.ukoln.ac.uk/metadata/dcmi/dcq-html>

<sup>33</sup> Lagoze, Carl, 2001. "Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description?", *D-Lib Magazine*, January 2001, Volume 7, Number 1.

## 2.4 Vad är en application profile?

Vi ska nu närmare gå in på begreppet *application profile* och vad det innebär. Först är det dock nödvändigt att förklara ett annat begrepp, nämligen *namespace*.

### 2.4.1 Namespaces

En *namespace* är en funktion utvecklad av *World Wide Web Consortium* (W3C) inom ramen för *Extensible Markup Language* (XML). W3C:s definition av en XML namespace lyder: ”An XML namespace is a collection of names, identified by a URI reference, which are used in XML documents as element types and attribute names.”<sup>34</sup> Duval, Hodgins, Sutton och Weibel framhåller att namespaces är en fundamental del av webbens infrastruktur och nödvändiga för att kunna hantera och tillämpa metadata utformad som application profiles, som de även kallar ”modular metadata”.<sup>35</sup> De ger exempel på vad en namespace är:

For example, the base protocol of the Web is HTTP, which is a namespace that guarantees that a given URI is globally unique. LCSH (Library of Congress Subject Headings) is a namespace managed by the U.S. Library of Congress according to rules governing the assignment of subject headings to intellectual artifacts. Any metadata element set is a namespace bounded by the rules and conventions determined by its maintenance agency.<sup>36</sup>

Rachel Heery och Manjula Patel, verksamma inom *UK Office for Library and Information Networking* (UKOLN), tar upp tre funktioner hos en namespace:

- 1) Identifierar den instans som handhar och kontrollerar en viss uppsättning element (eller en kontrollerad ordlista eller klassifikationssystem).
- 2) Understödjer definition för att unikt identifiera en uppsättning element.
- 3) Definierar unikt den bestämda uppsättningen av element eller den ordlista som används.

---

<sup>34</sup> Se W3C:s rekommendation för användning av XML Namespaces: <http://www.w3.org/TR/REC-xml-names/>

<sup>35</sup> Duval, E., Hodgins, W., Sutton, S. & Weibel, S., 2002. ”Metadata Principles and Practicalities”, *D-Lib Magazine*, Vol. 8, N.o 4, April 2002. Erik Duval är professor vid Computer Science Dept. vid Katholieke Universiteit Leuven, Belgien. Wayne Hodgins är chef för Worldwide Learning Strategies vid Autodisk. Stuart Sutton är Associate Professor i Information Studies vid Syracuse University. Stuart Weibel är chef för Dublin Core Metadata Initiative och verksam vid OCLC.

<sup>36</sup> Duval et al 2002.

För att klargöra distinktionen mellan dessa tre funktioner exemplifierar Heery och Patel med den modell för att registrera metadatascheman som utvecklades i ett projekt kallat DESIRE (siffrorna här motsvarar siffrorna ovan)<sup>37</sup>:

- 1) Registration authority.
- 2) Namespace concept.
- 3) Namespace.

Denna modell är tänkt att vara en hjälp för att kontrollera att metadatascheman är väl hanterade. Den fungerar i princip hierarkiskt: en namespace identifierar en uppsättning element. Uppsättningen kan finnas i olika versioner, vilka alla kräver en egen namespace. Alla versioner är dock samlade i gruppbezeichnung ”namespace concept” och både namespace och namespace concept är knutna till registreringsauktoriteten.<sup>38</sup> Vi kan förtydliga med hur dessa tre funktioner ser ut med Dublin Core som exempel:

- 1) Dublin Core Metadata Initiative (registration authority).
- 2) Dublin Core Metadata Element Set (namespace concept).
- 3) Dublin Core Metadata Element Set Version 1.1 (namespace).

Identifieringen sker alltså genom en URI (*Uniform Resource Identifier*), vilken uttrycks med en webbadress (URL) eller URN (se förkortningar). Denna identifiering gäller dels beskrivningselementen, dels om man inom ett element hänvisar till en kontrollerad ordlista, tesaurus eller klassifikationssystem av något slag. Även dessa måste sålunda identifieras via en URI.

Låt oss säga att vi i en application profile använder oss av ett antal av Dublin Cores element. För att kunna påvisa att elementen härrör just ifrån Dublin Core och för att profilens metadata rent tekniskt ska fungera måste vi i XML-syntaxen ange följande URL:

xmlns:dc = ”<http://purl.org/dc/elements/1.1>”

---

<sup>37</sup> Heery, Rachel & Patel, Manjula, 2000. ”Application profiles: mixing and matching metadata schemas”, *Ariadne*, Issue 25, September 2000.

<sup>38</sup> Heery & Patel 2000.



URL:en anger här att vi använder oss av Dublin Core Metadata Element Set, Version 1.1. På samma sätt måste alla andra element och termer identifieras.

### **2.4.2 Application profiles**

Begreppet *application profile* började diskuteras ordentligt år 2000, framför allt inom Dublin Core Metadata Initiative och en del projekt som var relaterade till dess verksamhet, som till exempel SCHEMAS-projektet, som är ett EU-finansierat initiativ och en del av *Information Society Technologies Programme* (IST). Kortfattat kan man säga att syftet med detta projekt är att skapa ett forum för personer och organisationer som praktiskt arbetar med att skapa metadatascheman för olika projekt. På SCHEMAS hemsida finns samlat ett antal scheman eller just *application profiles*. Meningen är att de som arbetar med att utforma metadata här ska kunna registrera egna scheman och undersöka andras, för att till exempel se hur andra har gjort för att lösa vissa problem.<sup>39</sup>

Vad som alltmer blev tydligt i arbetet med detta och andra projekt var enligt Heery och Patel att man kan urskilja två olika förhållningssätt till konstruerandet och upprätthållandet av metadatascheman. Å ena sidan finns ”standards makers”, de som främst är intresserade av att utforma en fast och sammanhållen uppsättning element, vilken kan bli betraktad som en standard. Det viktigaste målet för dem som tillhör denna kategori är modellens integritet och standarden i sig.

Å andra sidan finns ”implementors”, vilka är praktiker som förvisso använder standarder, men gör det på sitt eget sätt för att anpassa schemat till de specifika resurser och det projekt det är ämnat för. Heery och Patel framhåller att denna dikotomi visserligen till stor del är falsk, eftersom många som arbetar med metadatautveckling kan sägas tillhöra båda sidorna, men de vill ändå betona distinktionen, för att belysa de skillnader i prioritering som faktiskt existerar.<sup>40</sup>

Med andra ord har profiler utvecklats på grund av att man i det praktiska arbetet har upptäckt att olika projekt, resurssamlingar och ämnesområden har behov som enskilda standarder inte kan tillgodose. I stället för att skapa en helt egen ny standard för varje nytt projekt, så ser man *application profiles* som ett enklare och mer pragmatiskt sätt att angripa problemet. Som

---

<sup>39</sup> För mer information se hemsida: <http://www.schemas-forum.org>

<sup>40</sup> Heery & Patel 2000.

Makx Dekkers uttrycker saken: "[...] re-inventing the wheel so to speak, is not the optimal way of working."<sup>41</sup>

Heery och Patel definierar en application profile såhär: "We define application profiles as schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application."<sup>42</sup> Definitionen är något teknisk, och det kan vara klargörande att se på ytterligare en definition. Duval, Hodgins, Sutton & Weibel uttrycker det på ett enklare vis: "An application profile is an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema."<sup>43</sup> Detta förklarar vad det i grunden handlar om. I stället för att använda en enda given metadatastandard i ett visst projekt, där man vill beskriva ett antal resurser, kan man alltså välja beskrivningselement från olika standarder och sätta ihop dem till ett eget metadataschema.

Heery och Patel tar också upp några regler för application profiles. Man får inte skapa nya beskrivningselement som inte finns i en existerande namespace. Vill man skapa ett eget element, måste man även skapa en egen namespace, där man definierar elementet och tar ansvar för innehållet i namespaces. Några friheter som tillåts en skapare av en profil är till exempel att specificera om en kontrollerad vokabulär ska användas för att uttrycka innehållet i ett visst element. Man behöver då inte hålla sig till de vokabulärer som eventuellt finns specificerade i den metadatastandard man använder, utan kan välja någon helt annan, som bättre passar ens syften. Man kan också specificera vad ett elements innebörd ska vara, det vill säga elementets semantik. Här måste man dock vara försiktig, eftersom man inte får förändra innebörden helt och hållet, bara göra den "smalare", mer specifik.<sup>44</sup>

Thomas Baker, verksam vid German National Research Center for Information Technology, har identifierat regler eller snarare nödvändigheter som han menar att en profil bör uppfylla. De faller inom fyra olika kategorier:

- Definition of entity classes in the data model that underlies the application, identifying

---

<sup>41</sup> Dekkers, Makx, 2001. "Application Profiles, or how to Mix and Match Metadata Schemas", *Cultivate Interactive*, Issue 3, January 2001.

<sup>42</sup> Heery & Patel 2000.

<sup>43</sup> Duval et al 2002.

<sup>44</sup> Heery & Patel 2000.

the type or types of resources the application profile schema applies to, e.g. people, Web pages, books, image galleries;

- Formal declarations of elements and their semantics used by the application, including rules for their usage, e.g. declaring which elements are mandatory, optional, repeatable, which element combinations are allowed or mandated and what allowable formats for the values of elements are;
- Expression of controlled vocabularies for the value of elements, e.g. specifying which controlled vocabulary, classification scheme or thesaurus may be used as values for a particular element or restricting the allowable values for a particular element to an enumerated set;
- Human readable information about the application and usage guidance<sup>45</sup>

Det är med andra ord viktigt att elementen och deras innehåll definieras tydligt, för att man ska kunna försäkra sig om att profilen hanteras och används på ett kontrollerat och enhetligt sätt.

En för vårt vidkommande viktig funktion för en application profile uttrycks av Duval, Hodgins, Sutton & Weibel:

One of the benefits of this approach is that communities of practice are able to focus on standardizing community-specific metadata in ways that can be preserved in the larger metadata architectures on the Web. It will be possible to snap together such community-specific modules to form more complex metadata structures that will conform to the standards of the community while preserving cross-community interoperability.<sup>46</sup>

Det blir alltså i en profil möjligt för mycket ämnes- eller domänspecifik metadata att fungera interoperabelt med andra domäner i en webbmiljö. Denna funktion är något som naturligtvis är centralt för vårt syfte.

När det gäller själva arbetsprocessen med profiler, tycks det inte vara vanligt med en detaljerad dokumentation. Makx Dekkers, chef för Pricewaterhouse Coopers Consulting i

---

<sup>45</sup> De fyra kategorierna har sammanställts av Makx Dekkers. Se Dekkers 2001.

<sup>46</sup> Duval et al. 2002.

Luxemburg samt medlem i Dublin Core Advisory Committee, har emellertid utifrån workshops i SCHEMAS-projektet urskiljt en fyrstegsmodell för hur arbetet (oftast) går till. Den ser ut som följer:

- 1) Definiera vilka behov metadatan skall tillgodose.
- 2) Välj den mest lämpliga existerande metadatastandarden, och välj vilka element som täcker behoven.
- 3) Där det är möjligt, använd element från en existerande standard för lokala behov.
- 4) Definiera återstående element i en egen namespace.<sup>47</sup>

Dekkers menar att steg 1 och 2 är de avgörande. I praktiken tas de inte alltid i denna ordning, utan ofta väljs metadatastandard redan innan behoven har definierats. Detta behöver inte vara fel. I vissa fall kan en standard vara dominerande inom en viss domän eller ett ämnesområde. Det kan vara strategiskt och praktiskt välmotiverat att utgå från denna standard vid sådana tillfällen. Vi kommer för övrigt att anknyta till Dekkers modell i analysavsnittet (6.1) i uppsatsen.

Innan vi övergår till att visa några exempel på application profiles, ska vi bara helt kort nämna något om nödvändigt tekniskt ramverk för att profiler ska kunna fungera på webben. Ett sådant ramverk som väckt intresse inom metadatavärlden är *Resource Description Framework* (RDF), utarbetat av World Wide Web Consortium. RDF är en tillämpning av XML-syntaxen och fungerar som ett slags behållare eller struktur i vilken metadata kan infogas. RDF har även kallats en standard för metadata, men är ingen standard i den bemärkelse vi talar om i denna uppsats. Vi citerar Ulf Kronman och John Parnefjord och hänvisar den intresserade till dem för vidare läsning:

RDF är, precis som namnet anger, inte en metadatastandard i sig, utan ett utbyggbart *ramverk* för metadatastandarder. Inom detta ramverk kan man använda olika standarder för metadata som i RDF kallas för *scheman*. En finess med RDF är att man också kan använda flera olika scheman i samma beskrivning, något som gör att RDF blir utbyggbart och därigenom anpassningsbart för olika behov.<sup>48</sup>

---

<sup>47</sup> Dekkers 2001.

<sup>48</sup> Kronman, Ulf & Parnefjord, John, 2001. ”Resource Description Framework – metadata för framtidens Internet”, *Tidskrift för dokumentation*, 2001:1, s. 15-25. Citatet på s. 17. Kronman är webbutvecklare och Parnefjord bl.a. programmerare vid Karolinska Institutets Biblioteks webbyrå Vision. För en introduktion i RDF,

### 2.4.3 Några exempel på application profiles

Inom Dublin Core Metadata Initiative har arbetsgrupper tagit fram ett antal application profiles. Ett exempel är *DC Education*, som är en profil tänkt för beskrivning av resurser inom utbildningsområdet. Denna profil använder sig av Dublin Cores 15 element, men arbetsgruppen har även tagit fram två egna element som man har definierat i en egen namespace på Dublin Cores hemsida. Dessutom ingår tre element från metadatastandarden *IEEE Learning Object Metadata* (IEEE LOM), som är en domänspecifik standard för utbildningsresurser. En annan profil är *DC Library Application Profile* (DC-LAP), vilken som namnet antyder, ska användas i bibliotek och biblioteksrelaterade projekt. *DC Library* använder 13 av Dublin Cores 15 element, men har också 14 egna element, kallade *DC Library Metadata Element Set*. Intressant att notera är att man i profilen också återanvänder ett element från *DC Education*:s namespace.<sup>49</sup> Sålunda kan en application profile använda element som skapats specifikt för en annan application profile, och inte enbart element från standarder. Huvudsaken är ju att elementen finns definierade i en namespace.

Även utanför själva Dublin Core-organisationen baseras ofta en application profile på Dublin Cores uppsättning av beskrivningselement. Så är fallet med det EU-finansierade projektet Renardus, som pågick från januari 2000 till juni 2002, där partners från flera olika europeiska länder deltog, däribland Lunds universitet. Renardus-projektets syfte var att samla kvalitetsresurser på Internet från olika domäner och ämnesområden i ett gemensamt gränssnitt för möjlighet till samsökning och browsing. Renardus använder sig av en tämligen liten application profile, med bara åtta beskrivningselement, varav sju är hämtade från Dublin Core. Det sista är ett eget projektspecifikt element definierat i en egen namespace.<sup>50</sup>

---

se även Medeiros, Norm, 2000. "XML and the Resource Description Framework", *Online*, Vol. 24, Issue 5, Sep/Oct 2000, s. 37-40.

<sup>49</sup> För *DC Education*:

<http://dublincore.org/documents/2000/10/15/education-namespace>

För *DC Library*:

<http://dublincore.org/documents/2001/08/08/library-application-profile>

Båda profilerna finns även registrerade på hemsidan för SCHEMAS:

<http://www.schemas-forum.org/registry>

<sup>50</sup> Neuroth, Heike & Koch, Traugott, 2001. "Metadata Mapping and Application Profiles. Approaches to providing the Cross-searching of Heterogeneous Resources in the EU Project Renardus".

Jane Hunter, forskare i bl.a. metadata för multimedia vid University of Queensland, Australien, har presenterat en profil som kombinerar Dublin Core-element med element från standarden MPEG-7, med syfte att beskriva videoresurser. MPEG-7 är en standard som är mycket detaljerad och djupgående och utvecklad för multimediaresurser. Hunters profil är intressant därför att hon använder Dublin Core för dess fördelar, att till exempel understödja mediaoberoende återfinning av resurser över domän- och ämnesgränser. Hon framhåller att det är bättre att använda Dublin Core för den bibliografiska informationen för en videoresurs. MPEG-7 har en komplex struktur även för den typen av information. Element från MPEG-7 används mer effektivt för att uttrycka mediespecifika detaljer såsom tids- och rumskomponenter inom den audiovisuella resursen.<sup>51</sup>

Man kan alltså konstatera att det är vanligt att Dublin Core bildar grund i application profiles, och att en domänspecifik standard får komplettera, alternativt att egenskapade element läggs till profilen. Även om många profiler sinsemellan är mycket olika och skapas för varierande syften, skulle man kunna säga att denna kombination av typer av standarder är ett generellt drag.

## 2.5 Kulturarv (Cultural Heritage)

Vad innebär begreppet kulturarv? Hur ska det definieras? Detta är ett mycket mångfasetterat område och begreppet kan ges många olika betydelser och innebörder. Kulturarv kan t.ex. vara både *materiellt* och *immateriellt*. Det materiella, fysiska kulturarvet är påtagligt i form av exempelvis byggnader, konstverk, litteratur, film, föremål och fornlämningar. Det kan också vara immateriellt och ta sig uttryck i religion, språk, historia, traditioner, livsformer, idéer och samhällsstrukturer. Det finns även en uppdelning i *konstnärliga* kulturarv som teater och musik respektive *folkliga* som slöjd, berättande och dans.<sup>52</sup>

Den definition av kulturarv som ECHO-projektet använder kommer från Världsbanken:

Cultural heritage encompasses material culture, in the form of objects, structures, sites and landscapes, as well as living (or expressive) culture as evidenced in forms such as music, crafts, performing arts, literature, oral tradition and language. The emphasis is on cultural continuity

---

<sup>51</sup> Hunter, Jane, 2002. "An Application Profile which combines Dublin Core and MPEG-7 Metadata Terms for Simple Video Description".

<sup>52</sup> Skolverket, dokument för *Projektet Kultur för lust och lärande*.

from the past, through the present and into the future, with the recognition that culture is organic and evolving.<sup>53</sup>

EU-parlamentet lägger fram en utvidgad definition som innefattar nya former som t.ex. fabriker, maskiner, yrken som dött ut och till och med folkliga matrecept. Man tar även upp bevarandet av språk och dialekter som löper risk att försvinna.<sup>54</sup>

### **2.5.1 ECHO-projektet**

Uppslaget till inriktning på vår uppsats har vi fått genom ECHO-projektet (*European Cultural Heritage Online*) som vi kom i kontakt med genom Institutionen för lingvistik vid Lunds universitet. Det är ett EU-finansierat projekt inom kulturarvsområdet med syfte att integrera forskningsdata inom olika ämnen och göra dem tillgängliga och samsökbara på webben. Projektet går i sin första fas ut på att tillgängliggöra språkresurser inom lingvistisk forskning. I ett längre perspektiv skall andra ämnesområden tillkomma, till att börja med konsthistoria, vetenskapshistoria och antropologi. I ett ännu längre perspektiv finns en vision om betydligt fler ämnen från institutioner runt hela Europa, sammanlänkade i en digital sökrymd.<sup>55</sup> Målgruppen för ECHO är ganska bred. Man formulerar det på följande vis: ”ECHO will provide web-accessible multimedia content together with navigation facilities, hence making it attractive for researchers, teachers, students, journalists, and also for the general public user.”<sup>56</sup>

---

<sup>53</sup> Världsbanken, dokument: *Culture and Development Action Network. Working Group Meeting Brief, January 26-27, 1998.*

<sup>54</sup> EU-parlamentet, Utskottet för kultur, ungdomsfrågor, utbildning, medier och idrott. Dokument PE 286.688/1-76, 15 november 2000.

<sup>55</sup> *European Cultural Heritage Online (ECHO)*. 30 september 2002. Projektbeskrivning. (Opublicerat dokument.)

<sup>56</sup> *Ibid.*, s. 6.

## 3. Metadatastandarderna Dublin Core och IMDI

### 3.1 Dublin Core Metadata Initiative (DCMI)

<http://dublincore.org>

Dublin Core, som antagligen är den mest kända och spridda metadatastandarden, är en enkel standard innehållande 15 baselement eller *core elements*. Elementen är i första hand framtagna för att beskriva webbaserade resurser och göra dessa sökbara. Alla elementen är valbara och inga är obligatoriska, utan man väljer de element som man anser passar den resurs som skall beskrivas. Man får även upprepa ett element flera gånger i samma beskrivning. Exempelvis kan fältet *DC:Creator* upprepas om det finns flera upphovsmän till samma resurs. Standarden är utarbetad för att vara lättanvänd och lättillgänglig. Detta gör det möjligt för den enskilde producenten av ett webbdokument att själv kunna indexera det hon vill lägga ut på nätet och därmed göra det sökbart. Den semantiska vokabulär som *Dublin Core Metadata Element Set* tillhandahåller syftar till att beskriva ”kärn”-informationen i en resurs såsom ”innehåll”, ”upphovsman” och ”datum”, element som kan användas oberoende av ämnesspecificitet eller typ av organisation. Denna generalitet syftar även till att göra det möjligt att söka över domängränserna och kan ses som ett verktyg för att främja interoperabilitet mellan olika ämnesspecifika metadataformat. Dublin Cores generella baselement kan också ses som en utgångspunkt när det gäller att utveckla mer specifika metadataavokabulärer. Dublin Core-metadata läggs för det mesta in i själva dokumentet, men kan även läggas in i ett separat dokument.

Dublin Core började utvecklas efter en workshop för utveckling av den semantiska webben som hölls i Dublin, Ohio i mars 1995. Här samlades ett 50-tal personer från olika yrkesområden, bland annat bibliotekarier, arkivarier, nätverkstekniker och databasspecialister. Diskussionen gällde hur man skulle kunna utveckla en semantik som kunde användas för att kategorisera webbresurser och därigenom underlätta sökbarhet och återfinning. Resultatet blev ”Dublin Core metadata”, döpt efter platsen för workshopen. Sedan dess har det hållits ytterligare åtta workshops runt om i världen för att diskutera Dublin Cores utveckling och framtid.



Fördelar och mål med Dublin Core:

**Enkelhet:** Dublin Core skall kunna användas både av beskrivningsspecialister och av människor utan katalogiseringserfarenhet. De flesta elementen i Dublin Core har en allmän semantik som inte skiljer sig särskilt mycket från ett vanligt katalogkort.

**Semantisk interoperabilitet:** För beskrivning av Internetresurser används många olika beskrivningsmodeller, vilket gör att det är svårt att söka över domängränserna. Dublin Core är en generell metadatastandard och kan därför användas till att förena andra standarder och på så sätt hjälpa till att skapa semantisk interoperabilitet mellan olika områden.

**Internationell konsensus:** För att kunna återfinna webbresurser på internationell nivå krävs en väl fungerande infrastruktur. Dublin Core används i drygt 20 länder i Nordamerika, Europa, Australien och Asien och fungerar därför som en internationell standard.

**Möjligheter till utvidgning:** Dublin Core är mer allmänt och enklare uppbyggt än många andra mer detaljerade beskrivningsmodeller. Detta gör att möjligheterna till flexibilitet och utvidgning blir större genom att det är möjligt att ta in en mer sammansatt struktur och semantik från rikare beskrivningsstandarder.

### ***3.1.1 Dublin Core Metadata Element Set Version 1.1***

Detta är de 15 element som utgör Dublin Core. Definitionerna är hämtade från Dublin Cores hemsida och översatta av oss.

1. **Title:** Resursens namn.

2. **Creator:** Resursen upphovsman, den person eller organisation som är ansvarig för resursens innehåll.

3. **Subject:** Ämnesbeskrivning av resursens innehåll, kan innehålla både kontrollerad vokabulär, klassifikationssystem eller fria nyckelord.

4. **Description:** En kort redogörelse för resursens innehåll i fritext.

5. **Publisher:** Den enhet som svarar för att göra resursen tillgänglig, t.ex. ett förlag eller en universitetsinstitution.
6. **Contributor:** Någon som medverkat till resursens innehåll men som inte nämns i creator-fältet, t.ex. medförfattare eller illustratör.
7. **Date:** Datum förknippade med resursen, t.ex. tillkomstdatum, publiceringsdatum eller bäst-före-datum.
8. **Type:** Genre eller typ av innehåll i resursen, t.ex. hemsida, roman eller uppsats. Bör väljas från en lista med fördefinierade alternativ, *Dublin Core Types* (DCT1).
9. **Format:** Resursens fysiska eller digitala form. Bör väljas från en lista med fördefinierade alternativ.
10. **Identifier:** En unik identifikation av resursen, t.ex. en URL eller ett ISBN-nummer.
11. **Source:** Information om källan bakom resursen, varifrån den härstammar.
12. **Language:** Språket för resursens intellektuella innehåll.
13. **Relation:** En referens till en relaterad resurs.
14. **Coverage:** Resursens geografiska eller tidsmässiga täckning.
15. **Rights:** Information om rättigheterna till resursen.

### ***3.1.2 Dublin Core Qualifiers***

Dublin Cores 15 baselement kan specificeras med hjälp av *Dublin Core Qualifiers*. Det finns två kategorier av qualifiers, dels "element refinements", dels "encoding scheme". Ett "encoding scheme" består av en kontrollerad vokabulär, till exempel Dewey Decimal Classification, som används för att få en standardisering av de ord och begrepp som läggs in i beskrivningen i vissa element (se exempel 1 nedan). "Element refinements" innebär en avgränsning av innebörden i ett element (se exempel 2). Eftersom Dublin Core-elementen är

så öppna kan man behöva begränsa den semantiska vidden. Genom att på detta sätt skapa en struktur i den annars plana standarden ökar man precisionen där Dublin Cores element inte är helt tillräckliga för att göra en tillfredsställande beskrivning (detta kan dock minska elementens möjlighet till samsökbarhet mellan olika standarder/domäner). Qualifiers används alltså för att ge ett element ett nytt värde eller för att specificera informationen i elementet.

### **Exempel 1.**

Elementet *DC:Coverage* med qualifier med kontrollerad vokabulär, s.k. "encoding scheme":

*DC:Coverage.Temporal*: US civil war era; 1861-1865 (uttrycks enligt DCMI Period)

*DC:Coverage.Spatial*: Columbus (C,V) (uttryckt enligt TGN, *The Getty Thesaurus of Geographic Names*)

### **Exempel 2.**

Elementet *DC>Date* med qualifier med s.k. "element refinement":

*DC>Date.Created*: 2003-02-02

*DC>Date.Issued*: 2003-03-01

### **Exempel 3.**

Elementet *DC:Title* med qualifier:

*DC:Title*: Hamlet in Iceland; being the Icelandic romantic Ambales saga

*DC:Title.Alternative*: Ambales saga

(Se även tabell över qualifiers på sidan 36.)

Samma qualifier får aldrig upprepas inom ett och samma element. När det gäller kontrollerad vokabulär är det även tillåtet att använda sig av vokabulär från andra scheman för lokala eller domänspecifika applikationers behov. Om man inte vill förlora i interoperabilitet är det dock önskvärt att man använder registrerade kontrollerade ämnesord. Om man väljer en uppsättning kontrollerade ämnesord som inte är registrerade hos Dublin Core Metadata Initiative är det möjligt att registrera dem på organisationens hemsida.

En regel när man använder sig av qualifiers är att använda dem på ett sådant sätt att det är möjligt att, trots att man tappar i specificitet, läsa beskrivningen utan att ta hänsyn till ”qualifiern”, och ändå få fram relevant information om den beskrivna resursen. Detta kallas ”the Dumb-Down Principle”. Första förslaget till Dublin Core-qualifiers kom 1997 och de diskuteras fortlöpande och är fortfarande under utveckling.

På följande sida visas en tabell över Dublin Cores qualifiers. I den första kolumnen listas Dublin Cores 15 baselement. I den andra kolumnen står de ”element refinements” som Dublin Core rekommenderar för att avgränsa respektive baselements innebörd. I den tredje kolumnen finns de kontrollerade vokabulärer som Dublin Core rekommenderar för att uttrycka innehållet i elementen i kolumn 1 på ett standardiserat sätt.

## Tabell över Dublin Core Qualifiers, fastställda 2000-07-11

DCMES Element	Element Refinement(s)	Element Encoding Scheme(s)
Title	Alternative	-
Creator	-	-
Subject	-	LCSH MeSH DDC LCC UDC
Description	Table Of Contents Abstract	-
Publisher	-	-
Contributor	-	-
Date	Created Valid Available Issued Modified	DCMI Period W3C-DTF
Type	-	DCMI Type Vocabulary
Format	Extent	-
	Medium	IMT
Identifier	-	URI
Source	-	URI
Language	-	ISO 639-2 RFC 1766
Relation	Is Version Of Has Version Is Replaced By Replaces Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format	URI
Coverage	Spatial	DCMI Point ISO 3166 DCMI Box TGN
	Temporal	DCMI Period W3C-DTF
Rights	-	-

### 3.2 IMDI (ISLE Metadata Initiative)

<http://www.mpi.nl/ISLE>

Metadatastandarden IMDI har utarbetats för språkresurser. Det gjordes år 2000 som en del av projektet ISLE (*International Standards for Language Engineering*), som har en europeisk och en amerikansk gren. IMDI har utarbetats av projektets europeiska gren.<sup>57</sup> Medansvariga är också EAGLES (*Expert Advisory Group on Language Engineering Standards*), ett initiativ under Europakommissionen, vars syfte är att arbeta fram en digital infrastruktur för språk och språkresurser av olika slag. Man ska ta fram standarder, riktlinjer och teknologiska ramverk.<sup>58</sup>

Ett antal institutioner i Europa är knutna till EAGLES/ISLE:s metadataprojekt IMDI och i organisationen ingår främst forskare inom lingvistik och experter på lingvistikresurser. Ansvaret för hemsidan och en nyckelroll har Max-Planck-Institute for Psycholinguistics i Nijmegen, Holland. Ordförande i ISLE är Peter Wittenburg, som är teknisk chef vid sagda institut.

Anledningen till att man velat utveckla en metadatastandard specifikt för språkresurser är framförallt två saker: den ena är den explosiva ökningen av antalet resurser som har skett under senare år. Wittenburg och Broeder exemplifierar i uppsatsen *Metadata Overview and the Semantic Web* med situationen vid Max-Planck-Institute i Nijmegen, där 40 forskare varje år ger sig ut på fältundersökningar och gör ett stort antal inspelningar (s.k. ”sessions”) på video och/eller audio och att man nu (2002) har ca 10 000 sådana sessions i en online-databas.

Den andra anledningen är att samtidigt som man ser den kvantitativa ökningen av resurserna har resurserna i sig blivit alltmer komplexa och varierande i fråga om typ och format. Från att tidigare till största delen ha handlat om textbaserade resurser har ljud, bild och multimedia tillkommit i allt högre grad. Praktisk erfarenhet på institutet visar att forskarna har svårt att finna relevanta resurser och man står, som Wittenburg och Broeder uttrycker det, inför ”a serious resource management and discovery problem”.<sup>59</sup>

---

<sup>57</sup> Wittenburg, Peter & Broeder, Daan, 2002. *Metadata Overview and the Semantic Web*.

<sup>58</sup> ”Introduction to the EAGLES initiative” [odat.].

<sup>59</sup> Wittenburg & Broeder 2002.

En aspekt som bör betonas är att skaparna av IMDI uttryckligen vill att standarden ska användas för att skapa en sökrymd av sammanlänkade språkresurser på webben som är allmänt tillgänglig. Dessa resurser är i första hand intressanta för forskare inom olika discipliner, förutom lingvistik även datalingvistik, artificiell intelligens, antropologi, psykologi m.m. Men man föreställer sig även andra användare, såsom lärare, journalister och "the casual web-user".<sup>60</sup> Med andra ord, allt ifrån den vanlige webbsurfaren till den specialintresserade forskaren. En berömvärd ambition, men det sätter fingret på den aktuella problematiken inom metadatafältet idag, nämligen hur specialiserad metadata ska kunna integreras och göras användbar i ett större sammanhang på webben och hur alla olika användargrupper ska tillfredsställas.

### **3.2.1 Språkresurser – definition och konstitution**

För att kunna beskriva hur IMDI är uppbyggd är det nödvändigt att först förklara vad en språkresurs kan bestå av. Wittenburg, Broeder & Sloman definierar begreppet på följande sätt: "Language resources' are those databases which primarily document communicative acts of humans by some form of recording and/or descriptions, both directly as in corpora, or at higher levels of abstraction in lexicons and ontologies."<sup>61</sup>

Korpus (i plural corpora eller korpusar) är ett begrepp som ursprungligen innebär en samling autentiska texter. Begreppet används inte enbart inom lingvistik, utan även inom till exempel juridik, historia och teologi. Inom lingvistik har korpus på senare tid också börjat användas om annat än textsamlingar, såsom andra sorters registreringar av språkbruk – nedskrivna eller inspelade samtal, videofilmer, ögonrörelsemätningar m.m. Men det rör sig alltså även i dessa fall om en datasamling med autentiskt språkbruk. I modern lingvistik innebär det också att datasamlingen/korpusen är maskinläsbar.

Det finns många olika typer av korpusar, vilka alltså kan vara olika former av inspelat tal eller rena textsamlingar. Det kan röra sig om böcker, tidningar, dagböcker, offentliga tal etc, eller en datasamling inom en viss domän, helt beroende på forskarens eller institutionens inriktning. Som exempel på korpusar kan nämnas några av de som finns på hemsidan för W3-Corpora (*The World Wide Web Access to Corpora Project*): "The Air Traffic Control

---

<sup>60</sup> Wittenburg, P., Broeder, D., Sloman, B., 2000. *Meta-Description for Language Resources. A Proposal for a Meta-Description Standard for Language Resources. EAGLES/ISLE White Paper*. Se även Wittenburg & Broeder 2002.

<sup>61</sup> Wittenburg, Broeder & Sloman 2000.

Corpus”, ”Wellington Corpus of Spoken New Zealand English” (WSC), ”Japanese Speech Corpora of Major City Dialects” och ”CALLHOME Collection (Unscripted telephone conversations in six languages)”.<sup>62</sup>

Korpuslingvistik är att studera och analysera lingvistiska fenomen genom dessa datasamlingar eller korpusar. Forskaren kan undersöka existerande korpusar eller själv sammanställa en. Han/hon söker belägg i faktiskt språkbruk (inom varierande områden), något som kan ligga till grund för till exempel konstruktion av ett lexikon eller en grammatik.

Datortekniken har gjort att korpuslingvistiken har utvecklats mycket under de senaste decennierna. Att datasamlingar av språk kan göras maskinläsbara har möjliggjort att materialet snabbare och lättare kan hittas och genom olika verktyg presenteras i ett format som är lämpat för analysen.<sup>63</sup>

IMDI är uppdelat i tre olika scheman, ett för inspelningssessioner / korpusar (*IMDI Metadata Elements for Session Descriptions*), ett för lexikon (*IMDI Metadata Elements for Lexicon Descriptions*) och ett för publicerade korpusar (”published corpora”), som man kallar katalogmetadata (*IMDI Metadata Elements for Catalogue Descriptions*). Med publicerade korpusar menas helt enkelt att de finns öppet tillgängliga, idag vanligen via webben. Flera inspelningssessioner kan bilda en korpus och skillnaden gentemot katalogmetadata är att där beskrivs själva korpusen i sig, inte inspelningarna. I denna uppsats kommer vi att uteslutande behandla metadata-schemat för inspelningssessioner / korpusar, det vill säga Session Descriptions.

Det bör påpekas att IMDI Session Descriptions i första hand är tänkt för ljud-, bild- och multimediaresurser, vanligtvis som sagt inspelningar av olika former av språk. En grundläggande svårighet vid beskrivning av språkresurser inom området korpusar är att det kan finnas flera dimensioner av en och samma resurs, vad man kallar ”bundled resource”. Det innebär att en multimedial resurs kan innehålla följande: en inspelning (session) i både video och audio. Till att börja med finns själva inspelningen och dess innehåll (”the linguistic

---

<sup>62</sup> W3-Corpora är ett projekt som drivs av Department of Language and Linguistics vid University of Essex. Syftet är att tillgängliggöra korpusar på webben och tillhandahålla en introduktion i området korpuslingvistik. URL: [http://www.essex.ac.uk/linguistics/clmt/w3c/corpus\\_ling/content](http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content)

<sup>63</sup> För definitioner och beskrivning av korpusar och korpuslingvistik har vi använt W3-Corpora:s hemsida samt även fått hjälp av professor Sven Strömquist, Institutionen för lingvistik, Lunds universitet, och Marcus Uneson, amanuens vid samma institution.



event”), vilka som medverkar, vem som gjort inspelningen etc. Sedan kan det finnas flera olika videospår och ljudspår, samt inspelningar av till exempel ögonrörelser eller kroppsrörelser. Alla dessa kan således finnas på egna filer. Olika aspekter av en session kan vara av intresse för olika forskare.

Till detta kommer dessutom annotationer. Annotationer är manuellt eller automatiskt genererade textbeskrivningar. De manuella kan vara skrivna antingen i fritext eller genererade enligt en uppsättning koder. Att annotera ett material kallas ofta att ”tagga” det, eller ”märka upp” det. Det görs med hjälp av ett s.k. ”markup language”, till exempel SGML eller XML, eller en speciellt utarbetad kodning, som CES (*Corpus Encoding Standard*), som tagits fram av EAGLES. Annotationerna kan göras av helt olika skäl och alltså variera beroende på ämne och forskningsinriktning. Vanligen är dessutom annotationerna knutna till en viss del av inspelningen, som ett ljudspår eller ett videospår.<sup>64</sup>

Det säger sig självt att när en resurs av det här slaget ska beskrivas på ett så tillfredsställande sätt som möjligt, ställer det krav på mångdimensionalitet i metadatastrukturen.

### ***3.2.2 IMDI Metadata Elements for Session Descriptions Version 2.5***

[http://www.mpi.nl/ISLE/documents/docs\\_frame.html](http://www.mpi.nl/ISLE/documents/docs_frame.html)

Här följer en schematisk överblick över elementen i IMDI Session Descriptions. Av utrymmesskäl och på grund av att vi i vår avgränsning valt bort dem, har vi beslutat att inte redovisa *Sub-schemas* (ett antal scheman som ingår i vissa element och ligger under dem hierarkiskt), inte heller elementen för annotationer, *AnnotationUnit*. Definitionerna är översatta från IMDI, men ofta kommenterade och förtydligade av uppsatsförfattarna.

Eftersom IMDI är komplicerad kan en inledande sammanfattning av uppbyggnaden vara av nöden. I denna ingår inte heller de delar vi valt bort i avgränsningen. IMDI består av nio grupper av element med ett antal underelement, sammanlagt 70 stycken. Av de nio grupperna kan man säga att fem innehåller information som direkt hör till resursen eller inspelningen: *Session*, *Collector*, *Content*, *Participants*, *Media File*. De fyra andra grupperna innehåller information som i egentlig mening ligger utanför själva resursen: *Project*, *Source*, *Anonymous*

---

<sup>64</sup> Wittenburg, Broeder & Sloman 2000. Här även uppgifter från Marcus Uneson, Institutionen för lingvistik, Lunds universitet.

och *References*. Hierarkiskt sett finns det beskrivningselement på tre nivåer i IMDI (plus *Sub-schema*, vilket ligger på den lägsta nivån, som oftast är den fjärde, men ibland även den femte), men inga qualifiers i Dublin Cores bemärkelse. Med hjälp av hierarkin möjliggörs en djupgående beskrivning.

*Förklaring till tecken och förkortningar efter namnen på elementen:*

\* Elementet är obligatoriskt.

+ Indikerar en lista på ett eller flera element.

*string* Innehållet i elementet skrivs i en sekvens av alfanumeriska symboler inklusive mellanrum och interpunktion.

*sub* Till elementet hör ett sub-schema.

*group* Elementet är en gruppering av andra element.

*c* Elementet är bundet till ett specifikt kodningsschema, t.ex. ISO-standard.

*ccv (closed controlled vocabulary)* Innehållet i elementet måste väljas från en fördefinierad uppsättning värden.

*ov (open vocabulary)* Innehållet i elementet kan väljas från en fördefinierad föreslagen uppsättning värden eller kan definieras av användaren.

*ovl (open vocabulary list)* En lista med värden för innehållet i elementet kan väljas från en fördefinierad uppsättning värden eller kan definieras av användaren.

Här följer IMDI:s beskrivningselement:

(Element som är indragna åt höger står i en hierarkisk relation till det närmast ovanstående element som står längre ut åt vänster.)

*Session.Name (string) \**: Ett namn, ofta en förkortning av titeln, för att identifiera sessionen.

*Session.Title (string)*: Den fullständiga titeln på sessionen.

*Session.Date (c)*: Datumet då sessionens data skapades (t.ex. då en inspelning gjordes). Uttrycks enligt ISO-standard ISO8601.

***Location*** (group): Gruppering av information som har att göra med platsen där inspelningssessionen skapades. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Continent (ccv)*: Den kontinent där inspelningssessionen skapades. Uttrycks enligt en ordlista utformad av IMDI.

*Country* (ccv): Det land där inspelningssessionen skapades. Uttrycks enligt ISO-standard ISO3166-1.

*Region* + (string): Den region där inspelningssessionen skapades.

*Address* (string): Den adress där inspelningssessionen skapades (t.ex. en institution).

*Session.Description* + (sub): En beskrivning av sessionen i stort. Här ges en kortare beskrivning av den situation som sessionen visar. Exempel: "Ett samtal mellan mor, far och ett barn vid frukostbordet."

*Session.Keys* (sub): Nyckelord som uttrycks i par (*name* + *value*), för att beskriva domänspecifik information om sessionen. Exempel: "length = 182", "colour – red". Detta element är särskilt viktigt för stora projekt som behöver utveckla mycket domänspecifika nyckelord. Det är meningen att forskaren här också ska tillhandahålla en länk till en kontrollerad ordlista för de specifika nyckelorden.

**Project** (group): Gruppering av information om det projekt för vilket inspelningssessionen skapades. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Name* (string): Ett namn, ofta en förkortning av titeln, för att identifiera projektet.

*Title* (string): Projektets fullständiga titel.

*Id* (string): En unik referens för att identifiera projektet, oftast uttryckt genom något formellt system, t.ex. en URI.

*Contact* (sub): Information om person eller organisation ansvarig för projektet, och hur de kan kontaktas.

*Description* + (sub): En beskrivning av projektets mål och omfattning.

**Collector** (group): Gruppering av information om sammanställaren (skaparen) av inspelningssessionens data. Det är inte nödvändigtvis samma person som är ansvarig för projektet, eller som till exempel utför en intervju i en inspelning. I vissa fall kan dessa tre roller sammanfalla i en person, i vissa fall kan det röra sig om två eller tre olika personer. Denna kategori är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Name* (string): Namnet på personen som är ansvarig för sammanställandet (skapandet) av inspelningssessionens data.

*Contact* (sub): Information om personen ansvarig för inspelningssessionen, och hur han/hon kan kontaktas.

*Description* + (sub): Ytterligare information om personen ansvarig för inspelningssessionen.

**Content** (group): Gruppering av information om innehållet i inspelningssessionen. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Communication Context* (group): Denna kategori används för att beskriva den kommunikativa kontext i vilken inspelningen gjordes. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Interactivity* (ccv): Beskriver graden av interaktivitet mellan de deltagande personerna i inspelningen. Uttrycks genom en IMDI-specifik sluten kontrollerad vokabulär ("closed controlled vocabulary").

*Planning Type* (ccv): Indikerar hur mycket den/de deltagande i inspelningen ("the consultant") har planerat den lingvistiska aktiviteten ("the linguistic event"). Uttrycks genom en IMDI-specifik sluten kontrollerad vokabulär. Exempel: de tre termer som finns i vokabulären är "spontaneous", "semi-spontaneous" och "consultant/performer-planned".

*Involvement* (ccv): Indikerar hur mycket forskaren var involverad i den lingvistiska aktiviteten. Uttrycks genom en IMDI-specifik sluten kontrollerad vokabulär.

**Genre** (group): Denna kategori används för att beskriva innehållets diskursiva typ eller form. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Interactional* (ovl): Beskriver innehållets interaktionstyp. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: konversation, debatt, intervju, hälsningar och avsked etc.

*Discursive* (ovl): Beskriver innehållets diskurstyp. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: de två termer som finns i ordlistan är "Procedure" och "Explanation". Procedure definieras som en beskrivning av proceduren när man förbereder eller tillverkar någonting (t.ex. "how to make tortillas"). Förklaring definieras som ett uttalande om en praktisk, teoretisk eller historisk verklighet.

*Performance* (ovl): Beskriver innehållets typ av framförande. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: "song", "oral history", "narrative", "insult" etc.

**Task** (ocv): Beskriver den typiska uppgift eller situation ("task") som utförs/utspelas i inspelningssessionen. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: det finns en mycket typifierad terminologi för att beskriva dessa situationer, som "room reservation", "frog story", "wizard of oz", "travel planning" etc.

*Modalities* (ocv): Beskriver de uttryck eller aspekter av den inspelade lingvistiska aktiviteten som forskaren är intresserad av. Exempel: "speech", "gestures", "eye gaze", "facial expressions" etc.

*Languages* (group): Gruppering av information om de språk som används i inspelningssessionen. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Language* + (sub): En lista på de språk som används i inspelningssessionen. Det språk som huvudsakligen används nämns först. Uttrycks enligt IMDI:s Language Identifier Encoding, som innehåller flera tillåtna alternativ, som t.ex. ISO639 och RFC1766. Varje språk beskrivs närmare i Sub-schemat.

*Description* + (sub): En kortare beskrivning av de språk som används i inspelningen. Som ovan nämnts, beskrivs varje språk närmare i Sub-schemat för Content.Languages.Language.

*Description* + (sub): En fritextbeskrivning av innehållet i inspelningssessionen.

*Keys* (sub): Nyckelord som uttrycks i par (*name* + *value*), för att beskriva domänspecifik information om innehållet i inspelningen. Exempel: "length = 182", "colour – red", "Hausa – fluent", "English – basic". Detta element är särskilt viktigt för stora projekt som behöver utveckla väldigt domänspecifika nyckelord. Det är meningen att forskaren här också ska tillhandahålla en länk till en kontrollerad ordlista för de specifika nyckelorden.

***Participants*** (group): Gruppering av information om de deltagande i inspelningssessionen. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Description* + (sub): En beskrivning av relationer och interaktioner mellan de deltagande. Varje enskild deltagande beskrivs närmare nedan, under Participants.Participant.Description.

*Participant* (group): Gruppering av information om varje enskild deltagande i inspelningssessionen. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Type* (ov): Kategorisering av den deltagande i inspelningen. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: "interviewer", "consultant" etc.

*Name* + (string): Namn som används om den deltagande för att identifiera honom/henne. Det kan vara en förkortning eller pseudonym.

*Full name* (string): Den deltagandes fullständiga namn.

*Code* (string): En unik kod som används för att identifiera den deltagande. Det kan vara en förkortning eller pseudonym. Koderna

används främst i transkriptioner och annotationer för att identifiera de delar av inspelningen som hör till en specifik deltagande.

*Role* (ov): Den deltagandes roll i inspelningssessionen. Uttrycks genom en IMDI-specifik öppen vokabulär. Exempel: om en familj intervjuas, så ska i detta element relationerna specificeras. T.ex. mor, far, son, dotter etc.

*Language* + (sub): En lista över de(t) språk den deltagande behärskar. Det första språket i listan är den deltagandes modersmål. Uttrycks enligt IMDI:s Language Identifier Encoding. Varje språk beskrivs närmare i Sub-schemat.

*Ethnic Group* (string): Den deltagandes etniska tillhörighet.

*Age* (c): Den deltagandes ålder. Uttrycks enligt Codes for the Human Analysis of Transcripts (AGECHAT).

*Sex* (ccv): Den deltagandes kön. Uttrycks enligt IMDI:s slutna kontrollerade ordlista.

*Education* (string): Den deltagandes utbildningsnivå. Kan också användas för att uttrycka den deltagandes läskunnighet eller brist på sådan.

*Anonymous* (ccv): Används för att uttrycka om den deltagandes namn och fullständiga namn har ersatts av pseudonymer för att säkerställa dennes anonymitet. Uttrycks enligt slutna kontrollerad ordlista (värdet True innebär att namnen är pseudonymer, False att de inte är det).

*Description* + (sub): En beskrivning av den deltagande.

*Keys* (sub): Nyckelord som uttrycks i par (*name* + *value*), för att beskriva domänspecifik information om innehållet i inspelningen. Exempel: "length = 182", "colour – red", "Hausa – fluent", "English – basic". Detta element är särskilt viktigt för stora projekt som behöver utveckla mycket domänspecifika nyckelord. Det är meningen att forskaren här också ska tillhandahålla en länk till en kontrollerad ordlista för de specifika nyckelorden.

**Media File** + (group): Gruppering av information om själva mediafilen (resursen). Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Resource Link* (c): En länk till filen. Uttrycks genom en URL.

*Size* (string): Storleken på filen. Uttrycks vanligen i bytes.

*Type* (ccv): Resurstyp. Inte att förväxla med filformat. Uttrycks genom slutna kontrollerade ordlista, med termer från *Multipurpose Internet Mail Extensions* (MIME). Exempel: Audio, Video, Image etc.

*Format* (ov): Filformat. Uttrycks genom en öppen ordlista, med termer från *Multipurpose Internet Mail Extensions* (MIME). Exempel: WAV, MPEG, JPEG etc.

*Quality* (ccv): En numerisk indikator på kvaliteten på filen. Uttrycks som en skala från 1 till 5, där 1 står för låg och 5 för hög kvalitet. Bedömningen blir här förstås till stor del subjektiv.

*Recording Conditions* (string): Beskriver de tekniska omständigheter under vilka inspelningen för filen gjordes. Exempel: vilken utrustning som användes, vilken typ av mikrofoner och förstärkare etc.

*Position* (c): Start- och slutposition av en inspelningssession på filen. Det kan hända att en inspelningssession endast är en del av innehållet i en fil. Uttrycks enligt IMDI-specifik slutna kontrollerade ordlista, *Media Position Encoding*.

*Access* (sub): Specificerar villkor för tillgång till filen, samt eventuella rättigheter till innehållet.

*Description* + (sub): En närmare beskrivning av filen.

**Source** +: Gruppering av information om källan, varifrån filen har fått sitt innehåll. Det handlar ofta om videoband eller ljudband. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Id* (string): En kod för att identifiera källan. Det kan vara en hyllplacering i ett arkiv eller liknande.

*Format* (ov): Medietyp där ursprungsinspelningen finns. Exempel: CD, CD-ROM, DVD, Reel etc.

*Quality* (ccv): En numerisk indikator på kvaliteten på inspelningen. Uttrycks som en skala från 1 till 5, där 1 står för låg och 5 för hög kvalitet. Bedömningen blir här förstås till stor del subjektiv.

*Position* (c): Start- och slutposition av en inspelningssession på bandet / CD:n / DVD:n (eller vilken medietyp det nu rör sig om). Det kan hända att en inspelningssession endast är en del av innehållet. Uttrycks enligt IMDI-specifik slutna kontrollerade ordlista, *Media Position Encoding*.

*Access* (sub): Specificerar villkor för tillgång till källan, samt eventuella rättigheter till innehållet.

*Description* + (sub): En närmare beskrivning av källan.

**Anonymous** (group): Gruppering av information om fil där anonyma deltagandes pseudonymer konverteras till deras riktiga namn. Det är inte ett element i sig självt, utan måste följas av något av de följande underelementen.

*Resource Link* (c): Länk till konverteringsfilen. Uttrycks genom en URL.

*Access* (sub): Specificerar villkor för tillgång till konverteringsfilen, samt eventuella rättigheter till innehållet.

**References** (group): Gruppering av information om dokumentation och inspelningar som är relaterade till denna inspelningssession. Det är inte ett element i sig självt, utan måste följas av underelementet.

*Description* + (sub): Beskriver eventuella relaterade inspelningar och dokumentation.

### 3.2.3 IMDI:s ordlistor

Många av IMDI:s element skall innehålla ett eller flera värden uttryckta enligt specifikt utarbetade ordlistor, ”vocabularies”, som man kallar dem. IMDI skiljer på två typer av ordlistor, ”closed controlled vocabulary”, en sluten ordlista, och ”open vocabulary”, en öppen ordlista. De slutna ordlistorna tillhandahålls och kontrolleras av IMDI och är obligatoriska, vilket innebär att elementets värde måste väljas från den specifika listan. Att ändra eller lägga till i en sluten ordlista är förbehållet IMDI. De öppna ordlistorna är snarare förslag på värden. Ordlistan rekommenderas av IMDI, men det är tillåtet för den person som lägger in metadata att använda egna termer.<sup>65</sup>

### 3.2.4 Förändringar och versioner av IMDI

I november 2000 publicerades den första versionen av IMDI, *IMDI Metadata Elements For Session Descriptions, Version 2.0*. Sedan dess har standarden reviderats fem gånger, och den senaste officiella versionen är 2.5, publicerad i juni 2001. Ytterligare två revisioner har dock gjorts, 2.7 i juli 2002, som endast var för internt bruk och version 3.0 från november 2002, som i skrivande stund (februari 2003) ännu inte har publicerats.<sup>66</sup>

---

<sup>65</sup> *IMDI (ISLE Metadata Initiative). Part 1 A. Metadata Elements for Session Descriptions. Version 2.5.* June 2001.

<sup>66</sup> *IMDI (ISLE Metadata Initiative). Part 1. Metadata Elements for Session Descriptions. Draft Proposal Version 3.0.* November 2002. ”Appendix B: Revision history”, s. 39. Opublicerat dokument.



Att gå igenom alla förändringar som skett är knappast nödvändigt, men vi skulle vilja nämna ett par som vi har funnit vara av intresse, bland annat eftersom den version som har mappats till Dublin Core är en äldre, version 2.2. Den största skillnaden mellan 2.2 och 2.5 är att det gruppement som beskriver de deltagande i inspelningen, *Participants*, har förändrats. I stället för att specificera de olika roller som kan finnas i en inspelning, som *Researcher*, *Consultant*, *Contributory*, som egna element, ersattes dessa element med elementet *Participant*. Sedan tillfördes underelementen *Type* och *Role*, för att kunna specificera den deltagandes roll. Version 2.5 innehåller också ett separat Sub-schema med en fast uppsättning element och underelement. Tidigare angavs dessa element inom varje element i huvudschemat.

Den viktigaste förändringen i version 3.0 är att ett helt nytt gruppement har tillkommit, kallat *Written Resources*. Som namnet avslöjar är gruppementet avsett för textresurser, vilket faktiskt har varit standardens svaghet tidigare. Den har varit fokuserad på audiovisuella resurser. En annan förändring är att gruppementet *Collector* tagits bort. Sammanställaren eller skaparen av resursen läggs i version 3.0 in i *Participant.Type*, det vill säga som alla andra deltagande i resursen. Ett för biblioteksvärlden intressant tillägg finns under gruppementet *Content*. Ett underelement *Subject* har där tillkommit. Där rekommenderar IMDI användandet av en kontrollerad vokabulär som till exempel Library of Congress Subject Headings (LCSH). Värt att nämna är också att IMDI:s ordlista för elementet *Content.Genre* har förenklats. Man skriver i en kommentar: "In the IMDI metadata set for Sessions 2.5, an elaborate system for classifying the content of sessions was in place. Discussions with linguists actually indicated it was too complex. We propose in this draft a simplification of the content description scheme."<sup>67</sup> Det är ett exempel på att en dialog med dem som faktiskt ska arbeta med metadatan är viktig för att en metadatastandard ska utvecklas och formas på bästa sätt. Och inte minst tjänar det som fingervisning om nödvändigheten att vara medveten om att enkelhet ibland kan vara att föredra framför komplexitet, även för en ämnesspecifik standard.

---

<sup>67</sup> *IMDI (ISLE Metadata Initiative). Part 1. Metadata Elements for Session Descriptions. Draft Proposal Version 3.0.* November 2002, s. 30.

## 4. Beskrivning av en språkresurs i Dublin Core och IMDI

I följande avsnitt ska vi beskriva en språkresurs i respektive Dublin Core och IMDI, för att närmare kunna visa på en del skillnader mellan dem. Resursen är en audio- och videoinspelning av en man kallad John Doe, förkortat J., som är bosatt i Nigeria och tillhör den etniska gruppen Goemai, vars språk också heter Goemai. Han talar i inspelningen om olika typer av träd och användningen av dem. Resursexemplet är taget från en befintlig inspelning gjord av en forskare vid Max-Planck-Institute i Nijmegen. Namnen på de deltagande personerna har vi ändrat och fritextbeskrivningarna är förkortade. All text i beskrivningen är på engelska. Vi har inom parentes lagt in kommentarer till vissa beskrivningsselement.

### 4.1 Beskrivning av en språkresurs i Dublin Core

*Title:* The uses of trees

*Creator:* Eva Svensson

*Subject:* Goemai

*Subject:* Speech, pointing-gestures

(Ämnesorden som finns här under *Subject* är inte Dublin Core-rekommenderade. Dublin Core rekommenderar att ord från en kontrollerad vokabulär används, som t.ex. LCSH (Library of Congress Subject Headings). *Speech* och *pointing gestures* är definierade av IMDI som en öppen ordlista. Här blir de dock fria nyckelord.)

*Contributor:* John Doe.

(Att nämna John Doe, som är en deltagande och intervjuad i inspelningen, som medskapare till resursens intellektuella innehåll kan synas tveksamt. Vi har valt att nämna honom i detta element för att kunna jämföra med IMDI:s beskrivning i analysen.)

*Description:* Audio and video recording of a linguistic event. J describes the various trees in the vicinity. He names the trees, explains their uses, and compares different types of trees to each other. The researcher's interest in this task was to prompt the use of demonstratives in a semi-natural setting.

*Date.Created:* 2001-06-25.

(Vi valde att använda en Dublin Core Qualifier *Created* för att betona att det är tidpunkten för skapandet av resursen som åsyftas.)

*Type:* Sound.

(Dublin Core har en ordlista med rekommenderade termer för elementet *Type*. Ingen av dem är heltäckande för denna resurs. "Sound" syftar på en inspelning med ljud som primär typ. Men detta är ju en audiovisuell inspelning.)

*Format:* WAV.

*Identifier:* TreesSM\_1.wav

(Eftersom exemplet är hämtat från en intern databas är elementets värde inte uttryckt som en URL-adress.)

*Language:* Goemai; ISO639-2:ank

(Språket är uttryckt enligt ISO-standard som står för Goemai.)

*Coverage:* Africa

*Coverage:* Nigeria

*Coverage:* Kwande

(Innehållet i *Coverage* bör enligt Dublin Core uttryckas i standardiserad vokabulär.)

## 4.2 Beskrivning av en språkresurs i IMDI

*Session Name:* TreesSM

*Session Title:* The uses of trees

*Session Date:* 2001-06-25

### ***Location:***

*Continent:* Africa

*Country:* Nigeria

*Region:* Quán Pan Local Government Area  
Kwande  
Beyond the football field

### *Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* Description of the uses of trees. We looked at trees in four different sites...

### *Keys (sub-schema):*

IMDI file creator – Eva  
Check – TreesSM\_3.mpg is lost

### ***Project***

*Project Name:* Goemai

*Project Title:* Documentation of Goemai

#### *Contact*

*Name:* Eva Svensson

*Adress:* Postbus 310, 6500 AH Nijmegen.

*E-mail:* E-mail@mpi.nl

*Organisation:* Max-Planck-Institut für Psycholinguistics

### *Description (sub-schema)*

*Language:* RFC1766.x-sil-eng

*Text:* The Goemai project is part of...

### ***Collector***

*Name:* Eva Svensson

*Contact (sub-schema)*

*Name:* Eva Svensson

*Address:* Postbus 310, 6500 AH Nijmegen.

*E-mail:* E-mail@mpi.nl

*Organisation:* Max-Planck-Institut für Psycholinguistics

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* The collector of the session data does not appear on the participant screen...

**Content**

*CommunicationContext*

*Interactivity:* Non-interactive

*Planning Type:* Semi-spontaneous

*Involvement:* Non-elicited

*Genre*

*Discursive:* Explanation

*Task:* Frog-story

*Modalities:* Speech, pointing-gestures

*Languages*

*Language (sub-schema)*

*Name:* Goemai

*Id:* RFC1766:X-sil-ank

*Description (sub-schema)*

*Text:* Goemai is a Chadic language which is spoken by approximately 200.000 speakers...

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* Goemai is the only language that is used in the session...

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* J. describes the various trees in the ...

*Link:* TreesSMSketch.pdf

*Keys (sub-schema):* Topic - demonstrative

Topic - flora

## ***Participants***

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* J. was the only participant in the session...

*Participant*

*Type:* Consultant

*Name:* J.

*Full name:* John Doe

*Code:* J.

*Role:* Son

*Language (sub-schema)*

*Name:* Hausa

*Id:* RFC1766:x-sil-hua

*Description (sub-schema)*

*Text:* Hausa is a Chadic language...

*Name:* Goemai

*Id:* RFC1766:x-sil-ank

*Description (sub-schema)*

*Text:* Goemai is a Chadic language...

*Name:* English

*Id:* RFC1766:x-sil-eng

*Description (sub-schema)*

*Text:* English is the national language of Nigeria...

*Ethnic group:* Goemai

*Age:* 24

*Sex:* Male

*Education:* Secondary school

*Anonymous:* True

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* J. is one of my main collaborators in Kwande...

*Keys:* Dialect – Kwo  
Hausa – fluent  
English – basic

## ***Resources***

### *Media-File*

*Resource Link:* TreesSM\_1.wav

*Size:* 7000 KB

*Type:* Audio

*Format:* Wav

*Quality:* 3

*Recording condition:* Stereo

*Position:* (Start/end position of session on the media file.)

*Access:* Not available

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* This Wav-file was converted from an MPEG file...

### *Source*

*Id:* LBHGVD15jun01-2

*Format:* DV

*Quality:* 3

*Position:* (Position of session on the tape.)

*Access:* Not available

*Description (sub-schema):* (Description of the tape.)

## **References**

*Description (sub-schema)*

*Language:* RFC1766:x-sil-eng

*Text:* The sessions TreesLL and TreesAS have a similar content...

### **4.3 Kommentar till beskrivningarna**

Här ska endast vissa grundläggande skillnader påpekas i de båda beskrivningarna av samma språkresurs. Det första man kan konstatera är att IMDI:s beskrivning är mycket mer omfattande och oerhört detaljerad i förhållande till Dublin Core. Till exempel finns i IMDI inte mindre än 12 olika fritextbeskrivningar av olika aspekter av resursen, samt tio olika aspekter av elementet *Language*, språk. Det sistnämnda kanske inte är så konstigt, eftersom det handlar om språkresurser och forskning i språk. Men det är ändå uppenbart att mycket information går förlorad i Dublin Core. Främst är det svårt i Dublin Core att få en kvalificerad beskrivning av ämnesspecifika partier med lingvistisk terminologi, och beskrivning av den deltagande / intervjuade personen saknas i princip helt. Han finns medtagen endast som *DC:Contributor*, vilket egentligen är något missvisande, enligt vår mening. I hög grad saknas också administrativ metadata i Dublin Core, som beskrivning av själva filen och källan till inspelningen, och därtill hörande rättigheter. Å andra sidan är den typen av metadata inte den primära när det gäller att söka och återfinna material.



## 5. Application profile för beskrivning av språkresurser

### 5.1 Målgrupp och syfte

Det är viktigt att man analyserar och klargör målgrupp för och syfte med profilen. Detta påverkar förstås utformandet. I vårt fall utgår vi från ECHO-projektets definierade målgrupp, vilken har presenterats i 2.5.1 ovan. ECHO beskriver sin målgrupp som forskare, lärare, studenter, journalister, samt vad man kallar ”the general public user”.<sup>68</sup> Enligt vår mening är detta en mycket bred målgrupp, kanske alltför bred. Det blir svårt att samtidigt tillfredsställa både forskare och den stora allmänheten. Vi vill differentiera något mellan målgrupperna och menar att vår application profile i första hand vänder sig till forskare, lärare, studenter och journalister, det vill säga de grupper som kan tänkas söka resurser för arbete eller studier. Den målgrupp som sammanförs i uttrycket allmänheten definierar vi som användare som söker resurser på fritiden, med generellt sett lägre grad av intresse och kunskaper i ämnet. Vi utesluter dock inte denna målgrupp helt och hållet, utan menar att även allmänheten kan utnyttja den tänkta söktjänst eller databas som vår profil är utformad för.

Syftet med vår profil är att beskriva språkresurser på ett sådant sätt att de ger en tillräckligt ämnesspecifik beskrivning för att tillfredsställa i första hand forskare och andra yrkesintresserade, men även en intresserad allmänhet. Samtidigt skall profilen vara så generell att den möjliggör interoperabilitet med andra domäner i en webbaserad sökrymd.

Här vill vi klargöra att ovanstående syfte inte är att förväxla med syftet för hela uppsatsen. Syftet här tillhör resultatredovisningen och är en del av vår (fiktiva) application profile.

### 5.2 Redovisning av Application profile för beskrivning av språkresurser

Vår application profile använder sig av följande namespaces:

För Dublin Core-elementen:

*Dublin Core Metadata Element Set, Version 1.1*

<http://purl.org/dc/elements/1.1>

---

<sup>68</sup> ECHO. Projektbeskrivning 2002, s. 6.

För Dublin Core Qualifiers:

*Dublin Core Qualifiers* (2000-07-11)

<http://purl.org/dc/terms>

För Dublin Cores resurstyplista:

*DCMI Type Vocabulary*

<http://purl.org/dc/dcmitype>

För IMDI-elementen, IMDI qualifiers samt IMDI:s slutna och öppna ordlistor:

*IMDI Metadata Elements for Session Descriptions Version 2.5*

[http://www.mpi.nl/ISLE/documents/docs\\_frame.html](http://www.mpi.nl/ISLE/documents/docs_frame.html)

Här följer beskrivningselementen i vår application profile. Inget element är obligatoriskt.

Qualifiers får bara användas för de element där det uttryckligen sägs i kommentaren.

Qualifiers får inte upprepas i samma beskrivning om inget annat anges i kommentaren.

*DC:Title*

Dublin Cores definition gäller. Elementet är upprepningsbart.

*DC:Creator*

Dublin Cores definition gäller. Elementet är upprepningsbart.

*DC: Subject*

Qualifiers:

*Genre*

*Task*

*Modalities*

Till detta element hör tre qualifiers hämtade från IMDI:s element *Content*. Kontrollerad vokabulär hämtas från IMDI:s ordlista för gällande qualifiers. *Subject* får användas utan qualifiers. I detta fall rekommenderas LCSH men fria nyckelord är tillåtna. Elementet är upprepningsbart.

*DC:Description*

Dublin Cores definition gäller. Elementet är icke upprepningsbart.

#### *DC:Publisher*

Dublin Cores definition gäller. Elementet är upprepningsbart.

#### *DC:Date*

Dublin Cores definition för Date.Created gäller. Datumet skall gälla skapandet av resursen (inspelningstillfället). Elementet är icke upprepningsbart. Anges enligt ISO8601.

#### *DC:Type*

Dublin Cores definition gäller. Elementet är icke upprepningsbart. Innehållet uttrycks enligt Dublin Cores lista för resurstyper, Dublin Core Types (DCT1), eller enligt Multipurpose Internet Mail Extensions (MIME), som används av IMDI.

#### *DC:Format*

Dublin Cores definition gäller. Elementet är icke upprepningsbart. Om den beskrivna resursen innehåller flera filer beskrivs dessa separat. Format anges enligt Internet Media Types, som tillhandahålls av MIME.

#### *DC:Identifier*

Dublin Cores definition gäller. Elementet är icke upprepningsbart. Uttrycks oftast med en URL men även andra typer av URI är tillåtna.

#### *DC:Language*

Dublin Cores definition gäller. Elementet är upprepningsbart. Uttrycks enligt standarden RFC 1766 eller enligt ISO.

#### *DC:Relation*

Dublin Cores definition gäller. Elementet är upprepningsbart. Vi rekommenderar att elementets innehåll uttrycks genom någon typ av fastslagen standard (t.ex. URL eller ISBN) som ger en unik hänvisning till den relaterade resursen. Även fritext är tillåtet.

### *DC:Coverage*

Dublin Cores definition gäller. Elementet är upprepningsbart. Vi har valt att inte använda oss av qualifiers i detta element men rekommenderar att innehållet ändå uttrycks med Dublin Cores definierade ordlistor för elementets qualifiers Spatial och Temporal.

### *DC:Rights*

Dublin Cores definition gäller. Elementet är upprepningsbart.

### *IMDI:Participant.*

IMDI:s definition gäller. Elementet är upprepningsbart. Det måste användas med qualifiers.

### Qualifiers:

#### *Type*

IMDI:s definition gäller. Uttrycks enligt IMDI:s öppna ordlista.

#### *Full name*

IMDI:s definition gäller. Ingen ordlista används.

#### *Language*

IMDI:s definition gäller. Uttrycks enligt ISO-standard, ISO639-2 eller enligt RFC 1766. Denna qualifier är upprepningsbar.

#### *Ethnic Group*

IMDI:s definition gäller. Ingen ordlista används.

#### *Age*

IMDI:s definition gäller. Ålder uttrycks enligt IMDI:s regler med AGECHAT.

#### *Sex*

IMDI:s definition gäller. Uttrycks enligt IMDI:s slutna kontrollerade ordlista.

#### *Education*

IMDI:s definition gäller. Uttrycks i fritext.

### *Description*

Vi har valt att använda oss av den definition som ges för elementets underelement *Description.Text*, vilket innebär att innehållet uttrycks i fritext. Här kan forskaren lägga in övrig relevant information om den deltagande.

## **5.3 Beskrivning av en språkresurs med vår application profile**

*DC:Title:* The uses of trees

*DC:Creator:* Eva Svensson

*DC:Subject:* Goemai

*DC:Subject.Genre:* Explanation

*DC:Subject.Type:* Frog-story

*DC:Subject.Modalities:* Speech, pointing-gestures

*DC:Description:* Audio and video recording of a linguistic event. J describes the various trees in the vicinity. He names the trees, explains their uses, and compares different types of trees to each other. The researcher's interest in this task was to prompt the use of demonstratives in a semi-natural setting...

*DC>Date:* 2001-06-25

*DC:Type:* Audio

*DC:Format:* WAV

*DC:Identifier:* TreesSM\_1.wav

*DC:Language:* Goemai; ISO639-2:ank

*DC:Relation:* The sessions TreesLL and TreesAS have a similar content...

*DC:Coverage:* Africa

*DC:Coverage:* Nigeria

*DC:Coverage:* Kwande

*IMDI:Participant.Type:* Consultant

*IMDI:Participant.Full name:* John Doe

*IMDI:Participant.Language:* Goemai; ISO639-2:ank

*IMDI:Participant.Language:* Hausa; RFC1766:x-sil-hua

*IMDI:Participant.Language:* English; RFC1766:x-sil-eng

*IMDI:Participant.Ethnic Group:* Goemai

*IMDI:Participant.Age:* 24

*IMDI:Participant.Sex:* Male

*IMDI:Participant.Education:* Secondary school

*IMDI:Participant.Description:* J:s first language is Goemai. He also speaks Hausa fluently and basic English. J. is one of my main collaborators in Kwande...

#### **5.4 Kommentar till beskrivningen**

Eftersom vi tänker gå in djupare på profilen i analysavsnittet ska bara kort nämnas att vi som synes har koncentrerat oss på att komplettera Dublin Cores element med en mer ingående

beskrivning av de deltagande i inspelningar, IMDI:s element *Participant*, samt att erhålla mer ämnesspecifik terminologi under *DC:Subject*. I övrigt hänvisar vi till analysen.

## 6. Analys och diskussion

Vi ska i detta kapitel analysera och diskutera vår application profile och de problem som uppstått under arbetet med den. Vi har delat in analysavsnittet i underavdelningar, där vi diskuterar olika aspekter av profilen och arbetsprocessen. Många av problemen i ett arbete av det här slaget är relaterade till och påverkar varandra, vilket gör det svårt att skilja på vissa aspekter. Det kan därför hända att svårigheter med ett visst element tas upp flera gånger på olika ställen i analysen, men utifrån skilda infallsvinklar.

### 6.1 Allmän diskussion om profilen

I artikeln ”Application Profiles, or how to Mix and Match Metadata Schemas” (som vi tidigare tagit upp i 2.4.2) beskriver Makx Dekkers hur arbetsprocessen för utformandet av application profiles ofta ser ut. Han identifierar fyra steg:

1. Define metadata requirements
2. Select most appropriate existing standard metadata element set
3. Where possible, use standard elements for locally required elements, possibly narrowing semantics and adding local rules and vocabularies
4. Define remaining elements in private namespace<sup>69</sup>

Steg nummer 1, att definiera behoven, består som vi ser det av flera saker. Det handlar framför allt om att identifiera målgruppen för och syftet med profilen. Detta är naturligtvis något som varierar för varje enskilt projekt, och man kan inte säga att det finns några generella kriterier att utgå ifrån. Såsom tagits upp i 5.1 har vi för vår profil definierat målgrupp utifrån ECHO-projektet, och syftet är formulerat utifrån dessa målgrupper och ECHO:s strävan i ett längre perspektiv, att göra olika ämnesområdets resurser samsökbara.

I steg nummer 2, att välja den mest lämpliga metadatastandarden efter de behov man har identifierat, har vi valt Dublin Core och IMDI. Eftersom IMDI är den hittills mest utvecklade standarden för språkresurser och Dublin Core är den standard som de flesta application

---

<sup>69</sup> Dekkers 2001.



profiles baseras på, ansåg vi att dessa standarder var självklara val. Dekkers menar i sin artikel att det inte är ovanligt att steg 2 i praktiken kommer före steg 1, det vill säga att skapare av profiler utgår ifrån en eller flera standarder innan man definierar de specifika behoven. I synnerhet är det en vanlig strategi när det finns en standard som är dominerande inom ett visst ämnesområde eller domän.<sup>70</sup> Så är fallet med IMDI och i vår profil har vi utgått från valet av standarder, alltså börjat med steg 2.

Steg 3 innebär att man i första hand alltid bör välja element från existerande standarder för de lokala behoven, eventuellt med en snävare semantik och med lokala regler. Det finns några exempel på detta i vår profil, vilka diskuteras närmare längre fram.

Vad gäller det fjärde steget i Dekkers modell, att skapa ett eget namespace för element som inte finns i existerande standarder, har detta behov inte uppstått för oss.

## 6.2 Elementen i profilen

I vår application profile har vi 14 element, varav 13 är hämtade från Dublin Core. Två element från Dublin Core är inte med i profilen, *Source* och *Contributor*. *Source* valdes bort därför att vi ville att beskrivningen skulle gälla själva filen (som är den huvudsakliga resursen), inte källan varifrån filen härstammar. *Contributor* är onödig eftersom vi har IMDI:s element *Participant*, som beskriver de olika deltagarna i resursen mer ingående och på ett för syftet mer relevant sätt.

Av de 13 Dublin Core-elementen som vi använder är sju stycken definierade helt enligt Dublin Cores regler. Dessa sju element är: *Title*, *Creator*, *Publisher*, *Language*, *Relation*, *Coverage* och *Rights*. Resterande sex Dublin Core-element i vår profil har av oss fått vissa semantiska förändringar. Kortfattat är förändringarna såsom följer:

### *Subject*

Här har vi valt att lägga till tre qualifiers hämtade från IMDI:s semantik.

---

<sup>70</sup> Dekkers 2001.

### *Date*

Vi har gjort definitionen av elementet snävare genom att fastslå att datumet endast ska beskriva tidpunkten för inspelningssessionen, då vi ansåg att det är resursens viktigaste datum. Vi valde att ha ett enda datum för enkelhetens skull och för att underlätta interoperabilitet. Definitionen är sålunda tagen från en qualifier för elementet, nämligen *Date.Created*. Elementet är inte heller upprepningsbart, vilket alla element i Dublin Core vanligen är.

### *Type*

Vi har för detta element lagt till en ordlista för resurstyper som används av IMDI, vid sidan av Dublin Cores resurstyp lista. Elementet är inte upprepningsbart.

### *Description*

#### *Format*

#### *Identifier*

I dessa tre element är den enda förändringen att de inte är upprepningsbara. Den främsta anledningen till att vi gjort denna inskränkning är för att undvika att beskrivningar av olika filer (från samma inspelningssession) förväxlas. Varje fil får beskrivas separat och då finns inte heller någon anledning till att ovanstående element ska vara upprepningsbara. Mer om denna problematik kommer längre fram i analysen, där vi diskuterar elementet *DC:Format*.

Det enda IMDI-elementet i profilen är *Participant*. Till elementet hör i IMDI 13 stycken underelement. I vår profil har vi gjort åtta av dessa underelement till qualifiers. En av dessa qualifiers har vi beslutat skall vara upprepningsbar, nämligen *Language*. Övriga qualifiers är inte upprepningsbara.

## **6.3 Problem rörande semantik**

I detta avsnitt tar vi upp speciella problem som gäller metadatans semantik i skapandet av vår application profile. Vi analyserar problematiken utifrån några element i profilen.

### *DC:Subject*

Detta element är det mest problematiska i vår profil. I mappningsdokumentet (se bilaga 1) uttrycks i kommentaren att Dublin Core har två överlappande element som motsvarar IMDI:s

*Content*, nämligen *DC:Subject* och *DC:Type*. IMDI menar att inget av de båda är heltäckande för de olika kategorier som finns inom IMDI:s *Content*. Vi anser emellertid att *DC:Subject* ligger närmare *IMDI:Content* än *DC:Type*. *Type* syftar till resurstypen medan *Subject* syftar till det intellektuella innehållet.

Problemet med *DC:Subject* är att semantiken är väldigt bred. Dublin Core rekommenderar fem olika kontrollerade vokabulärer för detta element (se tabellen för Dublin Core Qualifiers, i avsnitt 3.1.3). Ingen av dessa vokabulärer kan erbjuda den mycket specifika terminologi som återfinns under *IMDI:Content*. Underelementen i *IMDI:Content* är däremot väldigt ämnesspecifika och använder sig av ordlistor utformade av IMDI. Det rör sig om element och ämnesord som kan ge forskare tydlig information om resursens innehåll och därmed också om det är av intresse för dennes forskning.

Vi har därför valt att till elementet lägga tre qualifiers från IMDI, *Genre*, *Task* och *Modalities*. Alla dessa är specifikt lingvistiska kategorier och tillför elementet en snävare semantik som man inte skulle uppnå i en vanlig Dublin Core-beskrivning. De qualifiers vi lagt till ger en fingervisning om vad vi vill att elementet ska innehålla. Det leder till att elementet får ett helt annat innehåll än om man endast skulle tillåta ämnesord från till exempel LCSH eller fria nyckelord. Om man tittar på den beskrivning av en språkresurs vi gjort i Dublin Core (se 4.1), har vi där använt *DC:Subject* två gånger, först för att beskriva språket, vilket vi ser som en självklarhet att ha med som ett ämnesord när det handlar om språkresurser. Vi har även lagt in termerna *speech* och *pointing gestures*, som är viktig information i den här typen av resurser. Det är dock tveksamt om dessa termer verkligen skulle läggas in i en Dublin Core-beskrivning där man inte har definierat den ordlista där de ingår. Som fria nyckelord är de ganska svårtolkade. Samma resonemang gjorde att vi tog bort två andra liknande ämnesord som vi ursprungligen hade med i Dublin Core-beskrivningen. Dessa var *explanation* och *frog-story*. Termerna är hämtade från IMDI:s öppna ordlistor, som är knutna till de qualifiers vi valt till *DC:Subject* i vår profil. Därför går de alltså bättre att använda i vår application profile än i Dublin Core.

#### *IMDI:Participant*

Det här elementet beskriver de personer som medverkar i en inspelningssession. Ett alternativ till *IMDI:Participant* skulle kunna vara Dublin Cores *Contributor*, som definieras som någon som medverkat till resursens innehåll. Här tycker vi att det finns en liten semantisk skillnad,

där *Contributor* indikerar något mer av aktivitet hos en bidragande, än *Participant*, som ju betyder deltagande. För den typ av resurser vår profil är ämnad att beskriva är *IMDI:Participant* det som passar bäst.

#### *IMDI:Participant.Language*

Detta element listar alla de språk en deltagande behärskar. Det är något som inte kan uttryckas i Dublin Core. *DC:Language* beskriver det språk som används i resursens innehåll, vilket ju inte automatiskt är detsamma som vilket/vilka språk en deltagande kan tala. *DC:Language* motsvaras i IMDI av *Content.Languages*. Det finns dock flera olika aspekter av *Languages* inom IMDI. Förutom vilka språk de deltagande talar, anges också på vilket språk olika beskrivningar inom standarden är skrivna på. Specificeringen av språken är förstås viktig när det handlar om just språkresurser och lingvistisk forskning. Vi har valt att i vår profil ta med två aspekter av språk: först det språk som talas på inspelningen, vilket beskrivs i *DC:Language*. Eftersom *DC:Language* är upprepningsbart ser vi inga problem ifall flera språk skulle användas i inspelningen. Den andra aspekten är de språk de deltagande behärskar, vilket beskrivs i *IMDI:Participant.Language*.

### **6.4 Problem rörande struktur**

Ett annat problem som uppstår vid beskrivningen av de deltagande är av strukturell art. Eftersom en språkresurs kan innehålla flera deltagande, måste de beskrivas en i taget. Detta för att man ska kunna koppla rätt information, till exempel *Age* och *Sex*, till rätt person. Det är ett problem som också kan lösas genom ett tydligt gränssnitt, både för den som arbetar med att lägga in metadata och för användaren.

Samma svårigheter kan man se gällande *DC:Format*, som beskriver filformatet. Det kan som tidigare diskuterats finnas flera filer inom en och samma språkresurs. Vi har i vår profil valt att göra elementet *DC:Format* icke upprepningsbart för att den information som hör till en specifik fil skall beskrivas separat och inte riskera att förväxlas med information om andra filer. Även här skulle kanske problemet kunna lösas med ett bra gränssnitt, så att flera filer skulle kunna rymmas i samma beskrivning. Gränssnittsproblematiken är dock något som ligger utanför denna uppsats.

*Participant* är i IMDI ett element som ingår i en hierarkisk struktur. Det överordnade elementet är *Participants* (i plural). Se exempel 1 nedan.

### Exempel 1.

#### Nivå 1 *Participants*

Nivå 2 *Description (sub-schema)*

Nivå 3 *Language*: RFC1766:x-sil-eng

*Text*: J. was the only participant in the session...

Nivå 2 *Participant*

Nivå 3 *Type*: Consultant

*Full name*: John Doe

*Role*: Son

*Language (sub-schema)*

Nivå 4 *Name*: Hausa

*Id*: RFC1766:x-sil-hua

*Description (sub-schema)*

Nivå 5 *Text*: Hausa is a Chadic language...

Nivå 3 *Ethnic group*: Goemai

*Age*: 24

*Sex*: Male

*Education*: Secondary school

*Description (sub-schema)*

Nivå 4 *Language*: RFC1766:x-sil-eng

*Text*: J. is one of my main collaborators in Kwande...

Om vi till exempel ska beskriva en deltagandes namn så ser det i IMDI ut på följande sätt:

### Exempel 2.

*IMDI:Participants.Participant.Full name: John Doe*

↑            ↑            ↑

nivå 1        nivå 2        nivå 3

Som framgår av exempel 1 ovan, finns det två element på nivå 2, nämligen *Description* och *Participant*. Vi har valt att i vår profil endast ta med *Participant* från denna nivå. Det innebär att *Participants* på nivå 1 inte längre fyller någon funktion. Vi bortser därför från nivå 1 i vår application profile. I vår profil ser en beskrivning av en deltagandes namn sålunda ut såhär:

*IMDI:Participant.Full name: John Doe*

↑            ↑

nivå 1        nivå 2

(nivå 2 i IMDI) (nivå 3 i IMDI)

Vi har alltså förändrat strukturen i detta fall, för att få ett enklare och mer lätthanterligt element till profilen. Ingen information går förlorad här, utan elementet har endast anpassats för den nya struktur den ska ingå i.

## 6.5 Hur strukturproblem kan påverka semantiken

När man förändrar strukturen på det sätt som beskrivs ovan, måste man också vara uppmärksam på semantiken. Det kan hända att ett underordnat elements innebörd är beroende av ett överordnat element, vilket kan skapa felaktigheter om man tar elementet ur dess ursprungliga hierarkiska struktur. Till exempel skulle det underordnade elementet (qualifier i vår profil) *Full name* inte kunna stå som ett eget element, eftersom det är definierat som namnet på en deltagande. *Participant* behövs alltså för att klargöra innebörden i *Full name*. Med andra ord är det nödvändigt att analysera strukturens/hierarkins påverkan på semantiken

i de enskilda elementen, när man överväger strukturella förändringar av det slag vi illustrerat i exemplet ovan.

I detta sammanhang vill vi framhålla vikten av att man i en application profile är noga med varifrån man hämtar sin definition av ett element eller qualifier. I vår profil har vi qualificern *Description* under elementet *Participant*. I IMDI finns detta underelement *Description* på nivå 3 och ska inte förväxlas med elementet *Description* på nivå 2, såsom nämnts i diskussionen ovan (se exempel 1). I IMDI kan *Description* inte användas fristående, utan åtföljs alltid av underelement från ett Sub-schema. För denna qualifier använder vi oss i profilen av en definition som härrör från ett av dessa underelement i sub-schemat, *Description.Text*. Definitionen innebär att innehållet uttrycks i fritext. I IMDI skulle en sådan fritextbeskrivning se ut på detta vis:

*IMDI:Participants.Participant.Description.Text: J. is one of my main  
collaborators in Kwande...*

I vår application profile är samma beskrivning alltså enklare:

*IMDI:Participant.Description: J. is one of my main collaborators in Kwande...*

Vi behöver inte använda underelementet *Text* till *Description* eftersom vi i definitionen av elementet redan har postulerat att innehållet ska uttryckas i fritext. I IMDI är fritext bara en av flera beskrivningskategorier för *Description*.

## **6.6 Interoperabilitet kontra ämnesspecificitet**

När man skapar en application profile för att uppnå kompatibilitet mellan ämnesområden är interoperabiliteten en grundläggande aspekt. Dublin Core är en standard med uttalat syfte att främja interoperabilitet mellan olika domäner. Sålunda har vi i profilen arbetat efter principen att använda Dublin Core-element i så stor utsträckning som möjligt. Profilen har 14 element, varav 13 är hämtade från Dublin Core, och därmed menar vi att vår profil i hög grad understödjer interoperabilitet.

När det gäller ämnesbeskrivningen i profilen, stod vi inför valet mellan *DC:Subject* och *IMDI:Content*. Som vi tidigare har påpekat fann vi att *DC:Subject* var för brett i semantiken för att kunna ge en tillfredsställande beskrivning av språkresurser för vår målgrupp. Alternativet var att använda *IMDI:Content* för att uppnå tillräcklig ämnesspecificitet, men det skulle påverka interoperabiliteten negativt. Lösningen blev att ta med *DC:Subject* som huvudelement och använda underelement från *IMDI:Content* som qualifiers, vilket vi menar kan vara ett sätt att balansera de båda syftena med profilen, interoperabiliteten och ämnesspecificiteten.

Det andra syftet med en profil med vår uttalade målgrupp är alltså att den ska ge en tillräckligt ämnesspecifik beskrivning för att tillfredsställa forskare och andra yrkesanvändare. Svårigheten med att definiera vad ”tillräckligt ämnesspecifik beskrivning” är kommer vi att diskutera närmare nedan. För att tillgodose behoven för profilens primära målgrupp, krävs att viss information i resursen beskrivs tämligen ingående. I fallet med *DC:Subject* bedömde vi att behovet av interoperabilitet vägde tyngre än behovet av ämnesdjup i beskrivningen. Vi uppnådde ett visst mått av ämnesspecificitet genom att vi definierade qualifiers från *IMDI:Content*.

I arbetet med en profil ingår att man måste prioritera mellan beskrivningselement och analysera vilka element som beskriver den mest nödvändiga informationen i resursen. Vi har som enda *IMDI*-element till profilen valt *Participant*. Elementet beskriver de deltagande i inspelningssessionen utifrån till exempel ålder, kön, språkkunskaper, utbildning, etnicitet och så vidare. Detta är viktig information för att kunna återfinna den typ av forskningsmaterial som en viss forskare är intresserad av. Vi ansåg att mycket av denna information måste vara med i vår application profile, men att det då inte var möjligt att använda Dublin Core med ett tillfredsställande resultat. En möjlighet i Dublin Core hade varit elementet *DC:Contributor*, men som tagits upp ovan valde vi *IMDI:Participant* av semantiska skäl. De åtta underelement vi valde att ha med som qualifiers i profilen bedömde vi helt enkelt som de mest nödvändiga.

Vad gäller *IMDI:Participant* är fallet något annorlunda än med de qualifiers vi lade till elementet *DC:Subject*. Terminologin i underelementen till *Participant* är inte av ämnesspecifik natur. Men däremot kan man kalla informationen i *IMDI:Participant* ämnesspecifik då den är nödvändig för sökning för vår målgrupp, i synnerhet forskarna.



När det gäller ämnesspecificitet kontra interoperabilitet är det fundamentalt att man definierar det vi i vårt syfte för profilen kallar ”tillräckligt ämnesspecifik beskrivning”. Här finns förstås inga objektiva kriterier. Vad som är tillräckligt är något som löpande måste diskuteras, utvärderas och omprövas. Det vi har bedömt som tillräcklig information här har ingen förankring i någon användarundersökning. Vi har helt utgått från våra egna bedömningar. I en verklig situation är det självklart nödvändigt att göra användarundersökningar, samt att ha fortgående diskussioner med ämnesexperter / forskare, vilka är de som i denna typ av söktjänst ska producera metadatan. Utifrån dessa diskussioner och undersökningar måste man sedan göra en avvägning för att även uppfylla profilens definierade mål angående interoperabilitet.

## 7. Slutsatser

En application profile eller ett metadataschema är aldrig ”färdigt”. Ett perfekt schema som löser alla problem nu och för alltid existerar naturligtvis inte. Webbens dynamiska karaktär innebär att man hela tiden måste finna nya lösningar för att kunna strukturera, organisera och beskriva ständigt föränderliga dokument- och resurstyper.

När man skapar en application profile finns inga generella lösningar. Allt är beroende av projektets kontext, den tänkta målgruppen och dess behov, syftet med profilen, vilken typ av söktjänst eller databas metadatan ska användas i (och här ingår sådant som gränssnittets utformning, sök- respektive browsningsmöjligheter, tekniska lösningar m.m.). Som Lynch och Gilliland-Swetland framhåller (se 2.2) är det meningslöst att tala om metadata utan att ta dess kontext, användningsområde och funktion i beaktande. I själva verket är application profiles konsekvensen av detta praktiska tänkande, det vill säga de är per definition beroende av kontext och användningsområde. Den kontext vår profil ingår i är ECHO-projektet och dess syften (se 2.5.1), det vill säga att integrera resurser inom olika ämnesområden. Varje ämnesområde ska kunna beskriva sina resurser på ett för ämnet relevant sätt men ändå kunna fungera interoperabelt med övriga områden. Application profiles är en möjlighet att få ämnesspecifik metadata interoperabel i ett större sammanhang. Detta är något som Duval, Hodgins, Sutton & Weibel lyfter fram som en av funktionerna i en application profile (se 2.4.2).

I vår första frågeställning frågar vi oss vilken typ av beskrivningselement i en application profile som kan hämtas från en generell metadastandard. Här har vi kommit fram till att de element som kan hämtas från en generell standard är de som beskriver de delar av resursen som inte kräver en ämnesspecifik semantik eller en utvecklad struktur. Detta kan illustreras med alla de Dublin Core-element som ingår i vår profil, möjligen med undantag för *DC:Subject*, som har fått ämnesspecifika qualifiers. Dublin Core-elementen beskriver aspekter av resurserna som inte är utmärkande för ämnet språk, utan skulle kunna användas även inom andra ämnesområden. För att uppnå syftet med interoperabilitet i en application profile försöker man så långt det går använda generella beskrivningselement. Detta har vi haft som grund för vårt arbete och vi vill hävda att det är ett generellt användbart tillvägagångssätt. Vi baserar detta även på andra application profiles som vi studerat under arbetets gång. Oftast

består profilerna till största delen av Dublin Core-element med endast några få element från andra standarder eller egna namespaces. Det finns förstås undantag, eftersom en profil skapas utifrån projektets särskilda egenskaper och behov.

Vad gäller vår andra frågeställning, vilka beskrivningselement som måste hämtas från en ämnesspecifik standard, har vi funnit att de element som behövs är de som erbjuder en för tillämpningen nödvändig ämnesspecifik semantik eller utvecklad struktur. Förutsättningen för att komma fram till ett svar på en dylik fråga är att man har definierat de ämnesspecifika behoven och vad en ”tillräckligt ämnesspecifik” beskrivning är. Behoven kan till exempel vara att man måste göra en snäv definition av innebörden i ett element för att kunna uttrycka nödvändig ämnesspecifik terminologi. Andra behov kan vara att kunna lyfta fram en för målgruppen viktig aspekt av resursen som en generell standard inte kan återge på ett tillfredsställande sätt. Det kan gälla både semantik och struktur. Ibland kan det krävas en hierarkiskt djupgående beskrivning för att strukturera information i resursen som tillhör samma element. Dublin Cores qualifiers är ett försök att ge standarden struktur, men när det gäller domänspecifika resurser behövs ofta element från en domänspecifik standard.

Vår tredje frågeställning tar upp hur stor betydelse skillnader i struktur och semantik mellan metadatastandarderna har i arbetet med en application profile. Som man kan se är skillnaden mellan standarderna i dessa aspekter mycket viktiga att ta hänsyn till när man utvecklar en ny application profile, en erfarenhet vi har fått i arbetet med vår application profile. Vi menar att de semantiska problemen sannolikt uppstår i de allra flesta arbeten med application profiles. Semantiken och dess problem är grundläggande i application profiles, eftersom elementen väljs utifrån sin innebörd (semantik) för att passa profilens syfte. Detta gör att semantiska problem är ofrånkomliga. Strukturella problem bedömer vi inte vara lika vanliga, eftersom dessa uppstår beroende på resursernas beskaffenhet och den specifika målgruppen, något som varierar från fall till fall. Våra problem har under arbetet med profilen i hög grad varit strukturella, men vi anser att vi inte kan dra några generella slutsatser utifrån detta, då våra problem stammar från beskaffenheten hos just de resurser vi valt att beskriva, det vill säga språkresurser. Resurser kan ha olika egenskaper beroende på ämnesområde, vilket innebär att problemen kan se annorlunda ut när man skapar en application profile inom ett annat område. Det kan även vara en skillnad beroende på vilken målgruppen är. Eftersom vår målgrupp i första hand är forskare och andra yrkesanvändare krävs en struktur som möjliggör för en mer ingående beskrivning. Om målgruppen är en annan, som inte kräver lika detaljerad

information, behövs inte en struktur av samma slag. I vår profil kan man se exempel på detta i beskrivningen av de deltagande i inspelningarna, vilket görs i elementet *IMDI:Participant*.

Här kan vi alltså se att det till viss del går att identifiera generella problem vid utformandet av en application profile. I vår problemställning frågar vi oss även om det går att skapa en application profile som kan fungera som vägledning för utveckling av ämneskompatibla beskrivningsscheman. För en profil med helt annat syfte och annan målgrupp kan vi inte se att vår profil är vägledande i någon större bemärkelse. Förutsatt att målgrupp och syfte är mycket närliggande, vilket ju är tanken i ECHO-projektet, så menar vi att vår profil i vissa avseenden kan vara till hjälp. Främst gäller detta profilens omfång och fördelning av ämnesspecifika kontra generella beskrivningselement. Möjligtvis kan vår profil även ge fingervisningar om vissa problemlösningar, som till exempel att kombinera ämnesspecificitet och interoperabilitet inom samma element, det vill säga att foga till ämnesspecifika qualifiers till ett generellt element. På en mer detaljerad nivå är det dock tveksamt om våra lösningar kan vara vägledande. Varje profil måste utgå från ämnets och resursernas egenskaper. Dessa kan se väldigt olika ut från område till område och påverkar utformandet av profilen även om det övergripande syftet och målgruppen är identiska.

## **Förkortningar**

**AACR2** - Anglo American Cataloging Rules, 2nd edition

**CES** - Corpus Encoding Standard

**DCMES** - Dublin Core Metadata Element Set

**DCMI** - Dublin Core Metadata Initiative

**DC-LAP** - DC Library Application Profile

**DESIRE** - Development of a European Service for information on Research and Education

**DLO** - Document-Like Object

**EAD** - Encoded Archival Description

**EAGLES** – Expert Advisory Group on Language Engineering Standards

**ECHO** - European Cultural Heritage Online

**FGDC** - Federal Geographic Data Committee

**HTML** – Hypertext Markup Language

**IEEE LOM** - Institute for Electric and Electronic Engineering, Learning Object Metadata

**IMDI** - ISLE Metadata Initiative

**ISO** - International Organization for Standardization

**IST**- Information Society Technologies Programme

**ISLE** - International Standards for Language Engineering

**MARC** – Machine Readable Cataloguing

**MIME** - Multipurpose Internet Mail Extensions

**MPEG** - Moving Pictures Expert Group

**OCLC** – Online Computer Library Center

**RDF** - Resource Description Framework

**RFC** – Request For Comments

**SGML** – Standard Generalized Markup Language

**TEI** – Text Encoding Initiative

**UKOLN** - UK Office for Library and Information Networking

**URI** - Uniform Resource Identifier

**URL** - Uniform Resource Locator

**URN** - Uniform Resource Name

**WAV** – ljudfil

**W3C** – World Wide Web Consortium

**XML** - Extensible Markup Language

## Käll- och litteraturförteckning

Alla URL:er kontrollerade 2003-02-23.

Benito, Miguel, 2001. *Kunskapsorganisation: en introduktion till katalogisering, klassifikation och indexering*. Borås: Tarancos bokförlag.

Björkhem, Miriam & Lindholm, Jessica, 2000. *Metadata för det digitala biblioteket: objektbeskrivning av elektroniska resurser*. Magisteruppsats i Biblioteks- och informationsvetenskap, Lunds universitet.

URL: <http://www.kult.lu.se/bivil/publikationer/fulltext00/2000-7.pdf>

Borgman, Christine L., 2000. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. Cambridge, Mass. & London: The MIT Press.

Dekkers, Makx, 2001. "Application Profiles, or how to Mix and Match Metadata Schemas", *Cultivate Interactive*, Issue 3, January 2001.

URL: <http://www.cultivate-int.org/issue3/schemas/>

Dempsey, Lorcan & Heery, Rachel, 1998. "Metadata: a current view of practice and issues", *Journal of Documentation* 54 (2), s. 145-172.

Duval, E., Hodgins, W., Sutton, S., Weibel, S., 2002. "Metadata Principles and Practicalities", *D-Lib Magazine*, April 2002, Volume 8, Number 4.

URL: <http://www.dlib.org/dlib/april02/weibel/o4weibel.html>

EU-parlamentet, Utskottet för kultur, ungdomsfrågor, utbildning, medier och idrott. Dokument PE 286.688/1-76, 15 november 2000.

URL: <http://www.europarl.eu.int/meetdocs/committees/cult/20001204/425655sv.doc>.

*European Cultural Heritage Online (ECHO)*. 30 september 2002. Projektbeskrivning. (Publicerat dokument.)

Fagerlind, Marita & Gisselqvist, Gunilla, 1999. *Metadata enligt Dublin Core*. Magisteruppsats i Biblioteks- och informationsvetenskap, Lunds universitet.

URL: <http://www.kult.lu.se/bivil/publikationer/fulltext99/1999-11.pdf>

Gill, Tony, 2000. "Metadata and the World Wide Web 2000", *Introduction to Metadata: Pathways to Digital Information*.

URL:

[http://www.getty.edu/research/institute/standards/intrometadata/2\\_articles/gill/content.html](http://www.getty.edu/research/institute/standards/intrometadata/2_articles/gill/content.html)

Gilliland-Swetland, Anne J., 2000. "Setting the Stage", *Introduction to Metadata: Pathways to Digital Information*.

URL: [http://www.getty.edu/research/institute/standards/intrometadata/2\\_articles/swetland/content.html](http://www.getty.edu/research/institute/standards/intrometadata/2_articles/swetland/content.html)

Gorman, Michael, 2002. "Metadata: Hype and Glory", i Jones, Ahronheim & Crawford (ed.).

Guenther, Rebecca, 2002. "MARC 21 as a Metadata Standard: A Practical and Strategic Look at Current Practices and Future Opportunities", i Jones, Ahronheim & Crawford (ed.).

Hedberg, Sten, [odat.]. "Metadata - kataloginformation på Internet".  
URL: <http://www.kb.se/Nvb/Katalog/kat1.htm>

Heery, Rachel & Patel, Manjula, 2000. "Application profiles: mixing and matching metadata schemas", *Ariadne*, Issue 25, September 2000.  
URL: <http://www.ariadne.ac.uk/issue25/app-profiles/intro.html>

Hudgins, J., Agnew, G., Brown, E., 1999. *Getting Mileage out of Metadata: Applications for the Library*. Chicago: American Library Association.

Hunter, Jane, 2002. "An Application Profile which combines Dublin Core and MPEG7 Metadata Terms for Simple Video Description", 2002-02-12.  
URL: [http://metadata.net/harmony/video\\_appln\\_profile.html](http://metadata.net/harmony/video_appln_profile.html)

*IMDI (ISLE Metadata Initiative). Part 1A. Metadata Elements for Session Descriptions. Version 2.5.* June 2001.  
URL: [http://www.mpi.nl/ISLE/documents/docs\\_frame.html](http://www.mpi.nl/ISLE/documents/docs_frame.html)

*IMDI (ISLE Metadata Initiative). Part 1. Metadata Elements for Session Descriptions. Draft Proposal Version 3.0.* November 2002. (Opublicerat dokument.)

"Introduction to the EAGLES initiative".  
URL: <http://www.ilc.pi.cnr.it/EAGLES96/edintro/node6.html>

Jones, Wayne, 2002. "Preface: Meting Out Data", i Jones, Ahronheim & Crawford (ed.).

Jones, W., Ahronheim, J.R. & Crawford, J. (ed.), 2002. *Cataloging the Web: Metadata, AACR, and MARC 21*. Lanham, Maryland & London: The Scarecrow Press.

Kronman, Ulf & Parnefjord, John, 2001. "Resource Description Framework – metadata för framtidens Internet", *Tidskrift för dokumentation*, 2001:1, s. 15-25.

Lagoze, Carl, 2001. "Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description?", *D-Lib Magazine*, January 2001, Volume 7, Number 1.  
URL: <http://www.dlib.org/dlib/january01/lagoze/01lagoze.html>

Lazinger, Susan S., 2001. *Digital Preservation and Metadata: History, Theory, Practice*. Englewood, CO: Libraries Unlimited.

Lynch, Clifford A., 2002. "Future Developments in Metadata and Their Role in Access to Networked Information", i Jones, Ahronheim & Crawford (ed.).

Medeiros, Norm, 2000. "XML and the Resource Description Framework", *Online*, Vol. 24, Issue 5, Sep/Oct 2000, s. 37-40.



Neuroth, Heike & Koch, Traugott, 2001. "Metadata Mapping and Application Profiles. Approaches to providing the Cross-searching of Heterogeneous Resources in the EU Project Renardus". URL: <http://www.nii.ac.jp/dc2001/proceedings/product/paper-21.pdf>

Rowley, Jennifer & Farrow, John, 2000. *Organizing Knowledge: An introduction to Managing Access to Information*. Aldershot: Gower Publishing Ltd.

Schwartz, Candy, 2001. *Sorting Out the Web: Approaches to Subject Access*. Westport, Conn. & London: Ablex Publishing.

Strunck, Kirsten, Lund, Haakon & Thorlund Jepsen, Erik, 1998. *Katalogiseringsteori*. København: Danmarks Biblioteksskole. URL: <http://ix.db.dk/epub/doksam/katalogiseringsteori.html>

Sjölund, Stefan & Wismén, Elon, 1999. *Dublin Core: ett schema för metadata*. Magisteruppsats i Biblioteks- och informationsvetenskap, Högskolan i Borås.

Skolverket, dokument för *Projektet Kultur för lust och lärande*. URL: [http://www.skolverket.se/teknikprogrammet/texteromteknik/pub\\_reform\\_kulturarv\\_ligger\\_i\\_tiden.shtml](http://www.skolverket.se/teknikprogrammet/texteromteknik/pub_reform_kulturarv_ligger_i_tiden.shtml)

Wittenburg, Peter & Broeder, Daan, 2002. *Metadata Overview and the Semantic Web*. URL: <http://www.mpi.nl/lrec/papers/lrec-pap-04-MD-overview-daan3.pdf>

Wittenburg, P., Broeder, D. & Sloman, B., 2000. *Meta-Description for Language Resources. A Proposal for a Meta Description Standard for Language Resources. EAGLES/ISLE White Paper*. URL: [http://www.mpi.nl/ISLE/documents/papers/white\\_paper\\_11.pdf](http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf)

Världsbanken, dokument: *Culture and Development Action Network. Working Group Meeting Brief, January 26-27, 1998*. URL: <http://wbln0018.worldbank.org/essd/essd.nsf/9b1cfc683a76b671852567cb0076a25e/06521784aeba7c21852567ed004d360d?OpenDocument>

Younger, Jennifer A., 2002. "Metadata and Libraries: What's It All About?", i Jones, Ahronheim & Crawford (ed.).

### **Hemsidor:**

Dublin Core Metadata Initiative:  
<http://dublincore.org>

ISLE Metadata Initiative, IMDI:  
<http://www.mpi.nl/ISLE>

DESIRE-projektet:  
<http://desire.ukoln.ac.uk>

SCHEMAS-projektet:  
<http://www.schemas-forum.org>

World Wide Web Consortium:

<http://www.w3.org>

W3C-Corpora (The World Wide Web Access to Corpora Project):

[http://www.essex.ac.uk/linguistics/clmt/w3c/corpus\\_ling/content/](http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/)

Netcraft:

<http://www.netcraft.com>

**Bilaga 1.**  
**Mapping mellan IMDI Version 2.2 och Dublin Core**

**1. Från IMDI till Dublin Core**

<b>IMDI Element</b>	<b>Comment</b>	<b>DC Element</b>	<b>Comment</b>
Meta-transcript.type	This element denotes that there are different MD sets for different data types such as corpora, lexica etc.	Type	This element best correlates with the intentions of IMDI. DC:Type is aimed to describe "general categories and aggregation levels" for content and the proposed controlled vocabulary indicates a high overlap.
Session.name	Is a short name of a session	Title.short	Could be a qualifier to the DC:Title element, since it refines Title
Session.title	Title of the session	Title	Almost perfect semantic match
Session.description	Description of a session	Title.description	Could be a qualifier to the DC:Title element, since it adds information to the Title element
Session.date	Date of the recording the session is based on	Date.main	Perfect match with the DC:Date element; since there are several dates involved there should be a qualification
Session.continent	Continent where the recording is made	Coverage.continent	Perfect match with DC:Coverage, could be qualified to denote the level of detail
Session.country	Country where the recording is made	Coverage.country	Same as above, but narrowed down
Session.region	Region where the recording is made	Coverage.region	Same as above, but narrowed down
Session.address	Sometimes it may be required to specify the	Coverage.address	Same as above, but narrowed down

	address		
Project.name	Short name or abbreviation of the project	-	There is no real equivalent in the DC set to describe the project. One could think of qualifying elements such as DC:Title, DC:Description, DC:Identifier; but it would mean to extend their semantics
Project.title	Title of the project	-	
Project.description	A description associated with the project	-	
Project.id	A unique identifier for the project	-	
Project.contact	A relevant address associated with the project	-	
Collector.name	The name of the person who was responsible for creating the resources in terms of recording and collecting. It is not meant to be person being interviewed etc.	Creator.name	Almost perfect match; perhaps qualified to distinguish between various IMDI entries
Collector.address	Address of the collector	Creator.address	Specification of an attribute of the creator; could also be solved by using for example RDF and adding other type of information taken for example from the Index element set
Collector.LINK	Web page of the collector	Creator.link	Same as above
Content.description	Free-text description of the content of the recording	Description	Almost perfect match with DC element
Content.type	Elements to refine the description of the content in terms of keywords; typically controlled vocabularies are associated with these elements	Subject.type	Almost perfect match with DC;Subject element; to make the distinctions as was required by IMDI members, qualifiers could be used
Content.genre		Subject.genre <sup>9</sup>	
Content.register		Subject.register	
Content.channel		Subject.channel	
Content.event		Subject.event	
Content.modalities		Subject.modalities	
Content.main.language <sup>10</sup>	Element to characterize the language the recording is about, the language being the subject of study	Subject.language	Same as above – just another qualifier; to make a distinction between main and other languages one could add another qualifier as a refinement
Content.language	Other languages		
Content.language.description	A free-text description associated with the languages used in the recording	-	No real equivalent in DC; could perhaps be achieved with RDF
Content.keys	List of attribute-value pairs to further describe the content	-	This concept of flexibility is not really supported by DC
Participants.description	A free-text description associated with the participants occurring in the recording	-	No real equivalent in DC; could perhaps be achieved with RDF
Participants.researcher.name	Name of the person used in the annotations	Creator.short-name; Contributor. Researcher. short-name	In DC there are two options dependent on the function of the person involved; whether this form of sub-qualification is intended by

			DC can be doubted
Participants. researcher. full-name	Full-name of the person acting as interviewer which is often the same as the collector	Creator; Contributor. <i>researcher</i>	In DC there are two options dependent on the function of the person involved
Participants. researcher. code	All elements further specify the participating researcher. IMDI people say that it is important to have these specifications often for quick inspection only.	Contributor. <i>researcher.code</i>	Could all be qualifiers to either the DC:Creator or the DC:Contributor element
Participants. researcher. role		Contributor. <i>researcher.role</i>	
Participants. researcher. language		Contributor. <i>researcher. language</i>	
Participants. researcher. age		Contributor. <i>researcher.age</i>	
Participants. researcher. ethnic-Group		Contributor. <i>researcher. ethnic-group</i>	
Participants. researcher. sex		Contributor. <i>researcher.sex</i>	
Participants. researcher. link		Contributor. <i>researcher.link</i>	
Participants. consultant. description		A free-text description associated with the consultants occurring in the recording	
Participants. consultant.name	Name of the consultant used in the annotations	Creator. <i>short-name</i> ; Contributor. <i>consultant. short-name</i>	whether this form of sub-qualification is intended by DC can be doubted
Participants. consultant. full-name	Full-name of the person acting as consultant which is often the same as the collector	Creator; Contributor. <i>consultant</i>	See above
Participants. consultant.code	All elements further specify the participating researcher. IMDI people say that it is important to have these specifications often for quick inspection only.	Contributor. <i>consultant.code</i>	Could all be qualifiers to the DC:Contributor element
Participants. consultant.role		Contributor. <i>consultant.role</i>	
Participants. consultant. first-language		Contributor. <i>consultant. First-language</i>	
Participants. consultant. language		Contributor. <i>consultant. language</i>	
Participants. consultant. ethnic-group		Contributor. <i>consultant. ethnic-group</i>	
Participants. consultant.sex		Contributor. <i>consultant.sex</i>	
Participants. consultant.age		Contributor. <i>consultant.age</i>	

Participants. consultant. education		Contributor. <i>consultant. education</i>	
Participants. consultant. anonymous		Contributor. <i>consultant. anonymous</i>	
Participants. consultant.keys	Mechanism to extend the description of the participant by attribute-value pairs	-	This concept of flexibility is not really supported by DC
Participants. contributory.name	Elements to describe a person which acts in the recording as a relevant contributory, but who is not the person being recorded	Contributor. <i>contributory. name</i>	Could all be qualifiers to the DC:Contributor element
Participants. contributory.code		Contributor. <i>contributory. code</i>	
Participants. contributory.role		Contributor. <i>contributory. role</i>	
Media-File.access	Structured element to describe the access rights of the media file	Rights; Publisher	Two DC elements are directly appropriate to elements of the IMDI structure; others could be mapped by using RDF mechanisms
Media-File.link	Unique identifier which normally is an URL	Identifier	The DC:Identifier element is used to contain a unique reference; IMDI sets contain a number of such unique references to related resources – this has to be solved by structural mechanisms
Media-File.size	the size of the session's media file	-	No direct equivalent; could be a qualifier of DC:Format although the semantics don't really fit
Media-File.type	the type of the session's media file such as Photo, Audio, Video	Type	The DC:Type element could be used here again; to denote the difference with the above usage it has to be embedded into some structure; also the controlled vocabulary is different to the one above which creates problems
Media-File.format	the format of the session's media file such as mpg, jpg, wav, ...	Format	Almost perfect match with DC element; however, the IMDI set contains several format specs i.e. structure mechanisms have to be used
Media-File.quality	the quality of the session's media file in global terms taken from a controlled vocabulary	-	No real equivalent in DC
Media-File.position	the start and stop time references of the session's media file with respect to its original material	-	No real equivalent in DC
Media-File.tool	Denotes the tools which could be started to operate on this file; perhaps not necessary if the format	-	No real equivalent in DC; Perhaps not necessary due to MIME types contained in DC:Format

	contains MIME types		
AnnotationU. access	Structured element to describe the access rights of the annotations	Rights; Publisher	Two DC elements are directly appropriate to elements of the IMDI access structure; others could be mapped by using RDF mechanisms
AnnotationU. language	The languages the annotations are written in; Can be several	Language	Almost perfect match with the DC element
AnnotationU. annotator	The person who did the annotations	Contributor. <i>annotator</i>	Could be solved by qualifying the contributor
AnnotationU. link	Unique identifier which normally is an URL	Identifier	The DC:Identifier element is used to contain a unique reference; IMDI sets contain a number of such unique references to related resources – this has to be solved by structural mechanisms
AnnotationU. size	the size of the session's annotation file	-	No direct equivalent; could be a qualifier of DC:Format although the semantics don't really fit
AnnotationU. type	the type of the session's annotation file such as orthographic, phonetic, morphologic, syntactic, translation, ...	Type	The DC:Type element could be used here again; to denote the difference with the above usage it has to be embedded into some structure; also the controlled vocabulary is different to the others which creates problems
AnnotationU. format	the format of the session's annotation file such as CHAT, Shoebox, AIF, ...	Format	Almost perfect match with DC element; however, the IMDI set contains several format specs i.e. structure mechanisms have to be used
AnnotationU. encoding	Element to describe which font or encoding scheme was applied	-	No direct equivalent in DC
AnnotationU. date	Date when a certain annotation tier was created	Date	DC offers the DC:Date field; since there are many dates this has to be qualified and perhaps to be embedded in structures
AnnotationU. lexicon	Reference to a lexicon file which can be used in parallel	-	No direct equivalent in DC
AnnotationU. anonymous	Just a switch to indicate whether real names have to be replaced by pseudo names	-	No direct equivalent in DC
AnnotationU. tool	Denotes the tools which could be started to operate on this file; perhaps not necessary if the format contains MIME types	-	No real equivalent in DC; Perhaps not necessary due to MIME types contained in DC:Format
MediaCarrier. access	Structured element to describe the access rights of the original media carriers	-	No real equivalent in DC;
MediaCarrier. Storageformat	Element characterizing the media format such as DAT, DV, VHS, Hi-8, ...	Format	The DC:Format element is an almost perfect match; it has to be embedded in a structure to

			differentiate with the other format usages
MediaCarrier. quality	Element characterizing the quality of the signals on the original material taken from a controlled vocabulary	-	No real equivalent in DC;
MediaCarrier. ID	Reference to a specific tape with a unique label	-	Not applicable since there is in general no mechanism to establish uniqueness
References	Descriptive element to point to refer to all sort of related documents	Relation	DC has the DC:Relation element to enter such references



## 2. Från Dublin Core till IMDI

DC Element	Comment	IMDI Element	Comment
Title <sup>2</sup>	Short name given to the resource	Session.title	unproblematic mapping
Creator	Entity primarily responsible for making the content of the resource (person, organization, service, ...)	Session.collector	In fact unproblematic mapping, the IMDI term is more neutral IMDI offers to enter a structured description
Subject <sup>3</sup>	Topic of the content of the resource; subject typically expressed as a sequence of keywords taken for example from a controlled vocabulary	Content.description Content.genre Content.register Content.channel Content.event Content.modality Content.language	DC has two overlapping elements to specify the content: Subject and Type; while Subject describes the content, Type is used to characterize the type or intellectual form the creator used to convey the content; yet we cannot exactly make mappings between DC and IMDI categories

Relation <sup>7</sup>	Reference to a related resource		Not used in IMDI since IMDI includes related resources which are partly encoded in one resource, but also could be part of several resources; the session concept could be broken up partly to use the DC:Relation element
Coverage <sup>8</sup>	Extent or scope of the resource; typically to include spatial location, temporal period or jurisdiction; best to be taken from controlled vocabularies	Session.continent Session.country Session.region Session.address	IMDI includes elements to specify a spatial location; they could be seen as qualifiers; there is no equivalence for the temporal period in IMDI
Rights	Information about rights held in and over the resource	Media-file.access; Annotation-unit.access; Media-carrier.access	The different resources described by IMDI sets have their own access policy statements; could be expressed in DC by structures

Description	Account of the content of the resource which could be an abstract, table of content, some reference to a free-text description etc.	Session.description Content.description	IMDI has several descriptions on various levels to allow people to express annotations they want to make with as goal to make the MD descriptions self-standing; mostly they have to be qualifiers of the corresponding elements
Publisher	Entity responsible for making the resource available (person, organization, service, ...)	Access.owner Access.publisher	IMDI provides the Access structure for this at several points; it covers availability and contact information
Contributor	Entity responsible for making contributions to the content of the resource (person, organization, service, ...)	Session.participants ....researcher ....consultant ....contributory	IMDI separates several types of contributors which are important for language resources; can be seen as qualifiers refining the semantics
Date <sup>4</sup>	Date associated with an event in the resources life-cycle	???.date	IMDI has several dates since language resources in general have for example many components which are created at different times; could be seen as qualifiers
Type <sup>5</sup>	Nature or genre of the content of the resource ideally defined by a controlled vocabulary	s. above under Subject	When looking careful to the comments given to this element, no IMDI element fits well; for similar elements see above under Subject
Format <sup>6</sup>	Physical or digital manifestation of the resource; may include things as media-type or information to indicate which software is necessary to operate on the resource (MIME type concept)	Media-file.format; Annotation-unit.format; Media-carrier.format	IMDI separates into three categories of format: - media files and annotation unit files, - media-carrier format which refers to the original audio/video carrier
Identifier	Unambiguous reference to the resource (URI, ISBN, ...)	Media-file.URL; Annotation-Unit.URL	IMDI metadata descriptions refer to a number of resources which form a bundle; this bundling can be expressed by using the relation mechanism of DC
Source	Reference to a resource from which the present resource is derived		Widely overlapping with Relation since "derived from" is a special type of relation; no correlate in IMDI although the original media-carrier could be seen as a source or high-level linguistic annotations could be seen as derived from an orthographic transcriptions
Language	Language the content of the resource is written in; can be several; specified by ISO or RFC codes	Annotation-unit.language	Meant is the language the resource is written in; in our case it is the language used for the transcriptions and annotations