Metadata för objekt- och tjänstebaserad säkerhet

HENRIK KARLZÉN

FOI är en huvudsakligen uppdragsfinansierad myndighet under Försvarsdepartementet. Kärnverksamheten är forskning, metod- och teknikutveckling till nytta för försvar och säkerhet. Organisationen har cirka 1000 anställda varav ungefär 800 är forskare. Detta gör organisationen till Sveriges största forskningsinstitut. FOI ger kunderna tillgång till ledande expertis inom ett stort antal tillämpningsområden såsom säkerhetspolitiska studier och analyser inom försvar och säkerhet, bedömning av olika typer av hot, system för ledning och hantering av kriser, skydd mot och hantering av farliga ämnen, IT-säkerhet och nya sensorers möjligheter.

FOI
Totalförsvarets forskningsinstitut
Informationssystem
Box 1165
581 11 Linköping

Tel: 013-37 80 00
Fax: 013-37 81 00

www.foi.se

FOI-R--2974--SE
ISSN 1650-1942

Användarrapport
Mars 2010

Informationssystem

Henrik Karlzén

Metadata för objekt- och tjänstebaserad säkerhet

Titel	Metadata för objekt- och tjänstebaserad säkerhet
Title	Metadata for object and service based security

Rapportnr/Report no	FOI-R--2974--SE
Rapporttyp Report Type	Användarrapport User report
Sidor/Pages	19 p
Månad/Month	Mars
Utgivningsår/Year	2010
ISSN	ISSN 1650-1942
Kund/Customer	Försvarsmakten
Projektnr/Project no	E53055
Godkänd av/Approved by	Martin Rantzer

FOI, Totalförsvarets Forskningsinstitut	FOI, Swedish Defence Research Agency
Avdelningen för Informationssystem	Information Systems
Box 1165	Box 1165
581 11 Linköping	SE-581 11 Linköping

Sammanfattning

Idag finns mängder av data tillgänglig för både den enskilde individen liksom olika typer av organisationer. För att göra datan mer lätthanterlig finns ett behov av att rubricera och kategorisera den, vilket görs av metadata. Användningsområdena för det senare är mycket vida även inom det delområde, datoriserad metadata, som är denna rapports huvudsakliga ämne. Rapporten är en sammanställning av en litteraturstudie och beskriver bland annat viktiga standarder för publicering, indexering samt kommunikation av metadata såsom Dublin Core (DC), Web Services Description Language (WSDL), SOAP, Universal Description Discovery and Integration (UDDI) samt Platform for Internet Content Selection (PICS). Dessutom redogör rapporten för påverkande faktorer för framtagandet av en modell för metadata samt vilka hjälpverktyg som existerar för att exempelvis automatisera skapandet av metadata, samt de senares brister. Förslag på framtida projekt och arbeten inom området samt tips på frågeställningar, angående hur metadatamodellernas allmängiltighet ska behållas utan att göra avkall på de beskrivande och informativa delarna, ges. Slutligen diskuteras även vissa säkerhetsaspekter såsom vad metadatan avslöjar om datan samt hur metadata kan missbrukas.

Nyckelord: metadata, säkerhet, objektbaserad, tjänstebaserad, UDDI, WSDL, Dublin Core, XML, autogenerering, sökmotorer, webbtjänster, SOA, SOAP, RDF, PICS, POWDER, W3C, meta-taggar, keyword stuffing, DCMES, DCMI, interoperabilitet, linked data, semantiska webben, Web 3.0.

Summary

Today, there is a lot of data available for both the individual as well as different types of organizations. To make the data more manageable there is a need to classify and categorize it, which is accomplished by using metadata. The usage area of the latter is very broad even within the subset, computerized metadata, which is this report's main topic. The report is a compilation of a literature study and describes, among other things important standards for publishing, indexing and communication of metadata such as Dublin Core (DC), Web Services Description Language (WSDL), SOAP, Universal Description Discovery and Integration (UDDI) and Platform for Internet Content Selection (PICS). In addition, the report describes the influencing factors for the development of a metadata model and the tools that exist for e.g. automated creation of metadata as well as the latter's shortcomings. Suggestions for future projects and work in the area and tips on issues to be studied, regarding how the generality of metadata models may be retained without sacrificing the descriptive and informative parts, are given. Finally, also discussed are some security aspects such as what metadata reveals about the data and how metadata can be abused.

Keywords: metadata, security, object based, service based, UDDI, WSDL, Dublin Core, XML, auto generation, search engines, web services, SOA, SOAP, RDF, PICS, POWDER, W3C, meta tags, keyword stuffing, DCMES, DCMI, interoperability, linked data, Semantic Web, Web 3.0.

Innehållsförteckning

1 Inledning

I detta inledande avsnitt ges de problemformuleringar som arbetet ska besvara, samt en kort bakgrund till varför frågeställningarna valts. Dessutom beskrivs arbetets metodik och avgränsning samt rapportens struktur.

Denna rapport är en del i projektet Objekt- och tjänstebaserad säkerhet. Projektet ämnar till att frikoppla vissa säkerhetsmekanismer från applikationer och istället göra nätverket ansvarigt för säkerheten. I projektet bedömdes att det behövdes mer kunskap om metadata för att kunna vidareutveckla denna idé. Metadata är även ett viktigt ämne för de nya informationsinfrastrukturer, som bland annat syftar till att bättre koordinera insatser mot terrorism [38], vilka är under utveckling inom exempelvis Sveriges [37] och USA:s [34] försvarsdepartement samt NATO [36].

Rapporten ämnar besvara vad metadata är och vilka standarder som finns inom området. För projektet relevanta användningsområden och existerande projekt beskrivs. Dessutom ger rapporten stöd för utvecklandet av en metadatamodell samt genomgår vilka verktyg som kan underlätta och automatisera framtagandet av metadata.

1.1 Metod och Avgränsning

Då detta är en introducerande text till ämnet har det inom den givna tidsramen främst bedömts finnas utrymme för en litteraturstudie och sammanställning av denna samt några reflektioner. Utvecklingen, både historiskt och för närvarande, har antagits domineras av standardiseringsorgan varför främst information från dessa granskats. Eftersom att flertalet av de nämnda aktörerna öppet publicerar sina standarder och rekommendationer på webben har litteraturstudien baserats på dessa dokument.

Några djuplodande detaljer har inte tagits med såsom specifika fältformat eller ingående beskrivningar av kommunikationsprotokoll annat än som belysande exempel då syftet är ge en bredare kunskap inom området.

Eftersom att författaren efterhand har bedömt att mängden information om metadata i allmänheten varit god medan densamma inriktad mot det egentliga projektområdet i stort varit bristfällig har arbetet på grund av sin grundläggande nivå inte kunnat gå alltför mycket på djupet. Ytterligare avgränsning görs i ”2.1 Vad är metadata?” och ”2.2 Vad används metadata till?”.

1.2 Struktur

I detta arbete beskrivs metadata grundläggande. Från att börja med att definiera vad metadata är och vad det används till går rapporten sedan in lite djupare på olika exempel på system som använder sig av metadata och vilka standarder som finns på området. Anknytningar till sökmotorer och framförallt webbtjänster utreds och en genomgång av vilka typer av verktyg som finns genomförs. Rapporten avrundas med en kort diskussion samt slutsatser.

2 Metadata

Den största delen av rapporten baseras som skrivits främst av en ren litteraturstudie. Eftersom att den mesta av den relevanta informationen i ämnet bedömts finans i resurser på Internet har det varit den främsta källan.

2.1 Vad är metadata?

Metadata beskriver data och kan användas för människor eller datorer för att exempelvis söka och processa information på webbsidor [2]. Medan metadata betyder data om data har webbkonsortiet W3C en mer webbspecifik definition: ”Metadata is machine understandable information for the web” vilket på svenska blir ”Metadata är maskinförståelig information för webben” [3]. Hierarkiskt strukturerad metadata kallas även för ontologi [1]. Metadata kan givetvis också vara data och i sin tur ha metadata och så vidare.

Man hittar metadata på många olika ställen: i form av så kallade meta-taggar i HTML-filer eller filändelser i ett filsystem, som rubriker i en text, som en boks ISBN-nummer, med många fler. Den här rapporten kommer främst att fokusera på metadata för data på datorer även om många av resonemangen och idéerna som beskrivs kan överföras på andra applikationsområden.

Metadata delas ofta in i följande tre kategorier [4]:

- Beskrivande - metadata om datans innehåll och information
- Strukturella - metadata om relationer till annan data
- Administrativa - metadata om format och upphovsrätt och annat tekniskt

Se Tabell 2 i avsnittet ”2.4 Dublin Core och RDF” för ett belysande exempel som använder denna kategorisering.

Det kan vara värt att notera att Metadata Corporation registrerade ordet metadata som varumärke i USA 1981. Eftersom termen dock har en annan utbredd, mer allmän, betydelse är det juridiska läget oklart [40].

2.2 Vad används metadata till?

Det finns många användningsområden för metadata, några kan direkt skönjas vid en granskning av föregående avsnitt. Metadata används även exempelvis för att bygga innehållsförteckningar och för att söka data och tjänster. Mest fokus kommer att läggas på att söka tjänster men det närbesläktade området datasökning, och därigenom vanliga sökmotorer för webben såsom Yahoo och Google, kommer även att avhandlas kortfattat.

2.2.1 Sökmotorer

Traditionellt sett har sökmotorer på Internet använt sig av metadata i meta-taggar i HTML-koden på de indexerade sidorna för att leverera relevant resultat till användaren. Funktionen började dock missbrukas på slutet av 90-talet genom att oseriösa sidor ändrade meta-taggarna på olika sätt för att sökmotorerna skulle tro att sidan var mer relevant än den egentligen var. Exempelvis kunde mängder av, för sidan irrelevanta, nyckelord läggas till i metataggarna och på själva sidan för att på så vis få sökmotorerna att lista sidan oftare

och högre bland sökresultaten och på så sätt få fler besökare. Denna metod benämns ”keyword stuffing” och är en del av ”spamdexing” [41].

Numera använder därför de flesta större sökmotorer andra metoder för indexering även om den äldre varianten lever kvar i exempelvis intranät och för forskningsändamål. De nyare metoderna använder mer teknisk metadata såsom andelen fungerande länkar och hur välskriven källkoden är samt sidans arkitektur och storlek. Dessutom spelar avancerade faktorer som antalet länkar från andra sidor in liksom exempelvis språk, uppdateringsfrekvens, bevakning på sociala media och antalet klick på de av sökmotorn sponsrade reklamlänkarna på sidan [5]. Även vanligt förekommande fraser på de länkande sidorna har använts som särskilda nyckelord för den länkade sidan, vilket dock också utnyttjats för att manipulera sökresultaten [6].

Det är intressant att notera att en av de första stora sökmotorerna, Altavista, tidigt stödde den gamla typens metataggar (1996) men var bland de sista att sluta stöda de mer missbrukade då de inte gjorde det förrän 2002 [42]. Nuförtiden (december 2009) är det i princip endast Yahoo kvar som till någon grad förlitar sig på keywords-taggen, det vill säga nyckelorden även om dess inflytande är begränsat [43]. En annan metatagg, Description, vilken innehåller en kort beskrivning av sidan, används till viss mån fortfarande och dyker normalt upp tillsammans med sidans rubrik, vilken också är en meta-tagg, i sökresultaten. Det är även värt att notera att olika sökmotorer stöder Dublin Core till olika grad och på olika sätt [7].

2.2.2 Webbtjänster

Eftersom att den här rapporten är en del av ett projekt med visst fokus på så kallad tjänsteorienterad arkitektur - förkortad SOA efter engelskans Service-oriented architecture - undersöks och diskuteras som skrivits även metadatas kopplingar till sökmekanismer för att hitta tjänster på Internet [27]. Den vanligaste typen av SOA är baserad på webbtjänster.

W3C:s Web Services Architecture Working Group har definierat webbtjänster på följande vis [8]:

”A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards”.

Webbtjänster används med andra ord för att kunna sprida ut tjänster på ett nätverk. Även om andra standarder än de som nämns i definitionen från W3C används inom SOA för webbtjänster så är de som nämnts i definitionen de vanligaste och eftersom att även branschorganisationen WS-I rekommenderar dessa kommer alternativ ej att beskrivas här. SOAP, som ursprungligen var en förkortning av Simple Object Access Protocol [9], över HTTP används för kommunikation inom arkitekturen, både för att hitta och använda tjänster medan Web Services Description Language, förkortat WSDL [10], beskriver vad de olika tjänsterna erbjuder. Även om det inte nämns i definitionen ovan används normalt även ett register för att hålla reda på vilka tjänster som finns tillgängliga. Den vanligaste typen för ett sådant register, vilket dessutom används till att publicera och hitta tjänster, är UDDI, vilket står för Universal Description Discovery and Integration [11]. SOAP, WSDL och UDDI, som exempelvis används i amerikanska försvarsdepartements metadataprojekt [34], kommer att beskrivas närmare i avsnitt 2.8.

Metadata kan också användas för att klassificera data exempelvis för att filtrera sådant som en användare inte är auktoriserad att ha tillgång till. Ett exempel på ett sådant system, PICS, beskrivs i 2.9.

2.3 Att skapa en modell för metadata

I målbeskrivningen vid utvecklandet av Dublin Core, vilken beskrivs närmare i avsnitt 2.4, nämns några olika faktorer, vilka beror på bland annat användningsområde, som är relevanta att tänka på vid utvecklandet av en modell för metadata [12]:

- Generalitet – Hur specifik ska metadatan vara? Ju mer specifikt desto färre olika typer av system och data som kan använda modellen men å andra sidan kan metadatan bli desto utförligare och mer beskrivande. Ett exempel på ett metadata-fält som är generellt men samtidigt inte innehåller tillräckligt specifik information är ”format”-fältet. Exempelvis kan två filer med samma format och filändelse skilja mycket med avseende på till exempel färgskalans antal bitar med anledning av de interna inställningar som finns i filen [4].

- Användningsområde – Hur bred är användarbasen? Kommer det att krävas stöd för olika språk, exempelvis både svenska och engelska?

- Medium-oberoende – Ska modellen vara användbar på olika typer av medium eller bara på exempelvis datorer?

- Modellstorlek - Hur stor ska modellen vara? Olika storlekar ställer olika krav på hur lätta de är att utveckla, använda, underhålla och vidareutveckla.

- Kompatibilitet - Bör modellen passa in i de standarder som finns? Kompabilitet kan, förutom att underlätta kommunikation med andra system, även göra den initiala lärotiden kortare för användare som känner till liknande modeller.

I det följande avsnittet om Dublin Core ges ett exempel på hur en metadatamodell kan se ut.

2.4 Dublin Core och RDF

Det finns en rad olika standarder för hur metadata ska se ut. En mycket utbredd sådan standard som bland annat används i amerikanska försvarsdepartements metadataprojekt [34] är Dublin Core, vilken beskrivs i detta avsnitt.

Dublin Core Metadata Initiative (DCMI) är en organisation som har som mål att skapa interoperabla standarder för metadata. Dublin Core-metadata (DC) är medieoberoende på så vis att det inte specifikt måste användas för exempelvis Internet eller ens datorer utan kan tillämpas på all data, eller *resurser,* även om det i praktiken mestadels används för datorresurser [14] och kan ses som en utökad version av så kallade metataggar i HTML. I detta sammanhang innehåller en vokabulär bland annat definitioner av metadatafält samt relationer mellan dessa [15]. Den grundläggande vokabulären för DC kallas Dublin Core Metadata Element Set (DCMES) och definierar 15 olika basfält, eller egenskaper, vilka vart och ett beskriver resursen vilken metadatan är kopplad till. DCMES som var den första DCMI-standarden, formaliserad år 1998 genom RFC 2413 vilken 2007 ersattes av RFC5013 vilken i sin tur även finns som NISOZ3985 samt ISO15836:2009, är en delmängd av alla DC-vokabulärer [16].

En mer avancerad variant av DCMI blev populär för användning med W3C:s Resource Description Framework (RDF), en generisk datamodell för metadata vilken tillsammans med resursidentifierare, eller URI:er efter engelskans Unified Resource Identifiers, skapar så kallad ”linked data”. Det senare är en del i att koppla samman data på Web 3.0, den nya så kallade semantiska webben, vilken i sin tur är en mer avancerad variant av webben som ger ökad integration mellan data på olika platser, både på Internet och mellan Internet och

den fysiska världen [17]. Detta ledde år 2005 till en abstrakt modell för DC vilken i sin tur band samman linked data och Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) - en standard för validerbar metadata [18].

Om en organisation väljer att anpassa sig efter Dublin Core kan den göra det till olika grad beroende på hur speciellt det egna systemet behöver vara och hur mycket kraft, tid och pengar den vill lägga ner för att öka interoperabiliteten med andra system. Det finns fyra nivåer av interoperabilitet där den första är minst interoperabel [18]:

1. Gemensam vokabulär
   På denna nivå används gemensamma definitioner i vanligt, så kallat naturligt, språk. Det ger en gemensam begreppsvärld utan behov av formalisering. För specifika system såsom intranät väljs normalt denna, den vanligaste, nivån.

2. Gemensam formell vokabulär
   För utökad interoperabilitet och stöd för linked data används den formella modell som tillhandahålls av RDF. Eftersom nivån ger struktur för exempelvis sökmotoroptimering och ofta kan läggas till i efterhand är det här den variant som växer snabbast.

3. Syntax-interoperabilitet för beskrivningsmängder
   Förutom kompatibilitet med linked data går det att validera metadata mot så kallade beskrivningsmängder, eller description sets. Än så länge befinner sig nivån dock på ett mer experimentellt stadium och valideringsverktyg är bristfälliga - se avsnitt 2.7 för vidare beskrivningar.

4. Profil-interoperabilitet för beskrivningsmängder
   På den här nivån, vilken är än mer experimentell än nivå 3, delar de användande parterna även ”världsbild” och restriktioner för att kunna göra en mer komplett validering.

För att kunna uppfylla standarden utan att helt göra avkall på sitt synsätt när det gäller metadata kan så kallade kvalificerare användas, vilka gör metadatan mer utförlig och specifik utan att riskera minskad interoperabilitet. Exempelvis kan man istället för att bara sätta ett datum i DC:s fält ”date” dessutom skriva till vad det är för typ av datum, till exempel ”senast ändrad”, ”publicerad” eller liknande. Om metadatan senare ska användas med mer enkel och strikt Dublin Core, där endast okvalificerad metadata får brukas, ignoreras helt enkelt kvalificerarna enligt DC:s princip om ”fördumning”, efter engelskans Dumb-Down Principle och typen av datum blir åter ospecificerad [14].

Inom DC finns som beskrivits flera olika avancerade vokabulärer och DCMI Open Metadata Registry är ett semantiskt modelleringsverktyg för att definiera termer och relationer mellan olika DC-vokabulärer samt vokabulärer fristående från DC [19]. Oftast är de senare varianter vilka är mer eller mindre direkt översättbara till Dublin Core, exempelvis genom automatiska verktyg för datoriserad metadata vilka beskrivs närmare i avsnitt 2.7. Olika vokabulärer kan dock vara olika passande för att beskriva just de aspekter av datan vilka önskas belysas. Detta har uppmärksammats i DCMI genom det så kallade Warwick-ramverket [14].

DCMES består mer specifikt av 15 olika fält som bedömts vara viktiga [13] och de delas in i tre kategorier: Innehåll, Upphovsrätt och Instans. Följande tabell namnger de olika fälten och beskriver de vars innebörd inte direkt följer av namnet [16]:

Tabell 1 – DC-metadatafält kategoriserade enligt [16]

Kategori	**Fält**
Innehåll	Titel, Ämne (nyckelord), Beskrivning, Typ (bok, webbsida, etc.), Källa (källa för den beskrivna resursen, om sådan finns), Relationer (till andra resurser), Täckning (om resursens tillämplighet är begränsad till viss plats eller tid).
Upphovsrätt	Skapare (resursens skapare), Utgivare, Medskapare, Rättigheter (ex. om upphovsrätt).
Instans	Datum, Format (exempelvis storlek eller utrustning som krävs för läsning av resursen), ID (ISBN, URI, etc.), Språk (engelska, svenska, etc.).

Kategoriindelningen i Tabell 1 är alltså DCMI:s egna. Med indelningen från det inledande avsnittet om vad metadata är, 2.1, får vi som jämförelse följande tabell:

Tabell 2 - DC-medatafält kategoriserade enligt [4]

Kategori	**Fält**
Beskrivande	Titel, Ämne, Beskrivning, Typ, Skapare, Utgivare, Medskapare.
Strukturella	Källa, Relationer, ID.
Administrativa	Täckning, Datum, Rättigheter, Format.

Amerikanska försvarsdepartementet har utvecklat en utökad variant av Dublin Core som integrerar de grundläggande fälten med ett som specificerar säkerhetsrelaterad metadata [35].

Metadata används som skrivits i många sammanhang. Exempel på en lång rad projekt av varierande storlek som använder sig av Dublin Core finns på [20].

2.5 Kopplingen mellan metadata och data

Normalt sett är metadata och data sammankopplade med hjälp av ”länk-information” i metadatan. I metadatan kan det uttryckligen stå var datan finns, exempelvis på hylla 3 i rum 2 i byggnad Z på H-gatan eller IP 150.227.5.137, eller så kan det finnas en referens till en annan metadata-källa, som till exempel ISBN-registret för böcker. I Dublin Core används fältet Identifier, det vill säga identifierare, för denna sammankoppling [29].

Metadata kan även ligga inbakad i datan eller direkt utanpå, vilket eliminerar vissa problem men skapar andra. Exempelvis är det lättare att uppdatera metadata tillsammans med data när den senare ska ändras. Däremot är inte längre all metadata på ett ställe och det går således inte lika lätt göra sökningar med hjälp av denna. Jämför även avsnitt ”2.9 PICS”.

Om data och metadata ligger spridda från varandra, såsom i första stycket i detta avsnitt, uppstår vissa problem om datan ändras eller flyttas. På något sätt måste metadatan uppdateras och därför måste informationen om att datan ändrats på något sätt föras vidare. Ligger data och metadata tillsammans är problemet mindre men själva uppdateringen måste ske i vilket fall. Annars riskerar metadatan att bli inaktuell - se exempelvis [21].

2.6 Säkerhetsaspekter

När säkerhetsklassad data beskrivs med hjälp av metadata kan även den senare kräva säkerhetsklassning. Ska metadatan trots det vara sökbar och läsbar, även för den som ej har nödvändiga rättigheter för själva datan, bör det genomföras en noggrann avvägning av vad för information som kan lämnas ut genom metadatan. Även om själva informationen inte direkt är hemlig kan den användas för att gissa sig fram till vad som står i den dolda, krypterade och säkerhetsklassade texten, en metodik som är välkänd [22]. Om en del fält i icke säkerhetsklassad metadata då inte kan fyllas i av säkerhetsskäl måste det avgöras om den trots allt är tillräckligt beskrivande. Det måste då även bestämmas om själva datans existens och eventuellt även metadatans dito ska vara säkerhetsklassad [29]. Ett fall där inte ens tillgång till information om metadatans existens tillåts beskrivs kortfattat i avsnitt ”2.9 PICS” där sökmotorer blockerar olämpliga resultat från att dyka upp i resultatlistan.

Som användare kan man inte alltid ta för givet att metadatan verkligen är korrekt och att beskriver den data som den verkar göra. Exempelvis kan filändelser och ikoner för program mycket lätt ändras utan att påverka själva programmet. Metadata kan dessutom finnas utan användarens vetskap [21]. Ett exempel är att viruset Melissas skapare lyckades spåras upp genom metadata i Microsoft Word. Gömd metadata finns i flera former, exempelvis i form av små färgade fläckar på vissa datorskrivares utskrifter [23].

2.7 Autogenerering och verktyg

Det finns många metadataprocesser som kan automatiseras eller åtminstone underlättas med hjälp av olika datorverktyg. Inom Dublin Core finns och forskas det mycket på följande kategorier [24]:

- Konvertering – Används för att översätta olika metadata-scheman till och från DC.
- Skapande av metadata – Icke-automatiska verktyg vilka hjälper webbmasters, katalogiserare med flera att skapa metadata från data.
- Extrahering av metadata – En automatisk process för att skapa metadata från resursen med hjälp av substantiv-igenkännande algoritmer och frekvensanalys av fraser och/eller ord.
- Skördning– Att samla in metadata som redan är kopplad till resursen. Kan ses som en mer generell variant av konvertering eller översättning.
- Validering – En process för att validera att den skapade metadatan är så generell och interoperabel som önskats.

Se [25] för fler verktyg relaterade till DC.

De olika verktygstyperna är olika väl utvecklade och medan det exempelvis finns bra verktyg för konvertering eller översättning saknas i dagsläget valideringsverktyg. I takt med att de högre interoperabilitetsgraderna av DC utvecklas och implementeras kan sådana verktyg komma att utvecklas. Många av de olika typerna finns inbakade i ett och samma verktyg. Ett exempel som både kan konvertera, hjälpa till att skapa och extrahera är DC-dot från universitet i Bath, England [26]. Metataggar i HTML-kod kan översättas till Dublin Core och saknas nyckelordstaggar extraheras sådana från den angivna resursen. Normalt väljs exempelvis ord som står i fetstil eller i form av länkar. Verktyg för extrahering, det vill säga autogenerering, är i dagsläget av varierande kvalité och klarar endast av att hantera textbaserad data [4].

2.8 UDDI, WSDL och SOAP

UDDI, vilket nämnts ovan, är ett XML-baserat, plattformsoberoende register för webbtjänster och dessas interaktioner. UDDI, utvecklat av OASIS – Organisation for Advancement of Structured Information Standards - består av kontaktinformation såsom var tjänsten och dess tillhandahållare kan hittas samt kategori av tjänst och tekniska bitar inklusive gränssnitt till tjänsten [11]. Själva informationen finns inbakad i WSDL-dokument och för att publicera eller läsa sådana dokument om en tjänst i registret används SOAP [10] - ett XML-baserat protokoll över HTTP . UDDI, vilket lanserades i en andra version år 2002 [11] blev snabbt populärt och utvecklingen drevs av flera stora aktörer inom datorvärlden, såsom IBM och Microsoft med flera [28].

WSDL är också XML-baserat och blev W3C-rekommendation 2007 [39]. Dokumenten ser ut ungefär som en blandning av HTML och klassiska programmeringsdokument till sin struktur med definitioner och typer. Gränssnitt mot den specificerade webbtjänsten beskrivs liksom vilka operationer (jämför metoder och funktioner) som är möjliga. Dessutom finns information om format och vilka kommunikationsprotokoll, såsom HTTP, som används för att nyttja tjänsten. WSDL, som från början utvecklades av IBM och Microsoft, uppdaterades väsentligt i och med version 2.0 [39].

2.9 PICS

Ett annat projekt inom metadatasfären var W3C:s PICS (Platform for Internet Content Selection) vilket startade på mitten av 90-talet.

PICS var ett projekt som skapade en mängd tekniska specifikationer för att stöda mjukvaruverktyg och tjänster som tillhandahöll så kallade etikettlistor med information om en webbsidas kategori. Från början fokuserades endast på nätcensur, även benämnt innehållskontroll eller Internetfilter, för att skydda minderåriga från olämpliga sidor på webben via främst HTTP, men själva tekniken kom senare att användas även för annan typ av censur eller filtrering. Etikettorganisationen, på engelska kallad ”label bureau”, vilken erbjuder tjänsten med etikettlistor utvärderar olika webbsidor och placerar dem i en viss kategori, exempelvis ”barnförbjudet”. Sedan hämtar verktygen, vilka prenumererar på etikett-tjänsten, listorna. När en användare sedan försöker nå en webbsida kontrollerar verktyget först om användaren ifråga bör få tillgång till webbsidan baserat på inloggningsinformation hos användaren samt etikettlistorna.

Lokala administratörer kan normalt även sätta egna filtreringsregler vilka används tillsammans med en eller flera etikettlistor från de olika etikettorganisationerna. Sådana verktyg blev snabbt vanliga och är det än idag. Nuförtiden används inte just PICS i samma utsträckning men Microsofts Internet Explorer hade stöd för PICS från version 3 till och med 6.

För att skapa etikettlistor kan organisationerna ifråga exempelvis söka igenom webbsidor efter nyckelord som sedan avgör kategori. Istället för att använda etikettorganisationer kan även webbsideskaparna själva utvärdera sin sida och publicera kategorin tillsammans med den övriga HTML-koden, som metadata. Då är det dock viktigt att tänka på att sidan kanske går att nå på andra sätt än via HTTP, exempelvis via FTP. Normalt blockerar verktygen även andra vägar att nå sidan men ligger kategoriseringsinformation lokalt på sidan måste även andra vägar in täckas av webbsideskaparna själva [30].

För att kontrollera att kategorin fortfarande är aktuell för en specifik sida måste utvärderingen, det vill säga kategoriseringen, upprepas om sidan ändras, såvida den kategoriserande parten inte litar på att sidan fortfarande är kvar i samma kategori. En del etikettorganisationer tar hjälp av en kryptografisk kontrollsumma för att kontrollera huruvida innehållet på en sida ändrats eller ej.

PICS har efterträtts av POWDER (Protocol for Web Description Resources), som använder RDF och blev officiell W3C-rekommendation 1 september 2009. POWDER är en mer utbyggd version vilken dock ej längre stöder att webbsideskaparna själva bygger in etiketter i koden för att underlätta underhållet av etiketterna [31][32].

Det finns en del andra aspekter rörande den här typen av filtrering. Självfallet kommer det ta något längre tid att ladda en webbsida eftersom det måste göras en åtkomstkontroll. För att hålla nere fördröjningen kan etiketterna cachas även om det då finns en risk för att ligga något efter i de fall sidan hinner uppdateras innan cachen. Önskas allt som inte uttryckligen tillåts blockeras, en så kallad default deny policy, går det oftast att blockera okategoriserade sidor tillsammans med dem som redan kategoriserats som förbjudna. Detta bör även det korta ner åtkomsttiderna [30].

Olika sökmotorer på webben använder sig av liknande filter och exempelvis Altavista använde ett tag PICS för att inte låta sökresultat som länkade till blockerade sidor ens dyka upp [33].

3 Slutsatser och framtida forskning

Metadata är data som beskriver exempelvis annan datas innehåll, struktur samt upphovsrättsdetaljer och används inom en lång rad områden. För projektet, vilken rapporten är en del av, är främst möjligheterna att söka efter tjänster och data intressanta. Det finns många relevanta standarder på området (DC, WSDL, SOAP, UDDI, RDF, PICS med flera), vilka används till att beskriva, publicera, hitta, indexera samt kommunicera metadata. Dessutom kan PICS eller liknande användas för att filtrera eller censurera data.

För att skapa en egen metadatamodell måste olika faktorer, såsom användningsområde och –medium samt projektstorlek, vilka påverkar interoperabiliteten, undersökas. För implementationen måste det dessutom avgöras var metadata ska lagras, hos datan för enkelhets skull eller i lättsökbara register även om det senare kan leda till vissa problem då datan uppdateras. Även vissa säkerhetsaspekter finns angående vad metadatan avslöjar om datan, både när det gäller den senares existens samt innehåll. Det bör också noteras att det inte alltid är möjligt att lita på metadatans information om datan eftersom denna ofta kan ändras relativt enkelt, ibland för ändrarens egen vinnings skull. Slutligen har arbetet som framledde till rapporten endast lyckats hitta verktyg med begränsad funktionalitet för exempelvis automatisering av metadatagenerering.

Framtida studier och arbeten bör fokusera på att ta fram de verktyg för bland annat automatisering av skapande av metadata som idag saknas på området samt vidareutveckla de existerande verktygen. Dessutom bör det undersökas vilka ytterligare problem som finns och som prognostiseras för att sedan ta itu med dessa på en teoretisk och praktisk nivå. Mer specifika frågeställningar är bland annat hur metadatamodellernas allmängiltighet ska bibehållas samtidigt som de specificerande, beskrivande och informativa delarna inte begränsas. Ur ett rent säkerhetsperspektiv bör olika typer av sårbarheter hos de i rapporten nämnda protokollen och specifikationerna utredas och belysas varefter existerande lösningar och alternativ till dessa bör analyseras. Bland annat bör specifikationer för säkerhetsrelaterade metadataattribut få mer fokus.

Källförteckning

[1]	*OWL Web Ontology Language*, W3C, februari 2004. http://www.w3.org/TR/owl-guide/ (hämtad 3 november 2009).
[2]	*Metadata Activity Statement,* W3C, augusti 2002. http://www.w3.org/Metadata/Activity.html (hämtad 3 november 2009).
[3]	*Metadata and Resource Description,* W3C, april 2001. http://www.w3.org/Metadata (hämtad 3 november 2009).
[4]	*Understanding Metadata,* National Information Standards Organization, 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf (hämtad 3 november 2009).
[5]	*Search Engine Ranking Factors 2009,* SEOMoz, 2009. http://www.seomoz.org/article/search-ranking-factors (hämtad 2 december 2009).
[6]	Zeller Jr T., *A New Campaign Tactic: Manipulating Google Data*, The New York Times, oktober 2006. http://www.nytimes.com/2006/10/26/us/politics/26googlebomb.html?_r=1 (hämtad 2 december 2009).
[7]	Mårtensson N., Vakhrameeva E., *Dublin Core i praktiken: En undersökning av hur Dublin Core används inom fem svenska söktjänster*, Högskolan i Borås/Institutionen Biblioteks- och informationsvetenskap (BHS), 2003. http://bada.hb.se/handle/2320/953 (hämtad 2 december 2009).
[8]	*Web Services Glossary*, W3C, februari 2004. http://www.w3.org/TR/ws-gloss (hämtad 5 november 2009).
[9]	*SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)*, W3C, april 2007. http://www.w3.org/TR/soap12-part1/ (hämtad 5 november 2009).
[10]	*Web Services Description Language*, W3C, mars 2001. http://www.w3.org/TR/wsdl] (hämtad 6 november 2009).
[11]	*UDDI Version 3.0.2*, OASIS, februari 2004. http://uddi.org/pubs/uddi_v3.htm (hämtad 6 november 2009).
[12]	Weibel S et al., *Dublin Core Metadata for Resource Discovery* (RFC 2413), Internet Society, september 1998. http://www.ietf.org/rfc/rfc2413.txt (hämtad 10 november 2009).
[13]	*A Framework of Guidance for Building Good Digital Collections 3rd edition,* National Information Standards Organization, december 2007. http://www.niso.org/publications/rp/framework3.pdf (hämtad 12 november 2009).
[14]	*DCMI Frequently Asked Questions (FAQ)*, DCMI, 2009. http://dublincore.org/resources/faq (hämtad 11 november 2009).
[15]	*Vocabularies*, W3C, 2009. http://www.w3.org/standards/semanticweb/ontology (hämtad 13 november 2009).
[16]	*Dublin Core Metadata Element Set*, Version 1.1, DCMI, juni 2008. http://dublincore.org/documents/dces (hämtad 13 november 2009).
[17]	*W3C Semantic Web Activity*, W3C, december 2009. http://www.w3.org/2001/sw (hämtad 17 november 2009).
[18]	*Metadata Basics*, DCMI, 2009. http://dublincore.org/metadata-basics (hämtad 13 november 2009).
[19]	*The Dublin Core Metadata Registry*, DCMI, 2008. http://dcmi.kc.tsukuba.ac.jp/dcregistry (hämtad 17 november 2009).
[20]	*Dublin Core Projects – Alphabetical,* DCMI, 2009. http://dublincore.org/projects (hämtad 19 november 2009).

[21]	Schneier B., *Schneier on Security: Metadata in MS,* Schneier on Security, november 2005. Officehttp://www.schneier.com/blog/archives/2005/11/metadata_in_ms.html (hämtad 25 november 2009).
[22]	Menezes A. et al., *Handbook of Applied Cryptography Chapter 1,* CRC Press, 1997. http://www.cacr.math.uwaterloo.ca/hac/about/chap1.pdf (hämtad 8 december 2009).
[23]	*Secret Code in Color Printers Lets Government Track You*, Electronic Frontier Foundation, oktober 2005. http://www.eff.org/press/archives/2005/10/16 (hämtad 27 november 2009).
[24]	*DCMI Tools Glossary,* DCMI, maj 2007. http://dublincore.org/groups/tools/glossary.shtml (hämtad 17 november 2009).
[25]	*Tools and Software,* DCMI, 2009. http://dublincore.org/tools (hämtad 17 november 2009).
[26]	Powell A., *Dublin Core metadata editor*, UKOLN, University of Bath, augusti 2000. http://www.ukoln.ac.uk/metadata/dcdot (hämtad 17 november 2009).
[27]	*Web Services Architecture,* W3C, februari 2004. http://www.w3.org/TR/ws-arch/ (hämtad 6 november 2009).
[28]	*Microsoft, IBM, SAP To Discontinue UDDI Web Services Registry Effort*, SOA World Magazine, december 2005. http://soa.sys-con.com/node/164624 (hämtad 3 december 2009).
[29]	Kunze J., Baker T., *The Dublin Core Metadata Element Set* (RFC 5013), Internet Society, augusti 2007. http://www.ietf.org/rfc/rfc5013.txt (hämtad 16 november 2009).
[30]	*PICS Frequently Asked Questions (FAQ),* W3C, januari 2003. http://www.w3.org/2000/03/PICS-FAQ (hämtad 8 november 2009).
[31]	*PICS Superseded by POWDER,* W3C, november 2009. http://www.w3.org/2009/08/pics_superseded.html (hämtad 8 november 2009).
[32]	*Protocol for Web Description Resources (POWDER) Working Group*, W3C, november 2009. http://www.w3.org/2007/powder (hämtad 8 november 2009).
[33]	*Platform for Internet Content Selection (PICS)*, W3C, november 2009. http://www.w3.org/PICS/#Innovations (hämtad 8 november 2009).
[34]	*Customer Conference 2007 NCES Developer Workshop,* Defense Information Systems Agency, Department of Defense, maj 2007.
[35]	*Department of Defense Discovery Metadata Specification (DDMS) Version 1.3,* Department of Defense, juli 2005.
[36]	*NATO Network-Enabled Capability,* NATO, 2010. http://nnec.act.nato.int/pages/about.aspx (hämtad 27 januari 2010).
[37]	*Development of NERE metadata specifications for technical and software systems,* Försvarets materielverk, maj 2007.
[38]	*Executive Order 13356 of August 27, 2004, Strengthening the Sharing of Terrorism Information To Protect Americans,* White House, augusti 2004. http://www.fas.org/irp/offdocs/eo/eo-13356.htm (hämtad 27 januari 2010).
[39]	*Web Services Description Language (WSDL) Version 2.0,* W3C, juni 2007. http://www.w3.org/TR/wsdl20/ (hämtad 2 december 2009).
[40]	*United States Patent and Trademark Office,* USPTO, 2009. http://www.uspto.gov/ (hämtad 6 december 2009).
[41]	*Spamdexing,* Answers.com, 2009. http://www.answers.com/topic/spamdexing-technology (hämtad 9 december 2009).
[42]	*Death of a Meta Tag,* SearchEngineWatch, oktober 2002. http://searchenginewatch.com/2165061 (hämtad 9 december 2009).

[43]	*Sorry, Yahoo, You DO Index The Meta Keywords Tag,* Search Engine Land, oktboer 2009. http://searchengineland.com/sorry-yahoo-you-do-index-the-meta-keywords-tag-27743 (hämtad 9 december 2009).