MAGISTERUPPSATS I BIBLIOTEKS- OCH INFORMATIONSVETENSKAP
VID BIBLIOTEKS- OCH INFORMATIONSVETENSKAP/BIBLIOTEKSHÖGSKOLAN
2003:102

Ontologier i kunskapsorganisation

Vägen från tesaur till den semantiska webben

Iréne Granström

Svensk titel:	Ontologier i kunskapsorganisation - Vägen från tesaur till den semantiska webben
Engelsk titel:	Ontologies in Knowledge Organization – From Thesaur to the Semantic Web
Författare:	Iréne Granström
Kollegium:	Kollegium 2
Färdigställt:	HT 2003
Handledare:	Johan Eklund

Abstract:

This thesis aims to analyse ontology, primarily as the word is used in the context of the Semantic Web. One of the main questions is how ontologies relate to classification and thesauri, two concepts well known within Library and Information Science.

The vision of the Semantic Web, a more intelligent addition to the current World Wide Web, has raised the question of how to deal with information scattered over a multitude of locations, stored in different formats, written in different languages et cetera. This issue is in many ways similar to the classical problem encountered in library science: how to classify and categorise large amounts of information. The use of ontologies is one of the main steps leading to the Semantic Web, as well as a tool which can be used in other areas of information organisation and management. The main use of an ontology is to provide a standardised way of describing an area of interest.

The results of the study show that there are several similarities between ontologies, classification and thesauri, but also that ontologies add some aspects compared to the others. The main differences are that ontologies require that relationships are defined formally in order to avoid ambiguities, and that they can handle concepts in addition to terms.

This makes ontologies more powerful, especially when using computer based systems. Utilising an ontology makes it possible for a computer to draw conclusions based on the data provided, something that is not possible when using a thesaurus.

Nyckelord: ontologi, klassifikation, tesaurer, semantisk webb, WWW, Oil

Innehållsförteckning

1. Inledning

Många av oss använder dagligen Internet och framförallt den del av Internet som kallas World Wide Web, WWW. Redan vid enkla sökningar efter specifik information inser man vilken enorm mängd information som finns tillgänglig på det som populärt kallas ”nätet” eller ”webben”, men att det faktiskt finns ett antal miljarder webbsidor är väl inget man reflekterar över varje dag i sin normala användning av WWW. Däremot tror jag att de flesta skulle bli gladare om det blev lättare att hitta vad man är ute efter i denna enorma informationsmängd. Dessutom vore det bra om man visste att man kunde lita på det som man hittar.

I denna uppsats kommer jag att behandla ett av de initiativ som finns för att försöka bringa ordning i den uppsjö av dokument, bilder, filmer och annat som finns tillgängligt på Internet, nämligen det som kallas den semantiska webben. Den semantiska webben är tänkt som en utökning av den nuvarande, inriktad på att datorer skall kunna tolka och utbyta information automatiskt och därmed kunna ge bättre svar på våra frågor. Visionen uttrycks så här:

> den semantiska webben är en vision: en dröm om att data på webben ska definieras och sammanlänkas så att de kan användas av maskiner, inte bara för att visas, utan även för automatisering, integrering och återanvändning av data via olika applikationer (Berners-Lee et al. 2001) (författarens översättning).

För att kunna uppnå denna dröm krävs ett antal steg på vägen. Ett av dessa är ontologier, vilket innebär att man på ett systematiskt sätt definierar hur ett område skall beskrivas, alltså relationen mellan termer. Detta är ju något som starkt påminner om metoder som används i traditionell informationshantering såsom den bedrivits inom biblioteks- och informationsvetenskapen, där man ju också i många sammanhang hanterar stora mängder information.

Uppsatsen har därför som fokus just ontologi, såsom det är tänkt att underlätta uppbyggnaden av den semantiska webben och hur detta begrepp står i relation till kunskapsorganisatoriska begrepp som klassifikation och tesaurer.

Tanken med att skriva en magisteruppsats om ontologier och framförallt i samband med den semantiska webben har vuxit fram efter hand. Ämnet är intressant och kopplingen till Biblioteks- och Informationsvetenskap relevant att belysa, eftersom uppbyggandet av en ontologi i hög grad påminner om tesaur- och klassifikations-uppbyggnad.

Ämnet är i hög grad aktuellt både för mig själv och för många andra, eftersom resultatet i förlängningen kan göra det lättare för oss att hitta rätt i informationsmängden när vi söker information, både på WWW och i andra sammanhang, såsom veteskapliga databaser och företagsinterna intranät. Jag tror att många av oss skulle uppskatta om det blev enklare att finna den information man söker. Om det dessutom går att bygga system som gör att man vet att man kan lita på informationen är det ännu mer eftersträvansvärt.

1.1. Bakgrund och problembeskrivning

Internet och WWW har gett oss tillgång till en otrolig mängd information och man räknar idag med att det finns flera miljarder dokument på WWW. Dessutom finns en stor mängd information på företagsinterna intranät och uppskattningsvis mer än 600 miljoner människor använder sig dagligen av information hämtad från intra- och Internet. Att informationsmängden ständigt ökar och dessutom med en avsevärd hastighet, gör att det blir allt svårare att organisera, hitta och underhålla den. Olika sök- och indexeringsfunktioner har försökt att strukturera vissa delar av det som finns på WWW, men det blir ändå bara en bråkdel av den tillgängliga informationen som blir sökbar. Utöver Internet och intranät finns dessutom ett stort antal andra informationskällor i form av databaser av olika slag, t.ex. över vetenskapliga tidskrifter eller mediciner och deras verkningar och biverkningar.

Ett av de stora problemen som gör att det är svårt att hitta på webben är att man skriver nästan all information i ett naturligt språk men detta kan inte datorer behandla och använda på ett effektivt sätt. Vi använder oss av datorbaserade system för att hitta informationen genom att ange sökord eller söksträngar, som vi önskar svar på. Om sedan inte datorn kan förstå vad ordet betyder, måste sökningen ske på syntax. Det betyder att vi söker efter förekomsten av ord i informationen men inte efter dess betydelse. På grund av detta blir sökningar efter information på webben i många fall ineffektiva.

Dessutom har vi de senaste 30 åren sett trenden att kunskap blir en allt viktigare konkurrensfaktor för många företag och organisationer, vilket gör att både hantering av redan inhämtad information och sökning av ny får en allt större betydelse. Det finns också ett antal begränsningar i de nuvarande systemen:

- *Sökning*: som redan nämnts sker sökning oftast via nyckelord, vilket innebär att det finns risk både för att få en stor mängd irrelevanta träffar (låg precision) och samtidigt gå miste om relevant information där t.ex. synonymer till sökorden använts (begränsad återvinningsgrad, recall)
- *Analys av sökresultaten*: idag krävs i de flesta fall att en människa går igenom och analyserar resultaten av sökningar för att man skall få ut ett användbart resultat.
- *Underhåll*: upprätthållandet av informationsmängden kräver ofta också mycket inblandning av människor eftersom det i många fall handlar om ostrukturerad information, vilken är svår att behandla automatiskt.

En annan aspekt är det begränsade användandet av automatiskt genererad information. Med hjälp av intelligent utformade mallar och andra verktyg för t.ex. webbsidor och dokument skulle mycket information som vore användbar i sök- och indexerings-sammanhang kunna genereras automatiskt. Idag sker oftast inte detta, utan det krävs att denna information läggs in för hand, vilket oftast betyder att det inte blir gjort.

En stor mängd information existerar också i form av bilder, ljud och multimediafiler, vilka ofta är ännu mindre strukturerade än traditionella textdokument.

Som svar på dessa utmaningar kommer då t.ex. visionen om den semantiska webben och de kompletteringar av nuvarande tekniker som behövs för att uppnå denna vision. Ett av de centrala begreppen i detta sammanhang visar sig vara ontologi och teoribildningen kring uppbyggnaden av ontologier. Den definition som ligger till grund för användningen av ontologi myntades i början av 1990-talet vid Stanford University av Thomas Gruber: ”An ontology is a specification of a conceptualization.” (Gruber 1993b).

Denna definition är flitigt både citerad och diskuterad. Det är ju inte helt uppenbart vad den innebär eftersom det samtidigt krävs att man är överens om vad specifikation och konceptualisering innebär, i kapitel 4.2 kommer jag försöka att förklara definitionen närmare.

Vad är då problemet som gör att detta uppsatsarbete är intressant? Jo, de starkast drivande krafterna inom både den semantiska webben och forskningen och utvecklingen av ontologier kommer från den datavetenskapliga disciplinen med ursprung både i den ”traditionella” webben och expertsystem. När man analyserar ontologibegreppet närmare finner man dock att det finns stora likheter med hur traditionella klassifikationssystem och tesaurer är uppbyggda. Det har också genererat en debatt huruvida man är på väg att uppfinna hjulet igen genom att inte tillvarata den kunskap som finns inom informationsvetenskap och kunskapsorganisation. Det kom t.ex. ett förslag från Dagobert Soergel 1996 att skapa en gemensam plattform för denna typ av arbete (Soergel 1996). Jag vill därför analysera vilka skillnader och likheter det finns för att tydliggöra relationen mellan disciplinerna och de synergieffekter som skulle kunna uppnås av ett närmare samarbete.

1.2. Syfte och frågeställning

Syftet med uppsatsen är att belysa ontologibegreppet såsom det används framförallt i utveckling av den semantiska webben och ställa detta i relation till klassiska kunskaps-organisatoriska begrepp som klassifikationsscheman och tesaurer. Flera av dessa begrepp används i olika sammanhang och kan därför ha något skilda betydelser, beroende på tillämpningsområde och vem som använder begreppet. Mitt perspektiv i uppsatsen är en analys av hur ontologier, såsom detta begrepp används i uppbyggnaden av den semantiska webben, förhåller sig till de i kunskapsorganisation existerande metoderna att strukturera information och information om information. Med avsikt att underlätta för läsaren ges därför en beskrivning av den semantiska webbens komponenter så att man får en förståelse för ontologibegreppets betydelse i sitt tänkta sammanhang.

Följande frågeställningar har upprättats:

1. Vilken relation finns mellan ontologi, tesaur och klassifikationsscheman?
2. Tillför ontologi något utöver det som finns i klassifikationsscheman och tesaurer?
3. Vilken betydelse har ontologier för den semantiska webbens tillkomst?

1.3. Avgränsningar

För att ge uppsatsarbetet en rimlig omfattning har vissa avgränsningar gjorts. Den semantiska webben är inte huvudmålet i uppsatsen, men för att förstå betydelsen av ontologier i den semantiska webben behövs en bakgrund, bl.a. i form av XML (eXtensible Markup Language), ett uppmärkningsspråk, och RDF (Resource Description Framework), ett ramverk för bl.a. metadata. En kort genomgång av dessa ingår därför i uppsatsen, men detta skall ses som nödvändig bakgrundsinformation för uppsatsen, inte som komplett information om hur man lär sig använda dessa verktyg.

För att ytterligare avgränsa mig har jag alltså inriktat mig på en av komponenterna i den semantiska webben, nämligen ontologi eftersom detta är ett mycket aktivt forskningsområde och det finns tydliga kopplingar till kunskapsorganisation. Begreppet ontologi kommer ursprungligen från filosofin och betyder ”läran om vad som är verkligt” (Nationalencyklopedin 2003). Ett så vittomfattande begrepp blir naturligtvis för stort i detta sammanhang, så jag kommer endast att behandla ontologi så som begreppet används i relation till informationshantering och den semantiska webben. Definitionen är i huvudsak den som Gruber (Gruber 1993b) har givit begreppet ontologier. Denna definition anser även Berners-Lee vara den som bäst passar i sammanhanget (Fensel et al. 2003).

En annan avgränsning är att jag inte kommer att gå in i detalj på de komponenter som bygger vidare från ontologierna, de så kallade ”Logic”, ”Proof” och ”Trust”, utan endast ge en översiktlig bild över hur dessa passar in i sammanhanget. Jag kommer inte heller gå in på metadata, även om detta används i diskussionen kring den semantiska webben. Detta är visserligen ett relevant område men det har behandlats av många andra och skulle också göra att uppsatsen blir alltför stor.

Eftersom ontologier används i flera sammanhang har jag som avgränsning använt mig av material kring ontologier som har anknytning till den semantiska webben och kunskapsorganisation.

1.4. Uppsatsens disposition

Kapitel 1 börjar med en inledning och går sedan över i en gemensam bakgrund och problembeskrivning som sedan mynnar ut i syfte och frågeställning. Här finns även uppsatsens disposition och de avgränsningar jag har valt att göra.

Kapitel 2 kallar jag för referensram, eftersom jag här beskriver de olika områden som jag grundar min analys av ontologier på, nämligen klassifikation, tesaurer och den semantiska webben. I detta kapitel återfinns även en grunddefinition av ontologi.

I kapitel 3 beskriver jag min metod samt hur jag gått tillväga med material- och informationssökning.

De två följande kapitlen utgör stommen i uppsatsen i form av beskrivning av ontologi, varifrån ordet kommer, hur det används i expertsystem och kunskapsorganisation (kapitel 4), samt vilken betydelse ontologier har för den semantiska webben (kapitel 5). Kapitel 5 innehåller också ett avsnitt om den semantiska webbens komponenter, eftersom detta är nödvändigt för att förstå varför ontologier behövs.

Kapitel 6 ger ett exempel på hur en tesaur har konverterats till en ontologi. Jag skriver kort om språket som använts och hur konverteringen har gått till.

Därefter kommer kapitel 7 med analys och diskussion samt kapitel 8 där uppsatsens resultat sammanfattas.

2. Teoretisk ram

I detta kapitel ger jag en referensram där jag lägger en grund till den analys som kommer senare i uppsatsen. Mitt perspektiv i uppsatsen är en analys av hur ontologier, såsom detta begrepp används i uppbyggnaden av den semantiska webben, förhåller sig till de i kunskapsorganisation existerande metoderna att strukturera information och information om information.

Jag har därför valt att som utgångspunkt ur kunskapsorganisatoriskt perspektiv ha klassifikation och tesaurer som en bas att utgå ifrån när jag söker kunskap om ontologi. Som kompletterande kunskapsbas behövs också den semantiska webben, eftersom det är denna som dels väckte mitt intresse för ämnet och dels är en av de starkaste drivkrafterna för utvecklingen av ontologier i dagsläget.

Detta kapitel innehåller också en grunddefinition av ontologibegreppet såsom det kommit att användas inom detta område.

2.1. Referensram

2.1.1. Klassifikation

Det finns likheter mellan klassifikationsscheman och ontologier, därför vill jag i detta kapitel beskriva hur ett sådant kan vara konstruerat. En andra orsak är att det av vissa forskare dras ett likhetstecken mellan klassifikation och ontologier. Därför kan det vara bra att titta närmare på dess uppbyggnad för att sedan analysera om det stämmer. Syftet med klassifikation är att organisera kunskapen i dokument så att den blir tillgängliga för dem söker efter den. Detta är också anledningen till att man skapar ontologier.

Klassifikation är något vi dagligen sysslar med utan att vi tänker på det. Vi sorterar upp saker som kläder, t.ex. strumpor på ett ställe och tröjor på ett annat. Denna typ av klassifikation kan också kallas taxonomi och urtypen för ett sådant system är Linnés klassifikation av arter. Att vi systematiskt ordnar kunskap, t.ex. samlingar av dokument, fyller två viktiga funktioner: det ger oss en ämnesöversikt över området, och det ger oss möjlighet att söka efter information om ett speciellt ämne utan att behöva söka igenom samtliga dokument. (Rowley & Farrow 1992, s. 192)

Bibliografisk klassifikation och klassifikation i det dagliga livet skiljer sig bl.a. genom att den bibliografiska i huvudsak organiserar kunskap i dokument för att det skall bli lättare att hitta för den som söker. Bibliografisk klassifikation använder egentligen samma teknik som klassifikation i det dagliga livet, men har huvudfokus på dokument och hur ämnen är representerade i dessa dokument. För andra typer av klassifikation är själva objekten oftast huvudintresset, inte informationen eller dokumentationen om dem. (Harvey 1999, s. 203)

Bibliografisk klassifikation använder koder för att beskriva hur ett objekt är klassificerat. Koderna som används kan vara siffror, bokstäver eller en kombination av båda. Den viktigaste funktionen för en kod är att visa på klassernas systematiska ordning. Ingången till ett klassifikationssystem går genom ett alfabetiskt ämnesregister, där det sedan finns hänvisningar till klassens beteckning (kod).

Relationen mellan klasser är viktigt att ha i minnet när man klassificerar. Ett dokument eller en bok handlar mycket sällan om endast en sak. Det gör att man måste definiera relationen mellan klasserna. Klasserna delas in i enkla eller sammansatta.

Vidare skiljer man mellan enumerativa, hierarkiska och facetterade system.

- Enumerativa system – de system som räknar upp ämnena och som löst grupperar relaterade ämnes objekt.

- Hierarkiska system – liknar enumerativa scheman med den skillnaden att de grupperar relaterade objekt i över- och underordnade klasser med målet att skapa en så naturlig uppdelning som möjligt.

- Fasetterat system – börjar ifrån en annan grund. Ämnena bryts ned i enskilda kategorier (facetter) och det finns en beteckning för varje fasett. Man kan t.ex. klassificera musik utifrån musikform, instrument och tidsperiod. Varje fasett får sedan sina speciella koder och den totala klassificeringen blir en kombination av dessa. (Harvey 1999, s. 205f)

I bibliografiska klassifikationscheman finns både generella och speciella scheman. Generella scheman täcker all dokumenterad kunskap. De är utvecklade för stora dokumentsamlingar som täcker stora ämnesområden för folkbibliotek. Exempel på sådana scheman är det svenska SAB, Dewey Decimal Classification (DDC) och Universal Decimal Classification (UDC). Speciella scheman är de som täcker ett specifikt kunskapsområde, exempel på detta är British Classification of Music. (Harvey 1999, s. 203f)

En viktig faktor att komma ihåg i detta sammanhang är att det i klassifikationsscheman inte finns några hänvisningar, vilket däremot är en viktig aspekt i både tesaurer och ontologier.

2.1.2. Tesaurer

I detta kapitel vill jag gå igenom vad en tesaur innehåller, för att sedan kunna analysera skillnader och likheter mellan en ontologi och en tesaur. Detta är också intressant eftersom man från bibliotekshåll vill bli mer delaktiga i utvecklandet av den semantiska webben och även vid utvecklandet av ontologier.

Ordet tesaur har sitt ursprung ifrån latinet och grekiskan och betyder ”ordskatt” (Baeza-Yates & Ribeiro-Neto 1999, s. 170). Den första tesauren är gjord år 1852 av

Peter Mark Roget. Titeln är ”Thesaurus of English Words and Phrases”. Rogets tesaur är av allmän karaktär, men det finns också tesaurer som är till för ett specifikt område. I detta sammanhang när det gäller ontologier, är specifika tesaurer mest intressanta, eftersom man oftast tänker sig att ontologier skall byggas upp från mindre, existerande ontologier som sedan kan kopplas till varandra. Tesaurer började användas igen under senare delen av 50-talet. Sedan 1974 finns det standardiserade regler för hur en tesaur skall byggas upp, ISO har en för flerspråkiga (2788) och en för enspråkiga (5964) (Chowdhury 1999, s. 125).

Huvuduppgifterna för en tesaur är enligt Aitchison:

> … det primära syftet för en tesaur är informationsåtervinning, det kan åstadkommas på olika sätt. Andra syften är en generell förståelse för ett ämnes område, förse området med en ”semantisk karta” genom att visa på relationer sins emellan, och hjälpa till att förse termerna med definitioner. (Aitchison et al. 1997, s.1) (författarens översättning)

En tesaur är en lista av viktiga ord inom ett avgränsat ämnesområde. För att skapa en tesaur använder man sig av kontrollerad vokabulär, en standardiserad lista för de ord som får användas vid indexeringen. En tesaur innehåller korshänvisningar som anger relationen mellan termerna i listan, och de grupperas idémässigt. Det skall också finnas en alfabetiskt uppställd förteckning över alla indexeringsord. Relationen mellan termerna visas både semantiskt och syntaktiskt. Termerna som används är ämnesord och dessa är godkända vad gäller grammatik och syntax. De kallas även för deskriptorer.

Det finns tre olika typer av semantiska relationer: ekvivalenta, hierarkiska och associerade relationer.

Ekvivalenta relationer

De ekvivalenta relationerna visar på relationen mellan en godkänd term och en icke-godkänd term när två eller flera termer finns för samma begrepp. Den godkända termen är den som är vald till att representera begreppet vid indexeringen, medan den icke-godkända är den som inte är utvald. Den icke-godkända termen visar på en ingångsterm som visar till en som är godkänd. Ekvivalenta relationer visas genom att skriva USE (används) och UF (används för). Ekvivalenta relationer inkluderar synonymer, antonymer, stavningsvarianter och förkortningar.

USE – står efter en icke-godkänd term och hänvisar till en godkänd term.
UF – Use For: står efter godkänd term och hänvisar till en icke-godkänd term.

Exempel:

elev
USE student

student

UF elev

Hierarkiska relationer

Hierarkiska relationer är de relationer som i huvudsak skiljer en tesaur från en ämnesordslista. Dessa relationer anger partitiva relationer (del av), ex. ben – bord, benen ingår som en del av bordet, och för att visa detta används BT och NT.

BT – Broader Term, hänvisar från en mer specifik term till mer allmän term.
NT – Narrower Term, hänvisar från mer allmän term till mer specifik term.

Exempel

databasspråk
NT frågespråk

databasspråk
BT programmeringsspråk

Associerade relationer

Associerade relationer visar på relationer, som varken är hierarkiska eller ekvivalenta, utan det rör sig om närbesläktade begrepp som associeras med en term, ex undervisning - lärare. För att visa detta används RT.

RT – Related Term

Syftet med dessa relationer är att sätta in ämnesorden, deskriptorerna i ett sammanhang. Ibland kan det dock behövas en förklaring av en term och då används Scope Note (SN). De flesta deskriptorerna följs inte av en Scope Note eftersom det går att förstå innebörden av termen i den aktuella tesauren, men när det inte tydligt framgår vad deskriptorn betyder, så skrivs en förklaring för betydelsen i den aktuella tesauren.

2.1.3. Semantiska Webben

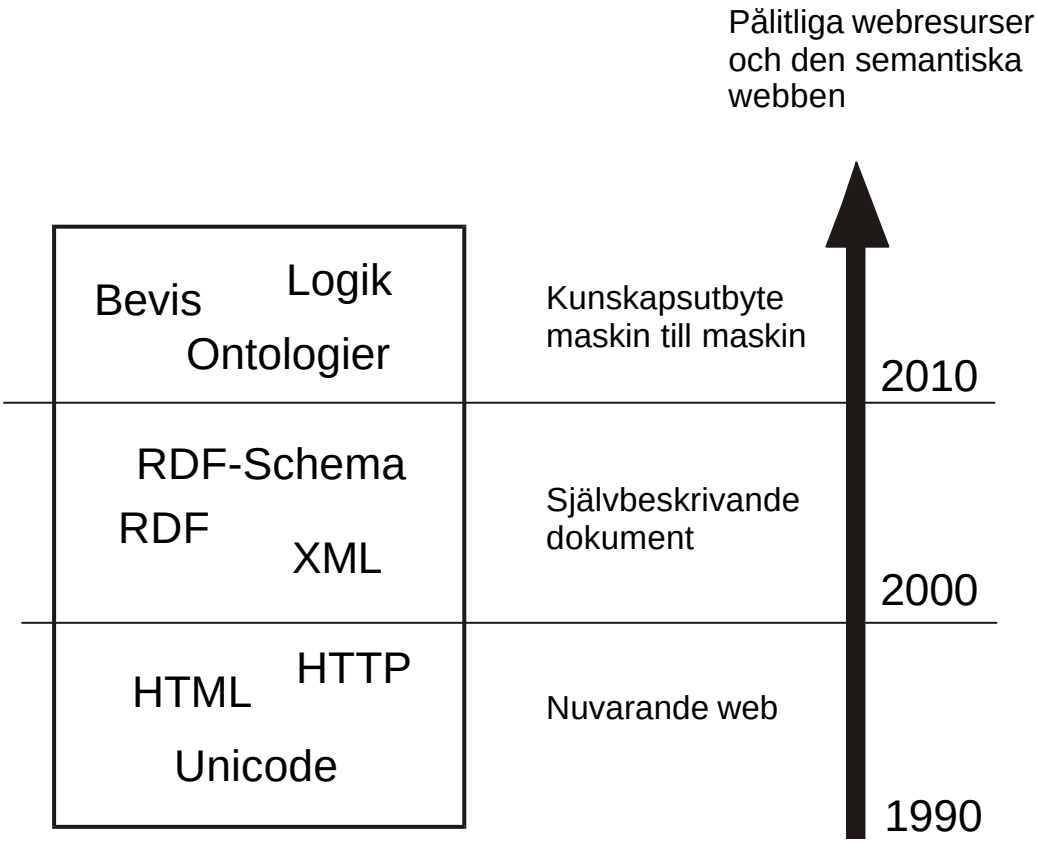
Ett sätt att försöka hantera den stora mängden av information är att utveckla ett system, som kan uttrycka information på ett sätt som datorer kan förstå och bearbeta. Detta är vad grundaren av World Wide Web (WWW), Tim Berners-Lee, har presenterat som vision för den så kallade semantiska webben. Den semantiska webben är inte en separat webb utan en utbyggnad av den nuvarande (Berners-Lee et al. 2001). Som tillägg till att vara läsbar av människor via webbläsare som Internet Explorer och Netscape, vill man se till att alla webbresurser också har metadata kopplat till sig. Metadata är data om data, dvs. extra information som beskriver vilken typ av data det

rör sig om och dessutom utformad på ett sådant sätt att datorsystem kan tolka den. Detta, i samverkan med en systematisk metod att beskriva ett område, vilket är där begreppet ontologi blir intressant, är tänkt att kunna lyfta webben och vår användning av den till en ny nivå.

I detta sammanhang bör också påpekas att den semantiska webben inte är tänkt att vara riktad mot speciella användningsområden, utan vara lika generell som den nuvarande webben är idag (Berners-Lee et al. 2001).

För att nå dit man vill, vilket alltså är högt ställda mål, finns ett antal komponenter eller steg som måste finnas och vissa av dessa påminner om områden inom kunskaps-organisation, fast i en ny tappning. Ett av dessa är ontologier. Detta kan vara lätt att glömma bort i den utveckling som sker eftersom den i mångt och mycket har kommit att domineras av personer med datavetenskaplig bakgrund.

För att försöka ge en bakgrund till var ontologier kommer in och varför detta är intressant, visas en illustration som beskriver i grova drag hur man har tänkt sig att vägen fram till den semantiska webben skall se ut. Längre fram i uppsatsen kommer en djupare genomgång av de flesta begreppen.



Figur 1. Planerad utveckling av den semantiska webben (Iselid 2001).

I princip kan man se tre huvudnivåer i utvecklingen, där den understa nivån existerar redan i den nuvarande webben. Mittennivån innehåller komponenter som existerar, men inte fått sitt stora genomslag än, möjligen med undantag av XML, eXtensible Markup Language (kommer att beskrivas i avsnitt 5.1.3) som får en allt större spridning, inte bara inom webbrelaterade områden utan också i andra användningsområden där man behöver utbyta information mellan olika system. Den översta nivån innehåller de mer komplexa koncepten för att uppnå målet att datorerna skall kunna tolka informationen. Det är på denna nivå vi återfinner ontologibegreppet, vilket alltså är till för att kunna skapa en gemensam terminologi. Man kan jämföra detta med hur det idag finns så kallade protokoll som reglerar hur datorer kommunicerar i nätverk. Detta för att göra det möjligt för alla typer av datorer att kommunicera med varandra och annan utrustning utan att för den skull använda samma operativsystem eller komma från samma tillverkare, vilket ofta var ett krav i den tidiga datorvärlden. På samma sätt tänker man sig att ontologier skall göra det möjligt att kommunicera information mellan olika domäner eller tillämpningsområden, t.ex. konstvetenskap och arkeologi, och dessutom skapa förutsättningar för att bearbeta denna information med målet att ge användaren de resultat han eller hon är ute efter.

2.2. Grunddefinition av ontologi

Mer detaljer om hur ontologier byggs upp och används kommer jag att bearbeta djupare längre fram i uppsatsen. Jag tror ändå att det kan vara på sin plats att ge en kort beskrivning om vad jag i denna uppsats använder som referensram för den fortsatta diskussionen.

Ordet ontologi kommer ursprungligen från filosofin, men i dessa sammanhang började begreppet användas inom artificiell intelligens (AI), ofta också benämnt expertsystem, för att underlätta kunskapsdelning och återanvändning mellan olika kunskaps-strukturer. En av de ledande inom området, Thomas Gruber, använder följande grunddefinition: “An ontology is a specification of a conceptualization.” (Gruber 1993b).

Man kan säga att ontologier eller det sätt som man använder ontologier, är till för att skapa ett enhetligt språk. Man har idag stora problem att dela information, eftersom olika datorsystem ofta hanterar data på olika sätt och i många fall kan det av kommersiella skäl vara av intresse att inte göra informationen lätt tillgänglig. Olika programspråk och nätverkssystem är också exempel på hinder för ett fritt informationsutbyte. För att överbrygga dessa hinder behövs standardisering på flera områden: representationsspråk, kommunikationsprotokoll samt en terminologi för beskrivning av innehåll och sammanhang, en typ av kontrollerat språk. De två första är oberoende av innehållet i den information som hanteras.

I början av nittiotalet började man studera ontologier som ett sätt att åstadkomma den eftersökta standardiseringen eller harmoniseringen i hanteringen av innehåll och sammanhang, både i kommunikationen mellan människor och datorer och mellan datorer (Gruber 1993a; Ding 2001). Detta ger också återkopplingen till de traditionella verktygen klassificering och indexering, vilka ju också handlar om att hantera

informationen baserad på dess innehåll. Kopplingen till biblioteksvärlden var dock inte så intressant i början av webbens utveckling, men den verkar nu ha blivit ett mycket viktigare ämne igen, eftersom mängden av information har blivit ohanterlig.

I denna uppsats använder jag mig av den definition som Gruber har givit. Eftersom det råder en viss oenighet om tolkningen av denna definition kommer jag dessutom att längre fram i uppsatsen gå igenom några av de alternativa eller kompletterande definitioner som förekommer.

3. Metod

3.1. Metodval

För att besvara mina frågeställningar har jag genomfört en litteraturstudie. Valet av litteraturstudie är gjort för att det för mig kändes som om det var ett bra sätt att skapa en överblick över det område som jag valt att studera. Att göra en praktisk studie och utvärdering av den semantiska webben är inte möjlig än, eftersom den inte existerar. Däremot går det att göra undersökningar av områden kring uppbyggandet men jag har valt att inte göra det utan riktat in mig på begreppet ontologi med utgångspunkt i artiklar som behandlar ontologier, så som begreppet används i diskussionen kring den semantiska webben. För den som är intresserad finns ett stort antal artiklar i litteraturen som behandlar olika metoder och verktyg som används för att bygga konstruera och utvärdera olika komponenter av den semantiska webben. Två bra startpunkter för detta är ”Spinning the Semantic Web. Bringing the World Wide Web to its full potential” och ”Towards the semantic web: ontology-driven knowledge management” (Davies et al. 2003; Fensel et al. 2003).

För att få kopplingen till biblioteks- och informationsvetenskapen tydlig har jag försökt att visa på relaterade områden inom kunskapsorganisation. Dessa områden är klassifikationsscheman och tesaurer. Jag vill se huruvida det finns en kopplingen mellan ontologier och de traditionella metoderna.

För att förklara vad jag ser som en litteraturstudie har jag använt mig av Hartmans bok ”Handledning”, där det ges en beskrivning. En litteraturstudie är när man väljer ut ett avgränsat material som är representativt för det område som skall studeras. Uppgiften, som jag som författare har, är att strukturera och sammanfatta det område som jag valt. Utifrån den frågeställning, som ställts, tas relevant material fram. Det är dock inte meningen att jag som författare skall förvränga, det material som använts, utan skapa och belysa den frågeställning som jag har. Detta kan inte jämföras med ett referat, eftersom jag som författare själv tillför en struktur och tillför ny kunskap. (Hartman 1990, s. 61)

För att besvara fråga 1 har jag använt mig av källor som behandlar ontologier i det sammanhang som jag har beskrivit ovan. Tesaurer och klassifikation anser jag ingår i utbildningen och jag har därför inte beskrivit dessa närmare, utan tagit fasta på de likheter och skillnader som finns jämfört med ontologier.

För att besvara fråga 2 används i stora drag samma litteratur som för fråga ett, men med fokus på de delar som, saknas i kontrollerad vokabulär, och vilken betydelse detta har för ontologibegreppets användbarhet.

För att besvara fråga 3 har litteratur kring ontologier i relation till den semantiska webben använts.

Jag räknar med att utifrån frågeställningarna och den litteratur jag valt kunna utveckla ett resonemang kring sambanden mellan begreppet ontologi och begrepp som tesaurer och klassifikation och då i relationen till hur det skall användas i den semantiska webben. Ur detta resonemang följer en analys av de argument som används i diskussionen kring relationen mellan de viktiga begreppen samt min egen ståndpunkt efter att jag besvarat mina frågeställningar.

För att tydligare exemplifiera sambandet mellan tesaurer och ontologier kommer jag att använda språket OIL (Ontology Inference Layer) för att föra över en tesaur till ett ontologiformat. Språket OIL ser ut att bli ett av språken som skall kunna möjliggöra den semantiska webben. Detta görs för att konkret belysa skillnader och likheter mellan tesaur och ontologier. Dessutom ger detta praktiska exempel en inblick i hur bl.a. XML och RDF vidareutvecklas till verktyg som kan användas för att webben skall utvecklas från en plattform fokuserad på presentation av information till en plattform för förståelse och bearbetning av information.

3.2. Material och informationssökning

Materialet jag har använt kommer huvudsakligen ifrån forskning kring begreppet ontologi.

För att kunna svara på frågeställningarna behövs material från tre olika kategorier:

1. Litteratur som behandlar begreppet ontologi ur ett informationstekniskt perspektiv.
2. Litteratur som ger en överblick över den semantiska webben och dess uppbyggnad.
3. Specifik litteratur som ur ett biblioteks- och informationsvetenskapligt perspektiv behandlar relationen mellan ontologier och de övriga begrepp jag anser vara relevanta.

Ontologiområdet behandlas framförallt i litteratur inom den datavetenskapliga disciplinen. Tidiga artiklar som ofta refereras till, kommer från Knowledge Systems Laboratory vid Stanford University, framförallt av Thomas Gruber. Dessa artiklar publicerades i början av 1990-talet och behandlar hur begreppet använts inom AI och expertsystem och den definition, som Gruber formulerade, är den som använts också i diskussionerna kring den semantiska webben och många andra områden. Den färskare litteraturen behandlar ett stort antal aspekter kring ontologier och hur de kan användas i olika sammanhang. Jag har här valt att fokusera på den gren som behandlar ontologier i samband med den semantiska webben, t.ex. Ying Dings artikel ”A review of Ontologies with the Semantic Web in view” (Ding 2001).

Den semantiska webben beskrivs bl.a. i en serie publikationer från W3C, utgående från Tim Berners-Lees ursprungliga presentation av visionen från 1994. Vid sidan av materialet från W3C finns en stor mängd litteratur, eftersom detta är ett aktivt forskningsområde. Det urval jag har gjort baseras på en bedömning av vilka namn som återkommer ofta i detta sammanhang och bör ha en ledande roll i utvecklingen. Ett

exempel är Dieter Fensel som publicerat ca 150 vetenskapliga artiklar inom detta och närliggande områden och dessutom är medförfattare till ”Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce” (Fensel 2001) och ”Towards the Semantic Web, Ontology-Driven Knowledge Management” (Davies et al. 2003).

För att förstå hur den semantiska webben skall byggas upp har jag också valt ut litteratur som behandlar XML och RDF eftersom dessa standarder behövs för användningen av ontologier. Det går att använda andra standarder för detta ändamål men som det ser ut idag så är det dessa som har valts ut av W3C. Exempel på relevant litteratur är Learning XML (Ray 2001) och W3C – rekommendation för XML (Bray et al. 2000).

För att få in det biblioteks- och informationsvetenskapliga perspektivet har jag framförallt valt artiklar från Journal of the American Society for Information Science and Technology (tidigare Journal of the American Society for Information Science), Journal of Information Science samt Journal of Documentation, eftersom det i dessa tidskrifter publicerats artiklar som behandlar de ämnen jag valt att bearbeta. Ledande personer i denna diskussion är Gilchrist, Soergel och Vickery. (Gilchrist 2003; Soergel 1999; Vickery 1997)

De databaser jag använt mig av är LISA, ERIC, Inspec och Library Literature & Information Science. Via dem fick jag fram både elektroniska och trycka källor med relevans för min uppsats. Biblioteket vid Högskolan i Borås har också bidragit till att finna material till uppsatsen. Genom att jag fann material via databaserna har jag sedan använt mig av artiklarnas referenslistor för att finna ny litteratur. Eftersom ämnet är så pass nytt i den användning som jag skriver om, finns det inte några monografier att tillgå, så därför har det mest blivit artiklar som jag använt mig av. Trots avsaknaden av monografier har jag inte lidit brist på litteratur, eftersom området är mycket aktivt. Snarare är det som så ofta numera att problemet är att sortera ut en rimlig mängd information för att göra arbetet hanterbart.

4. Ontologi

Detta kapitel innehåller en grundlig genomgång av ontologibegreppet, från dess ursprung i filosofin till användningen inom datavetenskapen, framförallt inom så kallade expertsystem. Jag ger här också en beskrivning av begreppets användning ur ett biblioteks- och informationsperspektiv, framförallt inom kunskapsorganisation.

4.1. Ursprung

Ordet har sitt ursprung i filosofin där det betecknar ”Läran om varande”, vilket är en del av metafysiken. Metafysiken går att härleda tillbaks till Aristoteles och hans lärjungar och ordet ontologi (ontologia) myntades på 1600-talet (Welty & Guarino 2001). I mer detalj kan man beskriva ontologi som läran om de begrepp eller kategorier som måste finnas för att man skall kunna ge en beskrivning av verkligheten, t.ex. för att kunna svara på frågan ”Vad är ett hus?”. Dessutom behöver denna beskrivning vara uttömmande, motsägelsefri och sammanhängande, vilket ställer stora krav på den. T.ex. kan man tänka sig att beskriva allt som existerar på jorden genom att börja med att dela upp i levande och icke-levande. Därifrån kan man sedan göra en allt finmaskigare uppdelning av allt som finns. (Nationalencyklopedin 2003)

Det finns också en något annan användning av ordet, nämligen följande: ”En teori som beskriver de begrepp, framförallt abstrakta sådana, som är tillåtna i ett språksystem.” (Webster's 1993) (författarens översättning).

Med denna definition i bakgrunden, är det naturligt att begreppet har kommit till användning inom områden som expertsystem, kunskapsrepresentation och på senare år även inom informationsåtervinning. Orsaken till att det blivit så intressant inom dessa områden, är att det utlovar en delad och allmän förståelse för områden som dessutom kan kommuniceras mellan människor och dataprogram. (Vickery 1997; Fensel 2001)

Begreppet ontologi har använts under en längre tid och inom ett antal områden. Sedan början av 1990-talet har forskningen kring ontologier varit mycket aktiv bl.a. inom områden som expertsystem, knowledge engineering, natural language processing och kunskapsrepresentation. Även inom information retrival, knowledge management m.m. har begreppet använts. (Ding 2001) Listan skulle med största sannolikhet bli ännu längre idag. Ett konkret exempel från biblioteksvetenskapen är Dublin Core Metadata Initiative från 1999 (McGuinness 2003, s. 174).

I de följande avsnitten beskriver jag användningen inom några av dessa olika områden.

4.2. Användning inom expertsystem

Under denna rubrik tar jag upp betydelsen av begreppet ontologi inom det område som jag nog betraktar som det ledande området vid framväxten av användandet av begreppet ontologi som det brukas idag, när det gäller ontologier på WWW. Inom

arbetet med expertsystem och artificiell intelligens, ett område som nog måste betraktas som mindre aktivt nu än för tiotalet år sedan, identifierade man i mitten av 1980-talet bristen på formella, alltså bevisbara, verktyg att använda för de system man byggde upp. Målet var att skapa så kallade expertsystem som skulle kunna underlätta för människor att fatta beslut i komplexa frågor. Man insåg då att det saknades konsekventa metoder att samla och formulera den nödvändiga information som efterfrågades. De metoder som användes var informella och ofta beroende av just den person som hade utformat systemet eller sammanställt informationen. I diskussionen kring dessa problem började man använda begreppet ontologisk analys som en metod att förbättra situationen.

I början av 1990-talet kom Thomas R. Gruber med en definition av ontologi som har kommit att ofta bli citerad. Hans definition lyder så här: ”An ontology is a specification of a conceptualization”. Vad betyder då ”conceptualization” i detta sammanhang? Jo, konceptualisering är en abstrakt, enkel syn av den värld vi av någon anledning önskar att representera. Den abstrakta synen behandlar då de företeelser vi antar existerar inom denna värld och förhållandena mellan dessa. Det går alltså att se en ontologi som en katalog över allt som finns i en värld, sambanden mellan beståndsdelarna och hur saker fungerar. Denna definition är alltså inte så väsensskild från den filosofiska, utan snarare lite av en omformulering för att passa in i ett annat användningsområde. (Gruber 1993b)

Denna definition av ontologi i detta sammanhang är dock inte oomstridd, utan det har förts en livlig debatt om detta. Bland de mest framträdande i denna diskussion har varit Guarino, som har analyserat och vidareutvecklat Grubers definition. Guarino gör detta genom att lista sju olika tolkningar av begreppet:

1. Ontologi som ett filosofiskt område
2. Ontologi som ett informellt konceptsystem
3. Ontologi som formell beskrivning av semantik
4. Ontologi som specifikation av konceptualisering
5. Ontologi som representation av ett koncept via en logisk teori (antingen baserat på formella egenskaper eller baserat på användningsområde)
6. Ontologi som vokabulär använd inom en logisk teori
7. Ontologi som en specifikation av en logisk teori

Definition 1 i listan skiljer sig mycket ifrån de övriga definitionerna och det är i huvudsak tolkning 2-7 som är av intresse för denna uppsats. Här kommer ett försök att beskriva Guarinos uppdelning. (Guarino & Giaretta 1995)

Definition 2 och 3 ser en ontologi som en semantisk konceptuell enhet som kan vara antingen informell eller formell. Vad menas då med semantisk konceptuell enhet? Ett sätt att se det, är att man skapar ett enat koncept kring en speciell betydelse av t.ex. ett ord eller ett område. För en ontologi skulle det betyda att man enats kring en betydelse av ett koncept. Skillnaden mellan 2 och 3 är just kravet på formalism. I definition 2 accepteras informella beskrivningar, jämför t.ex. med en ordbok eller WordNet, medan definition 3 kräver att beskrivningen är formell vilket gör att den är mer användbar i datorsammanhang.

Definitionerna 4-7 utgår från att ontologin behandlar de syntaktiska aspekterna, alltså vilken typ av symbolism som används. Parallellen här är naturligtvis meningsbyggnad, alltså hur man konstruerar ett språk, vilken ordning orden skall stå i o.s.v. I detta sammanhang bryr man sig mindre om ordens betydelse, utan fokus ligger på uppbyggnaden av språket eller beskrivningen. (Guarino and Giaretta 1995)

I den litteratur som finns är det mycket sällan man skiljer på dessa användningar, utan man rör sig ganska fritt mellan dem. Guarino gör ett försök att tydliggöra detta, men konstaterar samtidigt att det är svårt att komma fram till en entydig beskrivning av begreppet ontologi. Han föreslår att det borde gå att använda två kompletterande termer för att förtydliga det hela, nämligen *konceptualisering,* när man betonar de semantiska relationerna, och *ontologisk teori,* när det rör sig om den mer detaljerade beskrivningen som skall användas för att uttrycka ontologisk kunskap. (Guarino and Giaretta 1995)

Här kan man också se en skillnad mellan Grubers och Guarinos användning av begreppet ontologi och Guarino vill framförallt göra begreppet mer entydigt. I detta sammanhang finner jag ingen anledning att gräva djupare i denna diskussion. Man kan dock konstatera att detta ämne är långt ifrån slutdiskuterat och att det finns ett antal olika ståndpunkter på en ganska abstrakt teoretisk nivå. För den praktiska användbarheten är detta inte lika avgörande utan jag nöjer mig med att konstatera att debatten pågår, samtidigt som de flesta arbetar vidare med den definition de själva anser fungera för sitt ändamål.

Vilken användning är tänkt för en ontologi? Ett av de problem man brottades med, när man försökte konstruera olika typer av så kallade expertsystem (datorprogram som skulle kunna ge svar på frågor på samma sätt som när man vänder sig till en expert) var att, förutom att se till att kommunikationen mellan olika system fungerar, lösa problemet med att ”förstå”. Alltså, eftersom ett expertsystem arbetar med frågor och tar fram svar baserade på den kunskapsdatabas som systemet har tillgång till, måste det finnas specifikationer för hur denna kunskap är formulerad och arrangerad. En sådan specifikation kallas då ontologi och innefattar de termer, relationer, funktioner m.m. som man anser behövs för att beskriva ett specifikt område. För att bli praktiskt användbart krävs att man formulerar de ingående komponenterna (alltså termer, relationer osv.) på ett sådant sätt att man undviker missförstånd och tvetydigheter.

En ontologi utvecklas ofta i form av ett samarbete mellan flera experter inom ett område och man talar också om ett antal ”agenter” som kommer att använda ontologin. Dessa agenter är både människor och maskiner som samverkar inom ett specifikt område. Om en användare väljer att följa en specifik ontologi kallas detta att vara ”committed”, förbunden, till denna ontologi. Detta innebär att man kommer överens om att använda ontologin på ett konsekvent sätt.

Detta är en av de avgörande aspekterna för att en ontologi skall bli användbar för kunskapsutbyte och ett effektivt utnyttjande av den lagrade kunskapen. De agenter, som enats om att använda en ontologi, behöver inte alla sitta på samma kunskap, utan när en fråga kommer, svarar respektive agent utifrån den kunskap den har.

Användandet av en gemensam ontologi gör det möjligt att tolka informationen från samtliga agenter och sammanställa informationen på ett konsekvent sätt.

Utöver kunskapsutbytet finns det också en annan aspekt, som här brukar betonas, nämligen återanvändning. Tanken är där att man skall kunna bygga nya ontologier genom att kombinera redan existerande och därmed spara både tid och pengar. Detta ställer dock en del krav på hur ontologierna formuleras, och idealt skulle man vilja ha tillgång till många små och väldefinierade ontologier som sedan kan byggas ihop som legoklossar för att skapa nya. Som vanligt ser inte verkligheten ut så, utan det finns idag en uppsjö av olika typer av ontologier och ett antal olika sätt att beskriva dem. (Fensel 2001, s. 12f)

Några olika typer av existerande ontologier är följande (Fensel 2001, s. 12):

- *Domänontologi*. En sådan ontologi beskriver ett specifikt område, t.ex arkeologi eller hjärtkirurgi
- *Metadataontologi*. Ett exempel på detta är Dublin Core, som är till för att underlätta beskrivningen av elektroniskt tillgängliga resurser.
- *Allmän ontologi*. Precis som namnet avslöjar är tanken med denna typ av ontologi att försöka beskriva allmängiltig kunskap om t.ex. tid och rum. Denna typ av ontologi är därmed tillämpbar inom ett flertal områden vilket kan vara både en styrka och svaghet.
- *Representationsontologi*. Denna är inte heller knuten till ett specifikt område, utan beskriver generellt användbara ”representationsenheter” som går att använda i valfria områden. Om man vill veta mer om denna typ av ontologi rekommenderas att läsa om Grubers ”Frame Ontology” (Gruber 1993b).

För att göra det hela något mer gripbart kommer jag här att exemplifiera begreppet med två olika ontologier som båda har existerat en längre tid och ofta återkommer i litteraturen. Dessa två exempel är WordNet och CYC.

WordNet

WordNet (Fellbaum 1998; Gilchrist 2003) har utvecklats vid Cognitive Science Laboratory vid Princetonuniversitetet. Det är ett exempel på ett Internetbaserat referenssystem för ord från det engelska språket. De som byggt upp och utvecklat WordNet har använt sig av språkpsykologiska teorier för att lära sig hur vi som människor tänker och ordnar upp språket i vårt minne. WordNet är därmed en ontologi som beskriver det engelska språket och relationerna mellan orden i detta. Alltså kan WordNet klassificeras som en domänontologi enligt klassificeringen ovan.

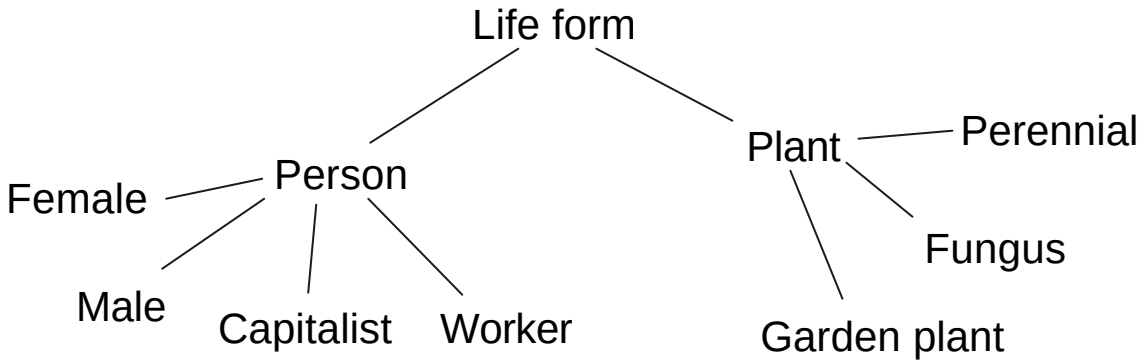
Grunden för WordNet är att orden organiseras i synonyma uppsättningar och varje uppsättning representerar ett underliggande lexikalt koncept. WordNet innehåller betydelsen av 100 000 ord och de är ordnade i en hierarkisk struktur. WordNet grupperar orden i fem huvudkategorier: substantiv, verb, adjektiv, adverb och funktionella ord. Inom varje kategori så organiseras orden i koncept (ex. ordets mening) och genom semantiska relationer mellan ord. Exempel på dessa relationer är

synonymer, antonymer, hyponymer (är-en relation). WordNet har blivit mycket populärt dels på grund av att det är tillgängligt på Internet, dels för att den är fri att använda. Det är också en ordbok som ger mer än bara en alfabetisk lista av ord. WordNet ger inga semantiska definitioner i formellt språk, utan de semantiska koncepten är definierade i naturligt språk. Detta gör att det är svårt att använda helt automatiserat, eftersom det finns utrymme för tolkningar. WordNet är ett exempel på en ontologi som är konstruerad för att användas inom ett specifikt område, i detta fall lingvistik. WordNet i sig själv är också bara avsett för det engelska språket, men det finns också en flerspråkig version som kallas EuroWordNet. (Fensel 2001, s. 14)

Jag visar ett enkelt exempel på hur WordNet är uppbyggt. I sammanhanget ”any living entity” hör följande ord (Detta blir i huvudsak på engelska eftersom WordNet är byggt för det engelska språket):

- Life form
- Organism
- Being
- Living thing

Ett av de sätt som WordNet använder för att bygga relationer mellan orden i databasen är att använda hyponymer, alltså att något är en form av något annat. T.ex. är personer och växter båda livsformer, och alltså är livsform en hyponym för personer och växter. På detta sätt bygger man upp relationer i trädstrukturer och med detta enkla exempel kan man bygga följande träd:



Figur 2. Exempel på trädstruktur i WordNet

Cyc

Den andra ontologin, som jag anser vara av vikt att ta upp, är Cyc. Initiativtagarna till Cyc kommer från en bakgrund inom AI och Cyc-programvaran har utvecklats sedan 1984. Målet har varit att göra så kallad ”sunt förnuft”-kunskap åtkomlig och användbar för dataprogram. Forskare inom AI ville få datorer att fungera på ett sådant sätt att de kunde ”tänka” som människor. Cyc började som en metod att formalisera denna kunskap från världen och förse den med semantik och formalism. Många hundra tusen

koncept har formaliserats för hand med logiska axiom och regler. På detta sätt har man byggt upp en stor databas över information som normalt betraktas som sunt förnuft (Fensel 2001, s 14f).

Cyc vet t.ex. att träd vanligtvis befinner sig utomhus, och att ett glas med vätska skall bäras upprätt osv. Cyc grupperar koncept i en övergripande ontologi, och sedan finns det djupare information som gör att det går att hålla ordning på om en viss information är tillämplig i alla sammanhang eller om det finns begränsningar i dess tillämpbarhet. Cyc är en kommersiell produktfamilj som utvecklas och distribueras av Cycorp Inc. Bland användarna av Cyc återfinns bl.a. amerikanska militären och Lycos. Att Cyc är en kommersiell produkt gör att det är svårt att få reda på mer detaljerad information om uppbyggnaden. Lite mer tillgängligt är OpenCyc.org, vilket är en fri version av Cyc. Denna innehåller bara en bråkdel av vad som finns i den kommersiella versionen, men kan ändå vara tillräckligt för att få en bild över hur systemet är uppbyggt. Dessutom kan man använda OpenCyc för att bygga upp sin egen information och dessutom dra nytta av det som redan finns. Systemet kan dessutom dra nya slutsatser baserat på kombinationer av det som fanns från början, och den information man själv lagt till. Ett exempel på slutledning, som Cyc kan göra, är att om systemet vet att person 1 visat sorg och saknad efter person 2, har person 1 levt efter att person 2 avlidit.

Till skillnad från WordNet är Cyc alltså redan från början anpassad för datoranvändning och därmed mycket formell i sin uppbyggnad.

I denna kontext är användningen av ontologier en detaljerad beskrivning för att göra ontologiska antaganden. Denna beskrivning görs lämpligen i form av ett formellt språk för att göra en automatisk hantering enklare.

4.3. Ontologi i biblioteks- och informationsperspektiv

Inom informationsvetenskap har ontologier inte använts som benämning, men sedan mitten av 90-talet har de även börjat dyka upp där. En av de första att uppmärksamma detta var Vickery (Vickery 1997). Man skulle kunna tycka att det egentligen vore mer naturligt om de hade kommit ifrån detta håll eftersom kopplingen mellan ontologier och kunskapsstrukturer som tesaurer och klassifikation är ganska uppenbar. Därför är det inte underligt att personer ifrån bibliotekshåll undrar varför de inte har blivit involverade i uppbyggnaden av ontologier, men kanske blir det ändring på detta i och med den utmaning som den semantiska webben innebär. Det har funnits förslag till att skapa plattformer för kunskapsutbytande som t.ex. SemWeb (inte att förväxlas med den semantiska webben). Detta förslag framfördes av Soergel (Soergel 1996; Vickery 1997). Förslaget till SemWeb kom också under mitten av 90-talet samtidigt som ontologier uppmärksammades allt mer.

Även inom informationsvetenskapen finns dock ett antal olika uppfattningar, bl.a. om relationen mellan tesaurer och ontologier, men också på en mer filosofisk nivå om förhållandet mellan olika ämnesområden och vetenskapliga inriktningar.

4.3.1. Några huvudaktörer i diskussionen

Dagobert Soergel

Dagobert Soergel vill i sin artikel ”The Rise of Ontologies or the Reinvention of Classification” likställa ontologier med klassifikation. Det tycks som om han drar ett likhetstecken mellan ontologier och klassifikation, och det skulle innebära att de utför samma uppgift (Soergel 1999). Detta kan nog på sätt och vis vara sant, men jag tror inte att det går att säga att en ontologi skulle vara helt detsamma som klassifikation. Jag kan dock hålla med honom om att i och med den växande informationsmängden på webben så har klassifikation blivit ett intressant forskningsområde för andra än bara inom biblioteks- och informationsvetenskap.

Ontologier är en typ av grund klassifikation som inte går på djupet utan visar på övergripande kategorier enligt Soergel. För expertsystem upptäcktes att det behövdes mer specificerade termer. Problemet var att klassifikationen var för grund, och att det behövdes terminologi i form av ord som ”leder in” till djupare förståelse av ett område/koncept. Detta gjorde att även tesaurer kommit till användning. Att man i detta sammanhang skapat ett nytt ord, alltså ontologi, för ett område som redan existerar, menar Soergel beror på en brist på kommunikation mellan de olika disciplinerna.

System som Cyc kan därför ha blivit uppbyggda med större ansträngning än vad som behövts, eftersom man inte utnyttjat redan existerande kunskap. Soergel skriver också att system som Cyc och WordNet skulle ha blivit lättare att använda om standardmetoder för tesaurer hade använts, istället för att konstruera nya. Här skulle SemWeb fungera som plattform, som skrivits tidigare, där det skulle kunna leda fram till ett allmänt tillgängligt system för olika funktioner som klassifikationsscheman, tesaurer, ordböcker och ontologier, där man lär av varandras erfarenheter (Soergel 1999). Tanken kring ett allmänt system för personer med liknande intresse är bra men det verkar tyvärr inte ha fått något genomslag.

Brian Campbell Vickery

B. C. Vickery var en av de första att uppmärksamma ordet ontologi i informations-vetenskapen. Vickery diskuterar ontologi utifrån olika synsätt inom ”knowledge engineering”. Vickery beskriver en ontologi som ett schema, som det ingår någon form av semantiska kategorier av viktiga koncept i, och dessa koncept representerar olika domäner. Han ser också en klar koppling till klassifikation och tesaurer, men skillnaden är dock den att det är tänkt för olika användningsområden.

Vickery redovisar också en del av diskussionen kring ontologiers ”granularity” eller ”grain size” – alltså till vilken grad ett koncepts hierarki skall fortsätta att dela upp sig, hur finfördelat skall det vara. Här finns i huvudsak två inriktningar, där Guarino företräder den riktning som hävdar att man måste ha en hög granularitet för att en ontologi skall bli praktiskt användbar. Den andra inriktningen, representerad bl.a. av Roberto Poli, betonar den övergripande användningen av en ontologi (Poli 1996).

Det finns enligt Vickery en konflikt mellan de olika grupper som arbetar med uppbyggandet av ontologier. Konflikten finns mellan dem som vill generalisera eller specificera ontologier. Forskare, som vill generalisera en ontologi skall kunna beskriva all kunskap från ett kompetensområde, medan de som vill specificera en ontologi, vill avgränsa den till en specifik domän/område. Här kan vi se en koppling till generell och speciell klassifikation inom biblioteks- och informationsvetenskap (Vickery 1997).

Oavsett detta är det ont om referenser från ”knowledge engineering” till biblioteks- och informationsvetenskap. Konsekvensen av detta är att man riskerar gå miste om mycket kunskap som kunde ha varit till hjälp. Även vad det gäller kopplingen till ontologier för syftet att vara en form av ordlistor, finns mycket få referenser till den kunskap som finns hos dem som arbetat med studier kring semantiska relationer. Det Vickery kommer fram till, är att problemet med semantisk analys i informations-processen för dem som utvecklar ontologier, är samma problem som man arbetat med under lång tid och som man fortfarande arbetar med inom informationsvetenskap, alltså hur man hanterar semantik på ett effektivt sätt och hur man gör för att kunna hantera detta maskinellt och minimera behovet av manuell analys (Vickery 1997).

De två olika konflikter eller diskussioner, som beskrivits ovan påverkar också de metoder som finns för att ta fram ontologier. Det är uppenbart att detta är ett område, som inte har hunnit mogna än, eftersom det finns ett antal olika metoder beskrivna i litteraturen. I vissa fall skiljer sig metoderna ganska mycket, medan det i andra fall endast är mindre skillnader. Denna flora av metoder ger också upphov till en antal olika verktyg som utvecklats på olika håll i världen, både från akademiskt håll och kommersiellt.

Alan Gilchrist

Alan Gilchrist har försökt att bringa klarhet kring orden tesaur, taxonomi och ontologier genom att analysera vilket användningsområde respektive begrepp eller term har. Jag har dock inte för avsikt att i denna uppsats närmare gå in på skillnaden mellan taxonomi och klassifikation, men det bör nämnas att man inom vissa områden sätter likhetstecken mellan taxonomi och klassifikation (Gilchrist 2003).

Taxonomi är ett begrepp som dyker upp då och då i dessa sammanhang och betyder ursprungligen vetenskapen om organismernas klassificering, dvs. beskrivning, namngivning (nomenklatur) och formell klassifikation av organismgrupper som taxonomiska enheter. I informationsvetenskapliga sammanhang har begreppet fått en något vidare betydelse och används t.ex. för att beskriva de tekniker som är grundläggande för att hantera automatisk indexering av webbplatser och skapandet av ämneskataloger, både för WWW och för t.ex. företagsinterna nätverk. Gilchrist ser i sin undersökning att det används både klassifikations- och tesaurtekniker i samband med taxonomier. (Gilchrist 2003)

Med den betydelse av taxonomi som Gilchrist exemplifierar ser han taxonomi som lite av ett mellansteg mellan tesaurer och ontologier. Han påpekar också att förekomsten

av de tre begrepp han valt att analysera, beror på vem som valt att använda respektive begrepp. Något förenklat ser han att det innebär följande uppdelning:

- Tesaur – används av informationsvetare inom kunskaporganisation.
- Taxonomi – används av systemvetare och programvaruutvecklare, framförallt i kommersiella tillämpningar för WWW och informationshantering för företag.
- Ontologi – har anpassats av datavetare för användning ursprungligen inom AI och expertsystem och på senare tid framförallt för den semantiska webben. (Gilchrist 2003)

En av Gilchrists slutsatser baserad på vad som beskrivits ovan, är att det finns en hel del överlapp mellan områdena, men detta uppmärksammas inte alltid, eftersom de som arbetar med dessa områden kommer från olika discipliner. I många fall pratar man bredvid varandra eftersom man använder olika termer för att beskriva samma eller liknande företeelser.

Han har dock förhoppningen att den semantiska webben kan vara ett projekt som är så stort att man mer eller mindre tvingas se sig om efter expertis och kunskap utanför det område man själv är aktiv inom.

4.3.2. Ontologi i relation till ämnesanalys och domänanalys

En annan aspekt som är intressant i detta sammanhang är om det finns en relation mellan de i biblioteks- och informationsvetenskapliga sammanhang förkommande begreppen ämnesanalys och domänanalys.

Ämnesanalys – Subject Analysis

Ämnesanalys innebär att vid klassificering av material bortse från det klassifikationssystem som används, och istället fokusera på att göra en analys av innehållet. Grundläggande för ämnesanalysbegreppet inom biblioteks- och informationsvetenskap är D.W. Langridges arbete från 1989 (Langridge 1989). Grundtanken i detta är att all information har en relation till redan existerande kunskap, åtminstone på någon nivå. En ämnesanalys görs stegvis (Langridge 1989, s. 136):

- Vilken kunskapsform handlar det om?
- Vilken disciplin inom denna kunskapsform?
- Vilket ämne?
- Vilken dokumentform?

Utgående från denna information gör man sedan en uppsummering och går till klassifikationssystemet. Langridge förutsätter här att det finns en viss uppsättning av kunskapsformer som kan anses vara permanenta och han listar följande (Langridge 1989, s. 33ff):

- Prolegomena (ex logik och tänkande)
- Filosofi (filosofi, etik etc)
- Naturvetenskap (fysik, kemi etc)
- Teknologi (materialteknik, elektronik etc)

- Humaniora (psykologi, sociologi etc)
- Samhällsvetenskap (utbildning, hälsovård etc)
- Historia (arkeologi, biografi etc)
- Moral
- Religion
- Konst
- Kritik (konstkritik, litteraturkritik etc)
- Personlig erfarenhet

Dessa är alltså hans toppnivå, vilken han anser täcker in all form av kunskap. Detta har naturligtvis ifrågasatts, inte minst hans uttalande om att detta är permanenta kategorier. Vi kan lämna denna debatt här, eftersom den inte påverkar relationen med ontologibegreppet.

Målet med ämnesanalysen är alltså att göra en utvärdering av ett givet dokument eller annan typ av information för att på så vis få fram de centrala aspekterna och därmed kunna göra en bättre klassificering.

Vilken är då relationen till ontologi? På samma sätt som med hjälp av ontologier försöker man i ämnesanalys representera verkligheten eller en del av denna genom att dela upp denna i olika områden med undernivåer. Naturligtvis finns också skillnader, framförallt eftersom ämnesanalysen innehåller just en analys, dvs. försöker värdera det analyserade dokumentet utifrån då ovan nämnda kriterierna.

Domänanalys – Domain Analysis

Domänanalys existerar som ämnesområde inom flera discipliner, bl.a. datorvetenskap där det används för att underlätta återanvändning av programvara. Domänanalysen används för att förstå hur olika system inom ett område fungerar, vad som är gemensamt och vad som skiljer dem åt. Framförallt fokuseras på att med hjälp av domänanalys kunna återanvända designprinciper och metoder, snarare än specifika block av programkod. (Arango 1994)

Inom biblioteks- och informationsvetenskap har domänanalys framförallt lyfts fram av Birger Hjørland (Hjørland & Albrechtsen 1995). Han har påpekat det som han ser som brister i tidigare forskning, nämligen att fokus legat för mycket på användarna av informationen och inte på sammanhanget i vilket informationen existerar. Jag går inte in på domänanalysen som sådan här, utan fokuserar på dess relation till ontologi. Det är nämligen så att det finns likheter mellan dessa företeelser. Inom domänanalysen diskuteras bl.a. behovet av att kunna skapa specifika klassifikationer. Många gånger är de generella klassifikationssystem som finns just för generella för att effektivt kunna klassificera ett speciellt område. Hjørland påpekar i detta sammanhang att man däremot inom datavetenskapen kommit längre inom detta område, men då använder begreppet ontologi för att beskriva aktiviteten att skapa en beskrivning av ett område. Framförallt liknar detta användningen av ontologi i betydelsen domänontologi (se kap. 4.2) (Hjørland 2002)

Hjørland påpekar också att skapandet av tesaurer också kan ses som en gren av domänanalysen, men att det finns mycket kvar att göra. T.ex. när det gäller att hitta nya metoder som gör det möjligt att automatisera processen för att ta fram en tesaur för ett område.

4.3.3. Två forskningsexempel

När man skall skapa en ontologi finns ett antal olika tillvägagångssätt, t.ex. kan man börja från scratch, kombinera existerande ontologier, utgå från någon form av kontrollerad vokabulär etc. För att exemplifiera i detta sammanhang har jag valt två studier där man utgår från en kontrollerad vokabulär respektive en tesaur för att skapa sina ontologier. Det senare arbetet ligger också till grund för det praktiska exempel som beskrivs i kapitel 6.

GEM

I studien ”Converting a controlled vocabulary into an ontology: the case of GEM” använder sig Qin och Paling av en kontrollerad vokabulär som heter Gateway to Educational Materials (GEM). GEM är ett initiativ från Department of Education’s National Library of Education i USA. Detta är till för att hjälpa lärare och personer som arbetar med utbildning att finna undervisningsmaterial. Det finns flera tusen metadatadokument i GEM-databasen och resurserna kommer ifrån flera olika områden som har med utbildning att göra. Det finns ett gränssnitt, som ger dem som katalogiserar möjlighet att använda sig av Dublin Cores 15 metadata element, och det finns 8 lokala element, som är gjorda speciellt för GEM. Det verktyg, som användes i Qin och Palings studie, var Ontolingua. Ontolingua är utvecklat vid Stanfords Knowledge Systems Laboratory och är ett webbaserat verktyg för att skapa och bearbeta ontologier. Verktyget är baserat på det språk Gruber tog fram för detta ändamål i början av nittiotalet och gör det möjligt att utnyttja ett bibliotek av ontologier för att skapa nya för det område man själv är intresserad av. Denna studie visar dels på fördelen med att kunna föra ihop redan existerande kunskapsstrukturer och dels på de problem man kan ställas inför. I studien fann man att GEM var en bra och väldefinierad startpunkt att bygga en ontologi från, men att det också krävs ett förhållandevis stort manuellt arbete för att komplettera med det som saknas i form av relationer mellan termer och en nerbrytning på en mer detaljerad nivå än vad som ursprungligen fanns. Det avgörande värdet ligger i att det går att göra en djupare semantik för att beskriva digitala objekt, både när det gäller begreppsmässigt och sambandsmässigt. (Qin & Paling 2001)

AAT

Ett annat exempel är ”From Thesaurus to Ontology”. Här använder sig B.J. Wielinga *et al*. av en Art and Architecture Thesaurus (AAT). Det man vill konstruera är en kunskapsrik beskrivning över konstobjekt på ett sådant sätt att denna beskrivning går att använda för den semantiska webben. De undersöker också vilka problem som är

relaterade till att hitta bakgrundskunskap för konstresurser. Man använder sig av en existerande tesaur, AAT, och bygger på den med djupare semantik och relationer så att det blir en ontologi (Wielinga et al. 2001). På samma sätt som i GEM-exemplet fann man att tesauren är en bra utgångspunkt för att generera en ontologi, men att det också krävs kompletteringar, t.ex. mer detaljerad information om de koncept man väljer att basera ontologin på samt uppbyggnad av de mer komplexa relationer som man har utrymme för i en ontologi jämfört med en tesaur. I artikeln påpekas också vikten av att vara noggrann i sitt val av tesaur för att inte i onödan försvåra konverteringen. Ett antal krav ställs på den tesaur man skall använda för att underlätta konverteringen:

1. Strikt hierarki, dvs. entydig koppling mellan objekt på övergripande och underliggande nivåer.
2. Beskrivande termer skall kunna kopplas till en specifik del av tesauren. Ett exempel: om ett objekt kan vara gjort av olika material, skall det finnas en del av tesauren som beskriver materialen.
3. Antalet möjliga värden för en egenskap bör vara begränsat och någon form av kontrollerad vokabulär är att föredra framför naturligt språk. (Wielinga et al. 2001)

Wielinga *et al* diskuterar också kring de olika språk och verktyg som finns för att beskriva ontologier och finner att det i dagsläget (2001 för deras del) inte finns något givet svar på detta, utan de olika alternativen har för- och nackdelar. De konstaterar dock i detta sammanhang att det är en fördel om den tesaur, man tänkt sig att använda, finns beskriven i något standardiserad språk, t.ex. XML eller RDF. Också när det gäller att föra in kompletterande information, gäller det att i möjligaste mån använda etablerade standarder för att göra ontologin användbar för andra. T.ex. kan Dublin Core eller standarder baserade på denna användas när man behöver komplettera en tesaur. (Wielinga et al. 2001)

5. Ontologi och den semantiska webben

Mitt ursprungliga intresse för det område denna uppsats beskriver, kom när jag kom i kontakt med Tim Berners-Lees vision om den semantiska webben. Det har också visat sig att en av de allra kraftigaste drivkrafterna bakom forskningen och utvecklingen av ontologier idag just är den semantiska webben och de krav den ställer på hantering av information. Detta kapitel går därför in i viss detalj på hur den semantiska webben är tänkt att fungera för att kunna beskriva och analysera behovet av ontologier och de krav som ställs på dessa. Det finns även andra områden som ser ett behov av ontologier, framförallt inom det som brukar kallas e-business, t.ex. hur företag kan sköta inköp elektroniskt och undvika mycket pappershantering. Detta innebär att leverantörernas kataloger görs tillgängliga elektroniskt, vilket ofta innebär stora informationsmängder. De krav, som ställs på ontologierna är dock likartade, så jag anser att den semantiska webben är en fullt tillräcklig referensram.

5.1. Den semantiska webben och dess komponenter

För att förstå betydelsen av ontologibegreppet så har det för mig varit viktigt att förankra det i ett sammanhang. Dels för att det gör det lättare att avgränsa uppsatsarbetet, dels för att det gör det lättare att beskriva begreppet genom att dra paralleller till närliggande områden.

Som redan nämnts i inledningen växer både antalet webbsidor och antalet användare av Internet och framförallt WWW. Även om det idag ser ut som om antalet användare inte ökar lika drastiskt längre, så fortsätter den kraftiga ökningen av antalet tillgängliga resurser på WWW. Denna utveckling måste ses som en tydlig bekräftelse på hur lyckade idéerna kring WWW och t.ex. HTML-språket har varit, inte minst med tanke på de relativt enkla verktyg, som kan användas för att vem som helst skall kunna göra i princip vad som helst tillgängligt för alla på WWW.

Just enkelheten i användandet av WWW är också det som skapar problemen med växtvärk. Det blir allt svårare att inte bara hitta den information man söker, utan också extrahera och tolka den funna informationen. Dessutom ställer det allt större krav på hur man gör sin information tillgänglig på webben, om man vill få den synliggjord via sök- och indexeringstjänster. Ett annat område, som kräver resurser, är att hålla informationen aktuell, något som inte minst gäller alla de företagsinterna intranät som existerar idag, parallellt med det öppna Internet.

Naturligtvis har denna utveckling inte undgått de personer och organisationer som ligger bakom webben som vi ser den idag, utan detta uppmärksammades redan i mitten av 1990-talet av t.ex. Tim-Berners Lee, av många ansedd som grundaren av World Wide Web. Vid den allra första konferensen om WWW (International World Wide Web Conference, CERN, Geneve) 1994 gav han uttryck för idéerna om att den då ganska nyligen skapade webben behövde kompletteras för att underlätta för datorer att förstå den information som tillgängliggjorts på WWW, alltså ett första uttryck för behovet av semantik på WWW (Berners-Lee 1994; Cailliau 1995). Han uttryckte då

behovet av att gå från den, visserligen revolutionerande idén att skapa hyperlänkar mellan olika typer av resurser och göra detta lättillgängligt via ett grafiskt gränssnitt, till ett system där datorerna inte bara fanns till för att lagra data, upprätthålla länkarna och presentera informationen utan också gavs möjlighet att skapa sammanhang och bearbeta information. Något senare uttryckte han utmaningen kring vad som då hade börjat kallas den semantiska webben på följande sätt:

> Utmaningen för den semantiska webben, är att tillhandahålla ett språk som uttrycker både data och regler för resonemang om data, och som tillåter regler från vilket som helst existerande kunskapsrepresentationssystem att användas på webben. (Berners-Lee et al. 2001) (författarens översättning)

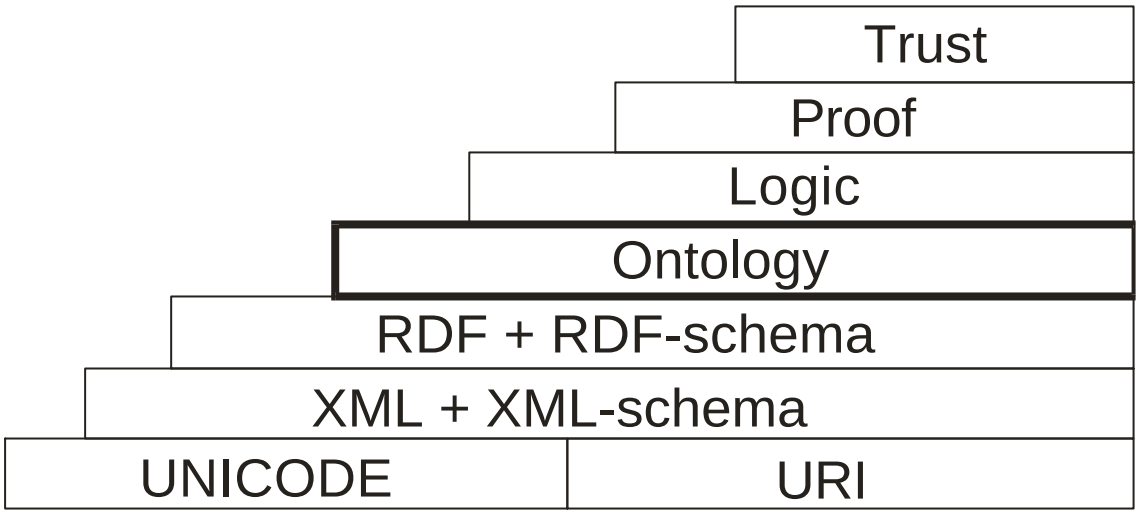
Den semantiska webben är alltså inte en separat webb utan en utvidgning av den redan existerande, med den skillnaden att informationen ges en väldefinierad mening. Människor och datorer ska kunna arbeta tillsammans på ett sätt som gör att man får fram information som är relevant för det man söker efter på ett bättre sätt (Berners-Lee et al. 2001).

Spindeln i nätet i detta sammanhang är W3C, World Wide Web Consortium, vilket bildades 1994. W3C ser som sin uppgift att vara ett öppet forum som leder den tekniska utvecklingen av WWW. Man har inom W3C satt upp tre punkter som mål för sitt arbete med WWW:

1. *Universal Access*. Att se till att webben är tillgänglig för alla, oavsett språk, utbildning, teknisk utrustning etc.
2. *Semantic Web*. Att utveckla programvara och programvarumiljöer som gör att varje användare kan utnyttja resurserna på webben på bästa sätt.
3. *Web of Trust*. Att kunna hantera de nya legala, kommersiella och sociala aspekter som uppkommer på grund av utvecklingen av WWW.

Arbetet leder framförallt fram till ett antal tekniska specifikationer för webbens infrastruktur. Till dags dato har W3C publicerat mer än 50 sådana specifikationer (W3C 2003b).

Den semantiska webben består av ett antal komponenter, eller lager, där man dels utnyttjar ett antal existerande tekniker och dels tänker sig att komplettera med ett antal nya. En förenklad modell över denna arkitektur ser ut så här (ontologilagret markerat med fet ram):



Figur 3. Schematisk bild över uppbyggnaden av den semantiska webben.

De fyra understa komponenterna är de som existerar idag och dessa beskrivs nedan. Jag kommer även att skriva lite kort om de översta lagren (Logic, Proof och Trust). Dessa lager befinner sig dock i idéstadiet ännu. Här bör också påpekas att det inte är tvunget att använda XML som beskrivningsspråk, men i dagsläget är detta den bästa kandidaten.

5.1.1. Unicode

För att datorer skall kunna hantera olika tecken krävs att de är representerade på ett sådant sätt att datorer kan tyda det. Datorer kan inte lagra tecken och bokstäver direkt, utan använder sig av olika tal/nummer för var och en av tecknen. Det finns flera olika system som kodar tecken till siffror, t.ex. ASCII. ASCII är en etablerad standard som funnits länge, men den har också vissa brister angående specialtecken mm., eftersom den i första hand togs fram för att koda de tecken som används i det engelska språket. Unicode däremot har tagits fram med målet att ge varje tecken ett unikt nummer, oavsett plattform, oavsett program, oavsett språk och standarden klarar idag att hantera 95221 tecken. Dessa tecken kommer dels från världens alfabet och dels från samlingar av andra symboler, t.ex. för kartor. Unicode har idag vunnit gehör inom de allra flesta programvaruföretag och andra stora spelare i sammanhanget. Unicode ses därmed som det bästa alternativet, när det gäller att välja en standard för teckenkodning för universellt användbara system, vilket naturligtvis både dagens WWW och den semantiska webben måste räknas som. Unicode ligger därmed till grund för användandet av t.ex. uppmärkningsspråk som XML och andra standarder som förutsätter att man kan transportera textbaserad data genom ett antal olika system utan att informationen förstörs eller förvanskas. (Unicode 2003)

5.1.2. URI – Uniform Resource Identifier

Internet är bl.a. baserat på att alla de datorer och andra resurser, som bygger upp nätet, har unika adresser och det finns olika typer av sådana. Så kallade nodnamn och IP-nummer adresserar en dator (kan se ut så här: 119.127.45.0), som i sin tur kan ha servrar för många olika tjänster. För att adressera en tjänst eller en del av en tjänst - ett dokument, en bild, ett program, en kataloguppgift - behövs en mer detaljerad adressering. En sådan utvecklades för webben, men gjordes så generell att den också går att använda till många andra tjänster på Internet - sådana som finns nu och sådana som utvecklas i framtiden. Samlingsnamnet för detta är URI, Uniform Resource Identifier**.** (W3C 2002)

Det vanligaste exemplet på URI idag är URL, Uniform Resource Locator, den adress man anger för att hitta en webbsida, oftast på följande form: http://www.rsv.se/index.html. Om man granskar denna adress närmare, ser man att den byggs upp av flera beståndsdelar:

- ”http://” visar att det så kallade http-protokollet (hypertext transport protocol) skall användas, vilket visar att det rör sig om en webbsida.
- ”www.rsv.se” är adressen till servern som webbsidan finns på.
- ”/index.html” Detta är söksträngen till den specifika sida man söker.

Som synes är uppbyggnaden av en URL ganska så generell och detta är hemligheten med URI:er. Allt som finns på webben går att beskriva genom att använda enkla grundblock som sedan sätts ihop. Detta är därmed en viktig aspekt för den semantiska webben, eftersom man där behöver ett verktyg för att beskriva allt som finns. Annars går det ju inte att utbyta information eller resonera om dessa företeelser.

Det finns dock vissa problem med att URI är ett så generellt begrepp, vi kan ge en URI till vad som helst t.ex. ett dokument på en server, en bok i en bokhylla eller en myra som kryper i landet. Det finns ingen som bestämmer över vad som kan ha eller inte ha en URI och det finns ingen som äger en URI, vilket gör att vem som helst kan skapa en URI till vad som helst. Detta gör att det till slut kan finnas många olika URI:er om samma koncept. I förlängningen betyder det att man inte kan vara säker på vad som menas med just den URI:n.

Alltså: en URI är inte ett sätt att hitta en resurs på Internet, men kan vara det. Däremot är det en beskrivning av resursen, ett namn på den. Resursen som sådan kan vara tillgänglig via Internet, men behöver inte vara det. En URI kan ge en beskrivning av var resursen finns, men behöver inte göra det.

För den semantiska webben är URI ett viktigt verktyg, men som synes finns en del oklarheter att reda ut för att säkerställa att den information, som ges i en URI, är användbar. Mer information om vad som händer inom detta område finns bland annat att finna hos W3C. (W3C 2002)

5.1.3. XML - Extensible Markup Language

XML är ett av de avgörande stegen som behöver tas för att den semantiska webben skall bli verklighet och en viktig underliggande komponent för ontologier.

XML utvecklades av XML Working Group, tidigare känd som ”the SGML Editorial Review Board”, och den gruppen bildades av the World Wide Web Consortium (W3C) 1996 (Bray et al. 2000). Detta visar XML:s ursprung i den tidigare standarden, SGML (Standardized General Markup Language), som i sin tur har utvecklats från Generalized Markup Language (GML), vilket utvecklades i början av 1970-talet för att kunna hantera dokument i den växande databehandlingsvärlden. Grundtanken är att införa strikta regler för hur dokument struktureras och därmed göra det enklare att hantera ökande mängder elektroniska dokument i den ökande automatiseringen av kontoren som satte fart framförallt under 80 och 90-talen. SGML spred sig t.ex. snabbt inom bl.a. EU:s administration och inom stora företag som IBM.

Ett problem med SGML är dock att standarden är mycket generell, vilket bl.a. gör att det är svårt att implementera programvaror för att skapa och hantera SGML-dokument. I andra änden av skalan återfinner vi en annan ML-teknik, nämligen HTML (HyperText Markup Language), vilken är baserad på en mycket mer begränsad uppsättning märkord. XML kan därmed sägas ha skapats för att fylla hålet mellan HTML och SGML.

XML är en uppsättning regler för hur man skapar ett uppmärkningsspråk men det går också att använda XML som det är för att strukturera data. Det är ett kodsystem som används för att representera information om vissa aspekter i elektroniska dokument. Med hjälp av XML kan man skapa uppsättningar av märkord som inte bara beskriver hur ett dokument skall se ut när det presenteras utan också beskriver vad informationen handlar om. I och med detta görs informationen sökbar i ett informationssystem (Bryan 1997).

En annan fördel med XML är att det är applikations- och plattformsoberoende. Med det menas att samma XML-dokument går att skapa, spara och läsa på olika operativsystem, och med olika programvaror från olika tillverkare.

Till skillnad från HTML är XML inte begränsat till en fast uppsättning elementtyper. Man kan med XML definiera sina egna element eller egenskaper, som man sedan kan använda i ett dokument (Ray 2001). XML ger därmed möjlighet att särskilja mellan t.ex. presentation och innehåll, där HTML i jämförelse endast beskriver presentationen, dvs om ett ord skall skrivas med ett speciellt typsnitt, på en speciell plats på sidan eller med en viss färg. Ett enkelt XML-exempel kan se ut så här:

<namn> Iréne </namn>.

Jag har definierat elementet med <namn> inom taggarna. Mottagaren och sändaren vet då att elementet betyder namn i detta sammanhang (Ray 2001, s. 36). Detta låter enkelt men riktigt så enkelt är det inte, eftersom det också måste finnas en programvara som

förstår ordet *namn* för att det skall fungera. För att kunna hantera detta har man använt dokumentmallar, så kallade DTD:er (Document Type Definition), vilka beskriver de attribut som finns och hur de får användas. Om man använde samma DTD för flera dokument underlättar man informationsåtervinning radikalt. En DTD kan antingen vara inbakad i XML-filen eller hanteras som ett separat dokument. Det finns idag ett antal fria DTD:er tillgängliga för alla (Bryan 1997).

För den semantiska webben har man valt att använda en vidareutveckling av DTD, nämligen XML-schema. Anledningen till detta är att DTD:er är begränsade. Det går t.ex. inte att särskilja dokument med samma namn men med olika skapare. Det finns heller ingen möjlighet att ange att innehållet i ett element ska vara av en specifik typ, t.ex. ett nummer, ett datum eller en valuta. Med XML-schema kan man göra detta och en annan fördel är att XML-schema till skillnad från DTD följer XML:s syntax, så man behöver bara arbeta med en syntax. Denna flexibilitet erhålls dock på bekostnad av DTDs korthet; om man anger samma regler i XML-schema och DTD blir DTD-varianten ofta betydligt kompaktare (W3C 2003c; Ray 2001).

5.1.4. RDF - Resource Description Framework

RDF är en annan komponent som man arbetar med för att den semantiska webben skall kunna fungera. RDF togs fram för att övervinna de problem som fanns med existerande metadatasystem som HTML META och Dublin Core. HTML META är förhållandevis spritt, men begränsat vad det gäller möjligheterna att beskriva dokument. Dublin Core är å andra sidan ett betydligt mer kompetent system, men har haft dåligt genomslag på WWW. RDF skapar ett ramverk för behandling av metadata vilket betyder att det går att kombinera olika typer av metadatasystem. Det begrepp som används för metadatasystem inom RDF är *schema*.

Eftersom datamängden på WWW är så enorm, är det helt orimligt att utföra manuell insamling och indexering annat än för en bråkdel av informationen. RDF är tänkt som ett led i arbetet att få datorprogram att ”förstå” sammanhang och därmed kunna utbyta information med varandra. En grundtanke med RDF är att möjliggöra automatiserad behandling av metadata (Kronman & Parnefjord 1999; Lassila & Swick 1999). I detta sammanhang är det också viktigt att skapa en mekanism som inte är bunden till en speciell applikationsdomän utan är områdesneutral.

Förutom ovan nämnda områdesneutralitet krävs också att RDF-beskrivningarna, för att vara maskinellt hanterbara, är mycket tydligt standardiserade utan att lämna utrymme för tolkningar. RDF definierar tre nivåer (Kronman and Parnefjord 1999):

1. *Struktur* – RDF datamodell
2. *Syntax* – RDF syntax, uttryckt i XML
3. *Semantik* – RDF schema, t. ex. Dublin Core

RDF – datamodell

RDFs datamodell består av tre grundläggande komponenter:

$$\left.\begin{array}{c} Resurser \\ Egenskaper \\ Värde \end{array}\right\} \rightarrow Uttryck$$

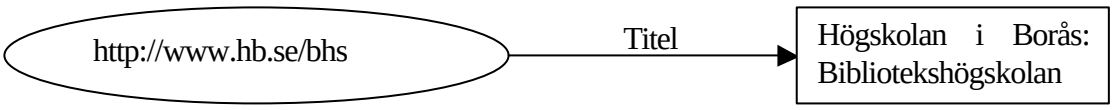
Resurser är identifierade som unika. Allting som beskrivs med RDF-uttryck kallas för resurser. En resurs kan vara en webbsida, en del av en webbsida, men kan även vara en tryckt bok. Resurser namnges alltid med en URI (se avsnitt 5.1.2). Ett vanligt exempel på URI är en URL:er (Uniform Resource Locator) vilka används t.ex. för att beskriva var webbsidor finns.

En URL består av tre delar:

1. Den metod som används för att nå resursen
2. Namnet på den dator där resursen finns
3. Namnet på resursen inklusive eventuell sökväg

Resurserna har ett antal *egenskaper* som kännetecknas av en aspekt, karaktär, attribut, eller en relation som man använt för att beskriva en resurs. Titel kan vara en sådan egenskap. Varje egenskap har sedan ett *värde* (Kronman and Parnefjord 1999).

Dessa tre komponenter: resurs, egenskap och värde bildar tillsammans ett *uttryck*. Här visas ett exempel på RDF-modellen:



Figur 4. Exempel på ett RDF uttryck

RDF – syntax

För att uttrycka RDF behövs en syntax som kan uttrycka och byta metadata. W3C använder sig av XML. XML beskriver endast innehållet och strukturen men talar inte om hur det skall se ut. Alltså kan man använda sig av samma XML-kod även om innehållet presenteras på olika sätt. Datamodellen i RDF uttrycks oftast med en seriell syntax: inledande <tagg> och avslutande </tagg> (Lassila and Swick 1999).

Det finns också en förkortad syntax med endast en inledande <tagg/>. Denna används bl.a. för att infoga RDF i html-sidor (Lassila and Swick 1999).

När man använder sig av XML kan RDF utnyttja funktionen *namnspace*, vilket gör att det går att använda sig av *scheman*. Exempel på scheman är ordlistor, klassifikations-system, tesaurer.

XML gör att man som användare kan skapa en bra struktur åt sitt dokument men säger inget om vad som menas med strukturen. Meningen står istället RDF för. RDF använder sig av något som kallas för ”trippels”. Varje trippel kan jämföras med ämne, verb och objekt för en grundläggande mening (Berners-Lee et al. 2001).

RDF – schema

När vi skriver eller förmedlar oss i vardagen, vill vi att vår omgivning skall förstå vad vi uttrycker. Det är därför viktigt att både skrivaren och läsaren har samma förståelse för det uttryckta. I detta sammanhang gäller det att det hela är mycket tydligt, eftersom WWW finns i ett globalt perspektiv och uttryck som t.ex. ”författare” måste vara mycket precist definierat så man vet precis vad som menas.

Meningen i RDF uttrycks just genom ett *schema* (W3C 2003a). Man kan tänka sig ett schema som en slags ordbok. Schemat definierar termer, som skall användas i ett RDF-uttryck, och det ger en specifik mening åt dem. Som jag skrivit tidigare så kan man använda olika slags scheman och Dublin Core är ett exempel.

Dublin Core är en metadatastandard som skapades vid ett möte mellan bibliotekarier och Internet-experter i Dublin, Ohio, USA år 1995. Hittills har man haft sex internationella träffar för att etablera standarden. Dublin Core består av 15 dataelement som t.ex. title, subject, description och source. Dublin Core är ett sätt att katalogisera, där upphovsmannen själv kan beskriva innehållet utifrån de 15 dataelementen. Målet har varit att Dublin Core metadata skall möta varierande behov hos experter inom olika discipliner, som har med olika typer av Internetinformation att göra. Önskemålen har varit enkelhet i förhållande till att skapa och upprätthålla metadata, gemensamt förstådd semantik, tillämpning efter existerande och kommande standard. Än så länge har inte standarden fått sitt stora genombrott, delvis beroende på att mängden information, som användarna måsta mata in, mer eller mindre manuellt ökar, när man vill komplettera t.ex. sin webbsida med metadata baserade på Dublin Core. Däremot finns förhoppningar om ett ökat användande i och med att man tar steget till nästa nivå som RDF, där Dublin Core passar bra in för att bygga upp RDF scheman (Hedberg 2001; DCMI 2003).

I RDF införs semantik med hjälp av hänvisningar till scheman. Ett sätt att uttrycka semantisk är t.ex att använda sig av Dublin Cores innebörd av ”title”. RDF scheman ärver egenskaper från de metadatasystem man väljer, t.ex. definitionerna av respektive egenskap men också begränsningar i användningen.

För att kunna skilja mellan olika element, som kommer ifrån olika scheman eller metadatasystem finns i RDF en identifierare som talar om vilket. Så skrivs t. ex. ”dc” för att visa när Dublin Core används.

RDF och Topic Maps

I detta sammanhang bör också nämnas Topic Maps, vilket är ett annat sätt att skapa associationer mellan objekt. Topic Maps och RDF liknar varandra på många sätt, men det finns också skillnader i hur de tekniskt bygger upp associationerna. Topic Maps har sitt ursprung i klassiska register och ordlistor och har därmed ett mer ”mänskligt” betraktningssätt, medan RDF härstammar från matematik och logik och alltså ett mer maskinmässigt förhållande till den information som skall beskrivas. Jag kommer inte att gå in på detta i detalj, eftersom det inte är av intresse för uppsatsen i stort. Jag nöjer mig med att konstatera att dessa två tekniker samexisterar och av en del bedömare anses komma att slås samman på sikt. I och med skapandet av RDF scheman går det redan idag att förhållandevis enkelt utbyta information som strukturerats med de två olika metoderna (Pepper 2002).

5.1.5. Logic

Lagret, som benämns logik, är ett av de lager som än så länge inte existerar mer än på konceptnivå och beskrivningen blir därför ganska kortfattad. Jag anser ändå att det är av värde att beskriva dessa koncept, eftersom det är relevant för hur ontologibegreppet passar in och behövs för att uppnå de mål man satt upp för den semantiska webben.

Logiklagret bygger delvis på att man utnyttjar slutledningsförmåga (inference), alltså att man kan generera ny information från den information eller datamängd man redan har. Ett exempel på detta är att datorn kan utnyttja att den får reda på vad som krävs för att ett objekt skall ges beskrivningen ”hund”. Systemet kan då själv gå igenom en datamängd och lista ut om det finns fler hundar i denna. På samma sätt underlättar detta hanteringen av information på flera språk, då man genom slutledning kan få systemet att förstå att allt som i en tysk databas beskrivs med ”farbe” motsvarar ”colour” i en engelsk.

Detta är ett av de områden som arbetas med idag, men gäller framförallt att konstruera språk som gör denna typ av logiska resonemang möjliga för datorsystem. Detta är också ett av de områden som debatteras flitigt och där det finns en hel del olika åsikter om detta är genomförbart eller inte och hur långt man i så fall kan komma (Berners-Lee et al. 2001).

5.1.6. Proof och Trust

Om logiklagret finns på plats, kommer det som en naturlig fortsättning att försöka bevisa riktigheten i den information systemet hittat eller resonerat sig fram till. Med tanke på hur det ser ut på WWW idag, är det knappast svårt att föreställa sig att det är lätt att hamna i situationer, där en källa hävdar att ”havet är blått” och en annan som lika fast försäkrar att ”havet är grönt”. Vilken källa eller information skall då systemet lita på och presentera för användaren?

Två av de aspekter, som är tänkta att användas i detta sammanhang, är kontext och digitala signaturer. Kontext innebär att man bygger upp en ”erfarenhetsdatas” av informationskällor, som man har erfarenhet från tidigare och därmed anser sig kunna lita på. Ett exempel från verkliga livet är om man har vänner med liknande smak för musik och film. Om någon av dem rekommenderar en film man själv inte sett, litar man mer på denna information än om det är en okänd person som hävdar detta.

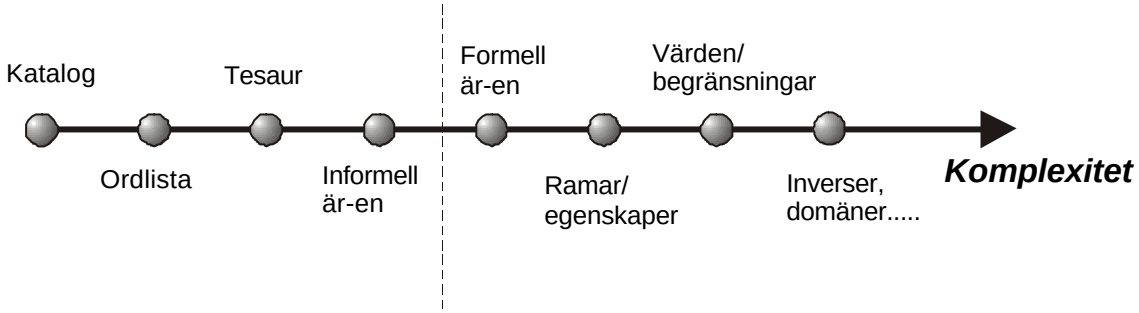
Digitala signaturer är egentligen datorvärldens motsvarighet till namnteckningar, och det finns redan idag system för detta. Ett populärt sådant är PGP (Pretty Good Privacy), som bygger på krypterade nycklar, unika för en specifik användare. Inom W3C finns idag en arbetsgrupp som definierar hur digitala signaturer skall implementeras i XML (Berners-Lee et al. 2001; W3C 2003d).

5.2. Varför behövs ontologier för den semantiska webben?

För att den semantiska webben skall bli verklighet är ontologier en mycket viktig del av uppbyggnaden. I detta kapitel tar jag upp ontologier i ljuset av den semantiska webben lite mer ingående.

Den ökade mängden information på t.ex. WWW gör att ontologier behöver ett allt mer uttrycksfullt och detaljerat språk för att kunna bli effektiv vid sökningar. Om ontologier skall fungera för den mängd information som finns på WWW, krävs att ontologierna är mycket väldefinierade. Problemet med en semantisk analys i informationsprocessen är välkänt inom informationsvetenskapen sedan länge, men nu har det blivit uppmärksammat inom andra områden i och med utvecklandet av ontologier (Vickery 1997).

McGuinness skriver om ontologier baserat på de behov som finns i utvecklingen av WWW och med den vision som Tim Bernes-Lee har genom den semantiska webben (McGuinness 2003, s. 171f). Bland annat tar hon upp aspekten, att ett helt spektrum av ontologier måste finnas, eftersom det finns olika behov av detaljering eller granularitet i olika sammanhang. För att det skall fungera så menar Fensel med flera, att det behövs många små ontologier, som kan kopplas ihop varandra, för att kunna användas om och om igen. När forskare har utvecklat ontologier så har tanken varit, att det skall finnas ett inre sammanhang och en begränsad mängd av växelverkan mellan modulerna. (Fensel 2001, s. 12f).



Figur 5. Spektrum av ontologier. (McGuinness 2003, s. 175)

Längst till vänster på skalan finns de enklaste formerna av ontologi med början i kontrollerad vokabulär, t.ex. kataloger och ordlistor. Ordlistor innehåller en viss semantisk information, men denna är oftast uttryckt i naturligt språk och därmed svår att hantera maskinellt.

Nästa steg längs skalan tar oss till tesaurer, där man har tydligare relationer än i ordlistor, men ofta saknar en tydlig hierarkisk struktur. Om man har en tydligare hierarki med klasser och underklasser, kommer man ett steg till på skalan. Ett exempel här är de strukturer som finns i ämneskataloger på WWW, t.ex. Yahoo! och Lycos. Ännu ett steg tas om denna hierarki är strikt uppbyggd så att om B hör till gruppen A, så är allt som är undergrupper till B också en del av A (McGuinness 2003, s. 175f).

Vi har med detta steg passerat en viktig punkt på skalan, eftersom man från och med de strikta klass/subklass-hierarkierna kan börja dra slutledningar, vilket är en viktig aspekt för ontologier. Fortsätter man åt höger kommer allt mer sofistikerade relationer och kompletterande information, bl.a. så införs egenskaper i och med de rambaserade strukturerna. Dessa egenskaper ärvs genom hierarkin automatiskt till subklasserna. Dessutom kan man koppla värden till egenskaperna och sätta begränsningar på dessa (t.ex. att en varas pris måste ligga inom ett visst intervall för att uppfylla ett speciellt kriterium). Allt mer komplexa samband byggs ju längre man kommer på skalan, men det viktiga här är att försöka visualisera den ökande komplexiteten.

McGuinness vill dra en skiljelinje vid den strikta hierarkin (formell är-en relation) och definiera detta som ett minimikrav för begreppet ontologi. Detta är dock en något kontroversiell fråga och det finns ett antal olika uppfattningar om detta. Vissa menar att kataloger, ordlistor och tesaurer är ontologier, medan andra som McGuinness säger att det behövs en tydlig hierarkisk uppbyggnad innan det kan kallas för en ontologi. (McGuinness 2003, s. 175f)

McGuinness ger också några riktlinjer på vad hon tycker att en ontologi behöver innehålla. För att kalla det för en enkel ontologi så behövs (McGuinness 2003, s. 177f):

- begränsad kontrollerad vokabulär (utökningsbart)
- entydig tolkning av klasser och relationer mellan termer

- hierarkisk klasstruktur

Det finns också andra egenskaper som kan nämnas, t.ex. begränsningar i vilka värden som får förekomma inom en klass, men dessa är inte nödvändiga för att uppfylla kraven på en enkel ontologi (McGuinness 2003, s. 178).

Varför behövs då ontologier för den semantiska webben? Ontologier ger tillgång till betydligt mer kraftfull hantering av information vid sökningar på WWW. Detta eftersom en ontologi, till skillnad från enklare strukturer, ger tillfälle för datorer att istället för att i huvudsak analysera antalet förekomster av ett ord eller kombinationer av ord, utnyttja de relationer som finns definierade i ontologin. Detta gör det möjligt att t.ex. hitta information om ett ämne även om denna inte är beskriven med exakt de termer som användaren använt i sin sökning. Genom att utnyttja ontologin kan systemet också hitta information om vad användaren söker även om denna är publicerad på ett annat språk.

En annan fördel är att genom att utnyttja relationer och begränsningar angivna i ontologin, så kan systemet sortera bort information som använder samma ord som användaren angivit som sökord, men i andra sammanhang. Detta blir alltså ett medel för att öka precisionen i sökningen.

Ontologier som bas för Logic, Proof och Trust

Kopplingen mellan logiklagret och ontologibegreppet är tydlig, eftersom ett av grundbegreppen för logiken är att kunna använda resonemang, alltså att systemet kan dra slutledningar från den existerande informationen och därmed generera ny information. Detta är en av de egenskaper som ontologierna tillför utöver det som finns i t.ex. RDF Schema. Proof bygger vidare på logiklagret i och med att det är logiken som används för att bygga upp bevisföringen. Proof är därmed också beroende av att ontologilagret existerar.

Begreppet Trust är mer svårgripbart och kräver också mer arbete för att fungera i praktiken. Ett exempel på hur ontologier och de andra underliggande nivåerna kan hjälpa till i detta sammanhang är något som kallas FOAF (Friend Of A Friend). Tanken med detta är att skapa ett ramverk för en användare att beskriva sig själv och de personer hon eller han har någon form av relation till. Med detta som grund går det sedan att bygga en struktur i vilken man också kan ange i vilken grad man litar på uppgifter som lämnats av respektive identifierad individ eller källa. (Golbeck et al. 2003)

6. Från tesaur till ontologi – ett exempel

Med detta exempel vill jag visa att det går att konvertera en tesaur till en ontologi, men också analysera de skillnader som finns mellan dessa två sätt att hantera information. Det finns ett stort antal tesaurer att välja på för denna typ av utvärdering och det finns också ett antal sätt att beskriva ontologier. Den tesaur jag valt att använda kallas AAT (Art & Architecture Thesaurus) och den är fritt tillgänglig via Internet (AAT 2000).

Idéerna till det praktiska exemplet kommer delvis från Wielinga *et al* (Wielinga et al. 2001), men med följande skillnader: jag har valt en annan del av AAT och beskriver ontologin med hjälp av ett annat språk. Dessutom anser jag att det är lättare att göra en tydligare analys av skillnaderna mellan tesaur och ontologi om man själv har gått igenom de olika stegen. Syftet är alltså inte att vidareutveckla det som Wielinga och andra gjort, eftersom detta är minst en magisteruppsats i sig själv.

6.1. Ontologispråk

För att en ontologi skall kunna användas och bearbetas av datorer krävs att den är beskriven med hjälp av ett fördefinierat språk. De språk som finns för närvarande är antingen logik-baserade (first-order), ram-baserade (frame logic) eller webb-baserade (RDF, XML, HTML). Ett språk, som använder sig av fördelarna från dessa tre områden, blir ett bra språk för att representera en ontologi. Det finns flera sådana språk, men de språk, som jag tittar närmare på, är OIL och DAML + OIL, eftersom dessa språk ser ut att bli de språk som kommer att användas vid skapandet av ontologier till den semantiska webben. Nästa steg i utvecklingen kallas OWL (Web Ontology Language) och är baserat just på kombinationen DAML + OIL. Jag kommer inte att gå in på dessa i detalj, men för att förstå hur ontologier skall användas i den semantiska webben så behövs det förståelse för hur dessa språk används. (Ding 2001)

OIL (Ontology Inference Layer eller Ontology Interchange Language, båda namnen används) är ett standardspråk framtaget under EU-projektet On-To-Knowledge (On-To-Knowledge 2002). Som nämns ovan är detta inte det första försöket att skapa ett språk för att beskriva ontologier, men här har man försökt dra nytta av erfarenheterna från användningen av de tidigare. Ett exempel är Ontolingua, vilket har använts under tiotalet år och det finns ett antal ontologier skapade med hjälp av Ontolingua tillgängliga i en databas (Ontolingua 2002). Detta språk är mycket flexibelt, men detta faktum, tillsammans med avsaknad av stöd för resonemang, gör det också svåranvänt. OIL utnyttjar erfarenheterna från WWW och startar från en liten mängd väldefinierade webbstandarder.

En relevant frågeställning i detta sammanhang är vad som skiljer t.ex. OIL från RDF Schema och varför det inte räcker med XML Schema eller RDF Schema för att beskriva ontologier. Sanningen är den att XML Schema och framförallt RDF Schema räcker till för att skapa en beskrivning av ett område, alltså något som liknar en ontologi. RDF Schema innehåller bl.a. klass/underklasstrukturer, men är i jämförelse med fullfjädrade ontologispråk begränsat. Några av skillnaderna mellan OIL och RDF

Schema är att i OIL har man tillgång till betydligt fler möjligheter att beskriva detaljer, ställa krav på att värden måste ligga inom specifika intervall, kombinera krav med booleska operatorer (and, or, not) osv. Den mest avgörande skillnaden är dock stödet för resonemang som finns i OIL. I RDF Schema kan vissa uttryck lämna öppning för tolkning och alltså förlorar man formalismen, bevisbarheten, som man eftersträvar i en ontologi. (Klein et al. 2003, s. 118f)

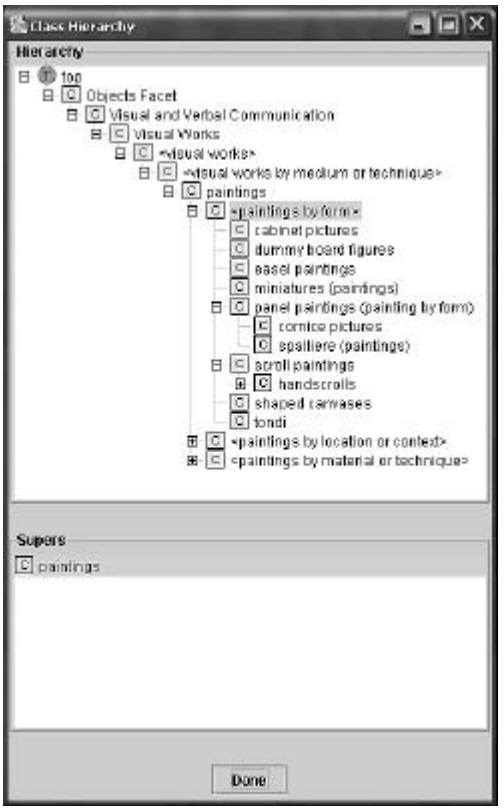
Det bör dock noteras att det finns en tydlig relation mellan XML Schema, RDF Schema och ett flertal av ontologispråken, t.ex. kan man betrakta OIL som ett tillägg till RDF Schema på samma sätt som RDF Schema bygger vidare på RDF. Ett OIL-baserat dokument kan till viss del tolkas av ett system som kan hantera RDF Schema, vad som händer är att de OIL-specifika delarna ignoreras och resten betraktas som ett ”vanligt” RDF-dokument. (Klein et al. 2003, s. 120)

6.2. Konvertering från tesaur till ontologi

6.2.1. Steg 1. AAT – Art & Architecture Thesaurus

AAT är en av de mest omfattande tesaurerna som beskriver konstverk och den är dessutom uppbyggd i en strikt hierarki, vilket underlättar när man skall skapa en ontologi. AAT innehåller mer än 125 000 termer för att beskriva många olika aspekter av konstverk och arkitektur från antiken till nutid. Jag har valt att fokusera på ett litet avsnitt av denna stora tesaur, nämligen ”paintings” och framförallt ”paintings by form”. I den vänstra delen av figuren nedan ses hur termerna är strukturerade strikt hierarkiskt med hjälp av BT och NT-relationer. ”Visual Works” är en Narrower Term, NT, i förhållande till ”Visual and Verbal Communication”.

Top of the AAT hierarchies
.... Objects Facet
..... Visual and Verbal Communication
..... Visual Works
..... <visual works>
..... <visual works by medium or technique>
..... ➡ paintings
..... <paintings by form>
..... cabinet pictures
..... dummy board figures
..... easel paintings
..... miniatures (paintings)
..... panel paintings (painting by form)
..... scroll paintings
..... shaped canvases
..... tondi
..... <paintings by location or context>
..... fore-edge paintings [N]
..... illuminations
..... mummy portraits
..... mural paintings
..... quadrature
..... quadri riportati
..... rock paintings
..... vase paintings
..... <paintings by material or technique>
..... acrylic paintings



Figur 6. *Till vänster ses hierarkin i AAT-tesauren och till höger är detta överfört till en ontologistomme med hjälp av verktyget OilEd.*

6.2.2. Steg 2. Komplettering av innehållet

I figur 6 ovan visas hur grundstommen av AAT har flyttats över till OIL med hjälp av OilEd. Än så länge har egentligen inte så mycket hänt. Innehållet är det samma och ingen kompletterande information har vare sig lagts till eller kunnat extraheras. För att göra ontologin mer användbar behövs alltså tillägg utöver det som finns i tesauren. Vad är då detta? T.ex. finns för ”cabinet pictures” följande information i tesauren:

cabinet pictures (<paintings by form>, paintings, ... Visual and Verbal Communication)

Note: Small easel paintings of the late 17th to 19th century, intended for hanging in small rooms and viewing at close range.

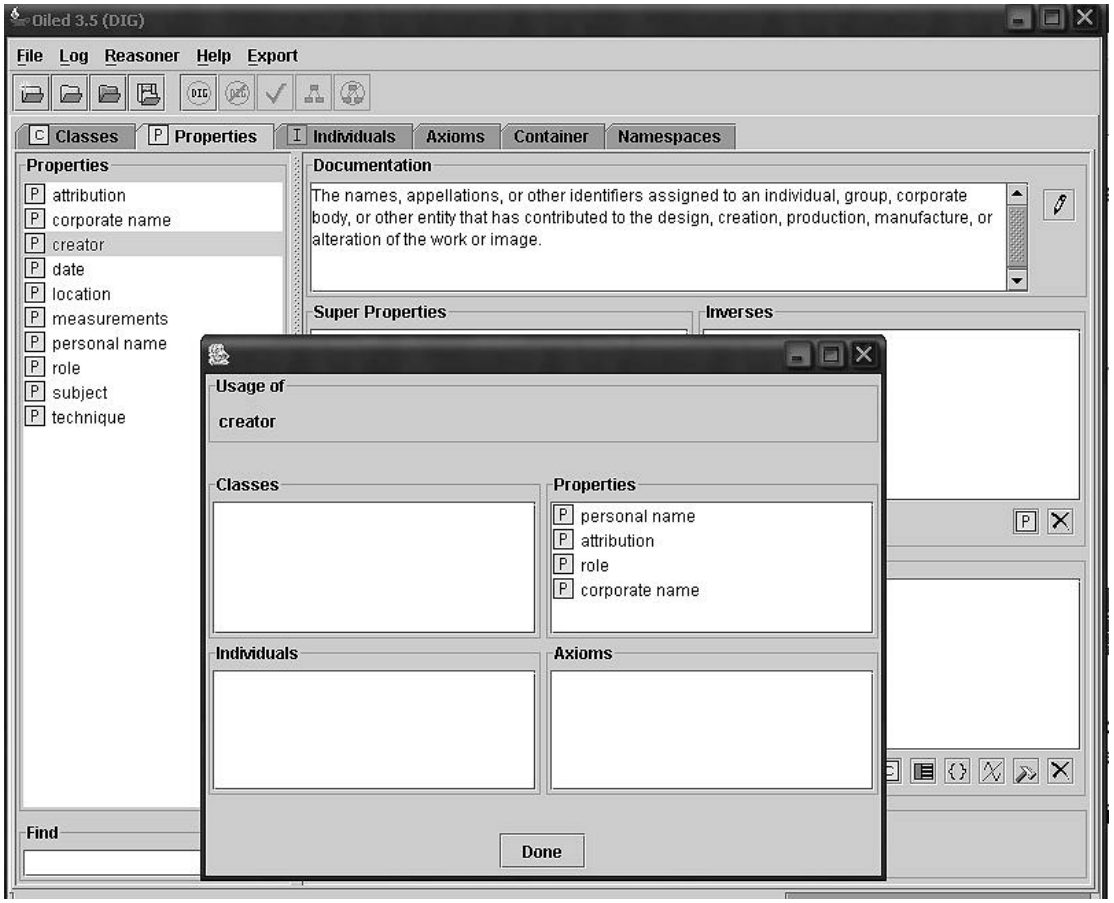
Terms:

cabinet pictures (**preferred**, C,U,D,English, American-P)
cabinet picture (C,U,AD,English, American)
cabinet painting (C,U,UF,English, American)
cabinet paintings (C,U,UF,English, American)
cabinet-picture (C,U,UF,English, American)
cabinet-pictures (C,U,UF,English, American)
painting, cabinet (C,U,UF,English, American)
paintings, cabinet (C,U,UF,English, American)
picture, cabinet (C,U,UF,English, American)
pictures, cabinet (C,U,UF,English, American)

Här finns synonymer som visar hur ingången till termen ser ut, men inte mycket mer. Det finns till exempel ingen möjlighet att göra kopplingar till vilken teknik som använts eller motiv.

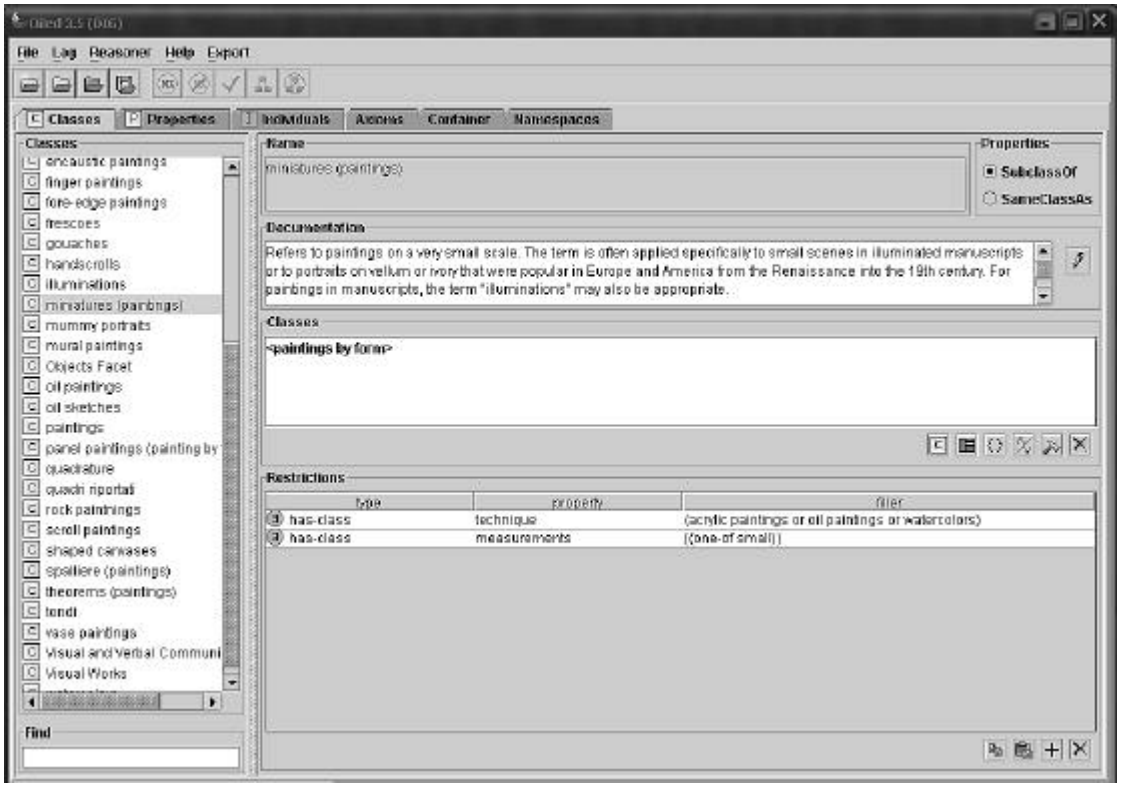
Nästa steg är att komplettera den information som finns i AAT och jag har valt att göra detta med hjälp av VRA Core Categories version 3.0, som är framtaget av Visual Resources Association Data Standards Committee. VRA Core Categories är baserat på Dublin Core-standarden och är avsett att användas för att beskriva konstverk och avbildningar av sådana. Den delen av tesauren jag valt att använda handlar om målningar därför fungerade VRA bra som utbyggnad. Naturligtvis skulle det gå bra rent principiellt att använda egna tillägg, men som grundprincip bör man alltid i möjligaste mån använda så etablerade standarder som möjligt. Detta för att öka sannolikheten att så många som möjligt finner ontologin användbar.

VRA innehåller 17 beskrivande element och jag har lagt in några av dessa i exemplet nedan. Dessutom kan elementen ha underliggande ”qualifiers”, t.ex. har man under ”Creator” följande undernivåer: ”Creator.Role, Creator.Attribution, Creator.Personal name, Creator Corporate name”.



Figur 7. Komplettering av AAT-tesauren med hjälp av element från VRA Core Categories.

Detta innebär att vi har börjat få möjlighet att få ut mer information om målningarna, i det här fallet skapare. I och med detta kan systemet nu leta reda på konstverk av samma konstnär om användaren är intresserad av detta. Ett annat exempel är att börja koppla andra aspekter till varandra, t.ex. att en miniatyrmålning också tillhör någon av klasserna som beskriver typ av målning, i exemplet nedan antingen olja, vattenfärg eller gouache (kan vara andra, men dessa har jag valt i exemplet). Dessutom har miniatyrmålningarna klassats beroende på deras storlek och det krävs att de har attributet ”small” för egenskapen ”measurements”.



Figur 8. Beskrivningen av ”miniatures”.

I OilEd kan man sedan välja att exportera informationen i ett antal olika format för användning i andra program, t.ex. som RDF Schema eller HTML. Exemplet med miniatyrer ovan ser ut så här exporterat till HTML och visat i en vanlig webbläsare:

---

class **miniatures (paintings)**

namespace:
file:/C:/Program/OilEd3.5-win/ontologies/paintings.daml#
documentation:
Refers to paintings on a very small scale. The term is often applied specifically to small scenes in illuminated manuscripts or to portraits on vellum or ivory that were popular in Europe and America from the Renaissance into the 19th century. For paintings in manuscripts, the term "illuminations" may also be appropriate.
type:
primitive
superclasses:

constraints:
restriction technique has-class (acrylic paintings or oil paintings or watercolors)
restriction measurements has-class one-of (small)

Figur 9. Informationen om miniatyrer exporterad som HTML

6.2.3. Steg 3. Fortsättning till fullständig ontologi

Det går sedan att fortsätta att komplettera informationen i den ontologi jag skapat och dessutom finns ett stort antal relationer och länkar som kan skapas. Jag har valt att inte gräva mig ner mer i teknikaliteterna kring detta, utan nöjer mig med att visa grunderna i detta arbete. Som visats ovan ger OIL-språket och ontologin som sådan möjlighet att tillföra information, som saknas i tesauren, och framförallt finns ett utrymme att skapa betydligt fler och mer komplicerade relationer och länkar mellan elementen. Detta skapar bland annat möjligheter till slutledning, dvs. det går att generera kompletterande information ur den redan existerande.

En annan fråga man kan ställa sig i detta sammanhang, är när man skall anse att en ontologi är fullständig. Det finns förmodligen inget svar på denna fråga, eller så finns det lika många svar som det finns ontologiskapare. Storlek och granularitet hos ontologier är nämligen ett av de ämnen som diskuteras flitigt i litteraturen och där det finns ett antal olika ståndpunkter.

6.3. Slutsatser

Vilka slutsatser kan då dras från detta korta exempel? I första steget ser man tydligt en del av de likheter som finns, t.ex. att man skapar en hierarki för att gruppera information om ett specifikt ämnesområde. I mitt exempel med AAT finns också en strikt hierarki med övergripande och underliggande nivåer. I andra tesaurer är denna hierarki inte lika tydlig, vilket gör det mer komplicerat att bygga en ontologi.

Nästa steg börjar peka på de konceptuella skillnader som finns. Här syns hur tesauren är förhållandevis grund i sin uppbyggnad. Det finns ingen koppling mellan de olika typer av målningar som listas och vilken teknik som använts eller vilken tidsperiod som är aktuell. Detta är kompletterande information som gör en automatiserad sökprocess mer effektiv och användbar.

Ett annat exempel på de begränsningar som finns i strukturen för tesaurer som uppmärksammas är att det inte finns något stöd för att göra slutledningar på det sätt som kan göras med ontologier. Tack vare att AAT kompletterats med elementen från VRA, kan ett söksystem som använder ontologin utnyttja att det definierats ett krav för miniatyrer att ha ett speciellt värde vad gäller storlek. Alltså kan sökningar som inte innehåller sökordet ”miniature”, men har storleksangivelse i detta intervall dra slutledningen att användaren söker information om miniatyrer och därmed snabbt sålla fram denna. Denna typ av koppling finns alltså inte i tesauren, utan måste läggas till. Detta har jag gjort för hand för en mycket liten del av tesauren, och det är lätt att se vilket arbete det skulle vara att göra detta för hela AAT.

Detta exempel belyser alltså en del av de skillnader som finns mellan tesaurer och ontologier. En av de mest framträdande är att vi ser hur det i tesauren saknas verktyg för att skapa de otvetydiga relationer och beskrivningar som krävs för att maskiner skall kunna bearbeta informationen på ett effektivt sätt. Tesauren kan innehålla

beskrivande text som kan användas av människor för tolkning, vilket också hittills har varit den huvudsakliga användningen av tesaurer. Detta innebär dock att det finns ett antal begränsningar i möjligheten att effektiv bearbeta informationen i datorer och det är här steget till ontologier är avgörande.

Några avslutande kommentarer: som påpekats ovan har konverteringen gjorts manuellt och hur detta skall kunna automatiseras är ett område som studeras aktivt. Det finns flera anledningar till detta, men framförallt finns ett behov av att enkelt kunna utbyta information mellan tesaurer och dessutom måste man kunna underhålla dem på ett effektivt sätt. Detta får också till följd att vi kan hoppas att tesaurer framöver blir lättare att beskriva i maskinläsbara format, vilket också skulle underlätta konvertering och komplettering för att skapa ontologier. Ett annat faktum som bör tas i beaktande är att valet av ontologispråk, i mitt fall OIL, påverkar strukturen på den resulterande ontologin och hur komplexa relationer som kan skapas.

7. Slutdiskussion

I detta kapitel kommer jag att gå igenom frågeställningarna utifrån den bakgrund som lagts i de tidigare kapitlen. På detta följer en allmän diskussion utgående från svaren på frågeställningarna och en blick framåt.

7.1. Frågeställningarna

Vilken relation finns mellan ontologi, tesaur och klassifikationsscheman?

Relationen mellan ontologi, tesaur och klassifikationsscheman kan tyckas ganska uppenbar vid en första anblick, speciellt relationen mellan tesaur och ontologi, där vissa till och med vill sätta likhetstecken dem emellan. En närmare analys av begreppen ger dock följande resultat:

Klassifikation och klassifikationsscheman är det äldsta av dessa och har använts av mänskligheten under lång tid. Klassifikationsscheman går också att se som ett första steg på väg mot en ontologi, t.ex. så som McGuinness beskriver (McGuinness 2003, s. 175f).

Tesaurer tillför i förhållande till klassifikationsscheman möjligheten att hänvisa, vilket skapar en första nivå av kopplingar mellan de ingående termerna. En tesaur befinner sig därmed ett steg längre fram på komplexitetsskalan. Tabellen nedan ger en översiktlig bild över relationen mellan de tre begreppen:

	Klassifikations-schema	**Tesaur**	**Ontologi**
Huvudsaklig användning	Organisation av böcker och annat informationsbärande material	Beskrivning av innehåll	Specifikation av hur ett ämnesområde (av någon form) skall beskrivas
Semantiskt innehåll	Inget	Begränsat	Stort
Ursprung	Kunskaps-organisation	Kunskaps-organisation	AI och expertsystem

Tabell 1. Schematisk jämförelse mellan klassifikationsschema, tesaur och ontologi.

Utgående ifrån ovanstående resultat ser man att det finns tydliga kopplingar mellan klassifikationsscheman, tesaurer och ontologier. Analysen av diskussionen kring detta ger till resultat några olika huvudlinjer:

1. Klassifikationsscheman, tesaurer och ontologier är alla olika aspekter av samma sak och vilken term som används beror på sammanhanget. Med denna ståndpunkt skulle i princip samtliga kunna benämnas ontologier.
2. Klassifikationsscheman, tesaurer och ontologier är olika steg på en komplexitetsskala och man måste särskilja dem beroende på deras semantiska innehåll.
3. Ontologi är bara ett nytt namn för klassifikation.

Den dominerande ståndpunkten i litteraturen får anses vara nummer 2 och är också den som jag själv anser passa bäst i sammanhanget. Om man intar ståndpunkt 1 går det visserligen att framföra goda argument för detta, men samtidigt hamnar man då i en situation där man måste införa ännu fler klasser eller former av ontologier för att beskriva vad de skall användas till. Ett problem med ståndpunkt 2 är var man skall dra skiljelinjen. När skall en specifikation av en konceptualisering betraktas som en ontologi? Ett förslag, som ser ut att ha ett förhållandevis stort stöd, är det som McGuinness presenterar, se figur 5, alltså att det krävs en formell är-en relation definierad för att man skall betrakta det som en ontologi (McGuinness 2003, s. 176). Ståndpunkt 3 är inte så vanlig. Den kan dock användas för att argumentera för att de, som arbetat med ontologier under de senaste tio åren, har gått miste om existerande kunskap då man inte insett att det man försökt åstadkomma redan till viss del existerat, men inom andra vetenskapliga områden (Soergel 1996; Vickery 1997).

Sammanfattningsvis alltså: Relationen mellan klassifikationsscheman, tesaurer och ontologier är framförallt att de samtliga används för att organisera information. Skillnaderna ligger i användningsområde och därmed kraven som ställs på hur komplext systemet måste byggas, framförallt vad det gäller det semantiska innehållet, dvs sammanhang och relationer mellan termer. Ur denna synvinkel är klassifikations-scheman längst ner på skalan. Även om de kan vara nog så komplexa i sin uppbyggnad så är de ändå mycket begränsade vad det gäller att ge information om samband mellan termer som inte har en strikt klass/underklassförhållande. Tesaurerna tillför denna möjlighet med sina definitioner av t.ex. relaterade termer. Ontologier tar ett steg till, vilket kommer att analyseras mer i detalj under nästa frågeställning.

Tillför ontologi något utöver det som finns i klassifikationsscheman och tesaurer?

Som belysts ovan, så är den huvudsakliga uppfattningen att ontologier ger användaren något mer än vad som finns i klassifikationsscheman och tesaurer, och det är detta jag utgår från vid analysen av denna frågeställning.

Vad är det då som en ontologi ger? För att finna svaret på detta kan man angripa frågan från olika håll. Antingen fokuserar man på den ökande komplexitet, som anges som en skiljelinje i ovanstående frågeställning, eller så betraktar man frågan utifrån varför ontologier har kommit till, dvs. det tänkta användningsområdet. Jag finner det mest naturligt att se på frågan ur det senare perspektivet, eftersom detta mer eller mindre av sig självt också ger förklaring till det krav på ökande komplexitet som ställs på en ontologi.

Målet med ontologier är att underlätta utbytet av information, vare sig detta rör sig om information tillgänglig för alla på Internet eller hantering av de stora informationsmängder, som finns i företagsinterna nätverk eller andra kunskapsbaser, såsom expertsystem och databaser över vetenskapliga tidskrifter. Ontologierna har målet att definiera hur informationen skall beskrivas. Grunden till detta behov är att det ofta finns stora informationsmängder tillgängliga, men dessa är ofta strukturerade och beskrivna på ett för just detta system eller denna organisation unikt sätt. För att kunna dra nytta av existerande information behövs alltså en struktur på en övergripande nivå som styr hur man beskriver ett område. Det handlar också om att kunna kombinera kunskap som finns i mindre områden till ett större. (Vickery 1997)

Det finns naturligtvis inte ett givet sätt att göra detta, utan ett antal olika lösningar kan anses vara likvärdiga. Hur löser man då detta? Jo, genom att välja ett alternativ intygar man att man avser att hålla sig till detta, vilket gör att andra användare vet det och kan rätta sig efter det vid utbyte av information.

I det som skrivits ovan finns egentligen inte så mycket som tyder på att en ontologi är något mer än en tesaur, eftersom en tesaur också handlar om att skapa enhetliga terminologier och praktiskt användbara hierarkiska strukturer för att beskriva information. Det finns dock ett antal skillnader, både i uppbyggnaden och i tillämpning. Jag väljer här att fokusera på skillnaderna mellan tesaurer och ontologier, eftersom dessa allmänt ses som mer närbesläktade än klassifikationsscheman och ontologier.

Både tesaurer och ontologier har relationer som en viktig beståndsdel, men en viktig skillnad mellan dem ligger i hur relationerna är definierade. I en tesaur är de relationer som finns ganska få och inte formellt definierade, vilket innebär dels att det finns stora begränsningar i vilka relationer som går att beskriva och dels att det lätt skapas tvetydigheter. RT (related to) får ofta användas för alla typer av relationer som inte går att beskriva med NT eller BT, t.ex. närbesläktade termer och olika typer av egenskaper. I en ontologi finns stöd för en stor flora av relationer, exakt vilka beror på vilket språk man väljer att implementera sin ontologi i, men generellt sett finns betydligt större möjligheter än i en tesaur. En annan skillnad vad gäller relationerna är att dessa i en ontologi måste vara formellt definierade, vilket gör att det inte finns utrymme för tvetydigheter. Detta är också en mycket viktig faktor, när det handlar om att kunna hantera detta automatiskt, eftersom alla oklarheter och tolkningsmöjligheter kraftigt begränsar möjligheterna för datorer att bearbeta informationen. (Wielinga et al. 2001; Ding & Foo 2002)

En annan avgörande skillnad ligger i tillämpningen av tesaurer respektive ontologier. En tesaur har som mål att beskriva relationerna mellan *termer*, medan ontologier hanterar *koncept*. I sin mest renodlade form skall ju en ontologi vara oberoende av vilka termer som används eller vilket språk som skall användas. Ontologier ligger därmed på en abstraktionsnivå över tesaurer, vilket också gör det hela mer svårgripbart. Naturligtvis måste de koncept ontologin beskriver också formuleras i någon form av termer för att göra det möjligt för oss människor att förstå vad en ontologi beskriver och kunna formulera nya eller utökningar av existerande. Någon form av koppling mellan termer och koncept behövs därmed när man arbetar med

ontologier, och det är bl.a. detta stöd man får med de verktyg som utvecklats. Detta är emellertid en abstraktionsnivå som helt saknas i tesaurer och vilket därmed utgör en avgörande skillnad. (Ding & Foo 2002)

Sammanfattningsvis finner jag alltså att det finns ett antal aspekter som gör att ontologier skiljer sig från tesaurer, framförallt de formella definitionerna av relationer och möjligheten att hantera koncept i tillägg till termer. Detta gör att en ontologi blir mer kraftfull som stöd för t.ex. ett söksystem, eftersom det går att analysera en frågeställning med hjälp av de koncept som finns i ontologin. Tack vare de formellt definierade relationerna kan systemet därmed dra slutsatser om vad som är relevant för användaren och sortera bort mycket av det som inte har med frågeställningen att göra, men som skulle ha fångats upp av ett tesaursystem p.g.a. osäkerheten och tvetydigheten i relationerna.

I det praktiska exempel jag arbetat med så syns även där, även om exemplet är kortfattat och inte går speciellt långt i skapandet av en ontologi, att det finns en tydlig nivåskillnad i den komplexitet som finns i en tesaur och i en ontologi. Det är också uppenbart från exemplet att ontologin ger utrymme för mer användbar information om ett område utan att för den skull skapa speciellt mycket mer information, utan framförallt genom att utnyttja den information som redan finns och göra kopplingar mellan olika nivåer i hierarkin.

Vilken betydelse har ontologier för den semantiska webbens tillkomst?

Som jag försökt beskriva i kapitel 5, så är ontologier en av de viktiga komponenterna som måste finnas för att Tim Berners-Lees och andras vision om den semantiska webben skall kunna bli verklighet (Berners-Lee et al. 2001). Den stora utmaningen ligger i att förändra den bakomliggande strukturen för WWW på ett sådant sätt att man rör sig från dagens system där informationen formateras, framförallt för att läsas och användas av människor. Datorernas roll i det hela är då framförallt att tolka kodningen och se till att informationen presenteras på det sätt som producenten tänkt sig. För att uppnå den ”intelligentare webb”, som man ser framför sig i den semantiska webben, måste datorerna inte bara förstå hur information skall presenteras utan också tolka innebörden i den information som presenteras. Grunden till detta ligger naturligtvis i användandet av metadata. Det finns dock problem med detta, dels beroende på att man hittills varit tvungen att mata in mycket av denna information manuellt, eller åtminstone redigera den för hand, dels på att det finns ett antal sätt att missbruka denna typ av ”osynlig” information på webbsidor. Det förekommer till exempel att det medvetet läggs in icke-relevant information i metataggar för att få högre rankning i sökmotorer eller automatiskt länka till andra sidor, ofta i form av pornografi.

De steg som finns under ontologierna i den semantiska t.ex. XML och RDF, har skapats för att göra det möjligt att ta steget vidare från dagens HTML-baserade webbsidor och göra det enklare att generera den typen av information som behövs för att datorerna skall kunna börja bearbeta informationen. XML är också användbart i många andra sammanhang där man vill hantera information som ständigt kompletteras

eller förändras. Ju fler tillämpningsområden som finns för t.ex. XML, desto mer information finns också enkelt tillgänglig för den semantiska webben.

Man måste i detta sammanhang också komma ihåg att datorer enbart utför de instruktioner de fått av oss människor genom de program som körs. En dator kan bearbeta enorma mängder information på kort tid, men det betyder inte att datorn i sig tillägnar sig denna information utan den följer bara sina instruktioner. Detta är anledningen till att de olika lagren i den semantiska webben måste finnas och dessutom vara väldefinierade, eftersom det i huvudsak är datorerna som kommer att använda de olika funktionerna i den semantiska webben. För en mänsklig användare skall egentligen den semantiska webben vara osynlig. All bearbetning av information ska alltså ske i bakgrunden och endast resultatet ska presenteras för användaren. (Berners-Lee et al. 2001)

Var kommer då ontologierna in? Idag genereras en enorm mängd information som publiceras på WWW. Denna information skapas med hjälp av olika verktyg, med olika målgrupper, på olika språk, av personer med oerhört varierande datavana osv. Hur skall då den semantiska webben kunna väva ihop detta? Jo, genom att definiera hur man beskriver ett område skapas en möjlighet att knyta ihop information från många olika källor. En ontologi skapar alltså förutsättningar för hantering och bearbetning av information som finns utspridd på många ställen och lagrad i olika former. I och med de relationer och annan kompletterande information som finns definierad i ontologin kan datorerna också dra en del av de slutsatser som i nuläget kräver en människa. (Davies et al. 2003, s. 4)

Detta är dock inte helt trivialt, dels finns det som vi sett tidigare en uppsjö av olika uppfattningar om hur en ontologi skall vara konstruerad, dels finns det ett antal olika språk och verktyg för att konstruera dem. Dessutom finns ofta inget självklart förstahandsalternativ, när det gäller att beskriva en del av verkligheten, ett sätt är ofta lika bra som ett annat. En annan aspekt är att informationen på WWW är under ständig förändring. Här står man alltså inför en hel del problem att lösa och en del av dessa har inget givet svar idag, men det finns en del förslag:

- *Olika ontologier likvärdiga* – i den semantiska webben krävs att man väljer en och endast en ontologi, man gör ett ”committment”.
- *Information existerar på olika språk* – de övergripande målen med en ontologi, att definiera koncept och relationer dem emellan, skall vara oberoende av språk. Däremot finns en koppling till, språk när man skall göra ontologin begriplig för människor. En mappning till t.ex. svenska eller engelska behövs alltså, när en människa skall förändra eller utveckla en ontologi och exempel finns idag på hur detta kan fungera.
- *Kontinuerlig förändring* – ontologier kan inte vara statiska utan det måste finnas stöd för ett antal olika aspekter på förändring:
    - Området som ontologin beskriver förändras. Exempel: En ontologi som beskriver EU behöver förändras när nya lagar eller förordningar tillkommer.

  - Tillämpningen förändras. Exempel: En ontologi som beskriver trafiksystemet i Göteborg behöver anpassas om man byter användningsområde från cykeltrafikanter till spårvagnsförare.
  - Specifikationsförändring. Exempel: Ny version av det beskrivningsspråk som används, t.ex. OIL.

  Detta är naturligtvis inte okänt för dem som arbetar med ontologier, och de flesta verktyg för att hantera ontologier har därför funktioner för att göra dessa aspekter möjliga.
- *Floran av verktyg och ontologispråk* – Här kommer utvecklingen troligen att bero delvis på hur starka W3C kommer att vara i denna fråga. Med erfarenhet från vanliga WWW finns förhoppningar att de rekommendationer W3C utfärdar kommer att styra utvecklingen, vilket skulle tyda på XML, RDF och OWL, men detta får tiden utvisa.

Sammanfattning: För den semantiska webben är ontologier en nödvändighet, eftersom det krävs denna typ av övergripande struktur för att datorer skall kunna bearbeta och analysera information och därmed minska behovet av manuell analys av t.ex. sökresultat. Man kan se i litteraturen att det finns en enighet om att ontologier eller något liknande är av stor nytta både för den semantiska webben och för andra liknande tillämpningar. Däremot finns än så länge ingen övergripande standard eller entydig definition om vad begreppet innebär, vilket betyder att det finns en pågående debatt om definitionen som sådan, samtidigt som de mer pragmatiska arbetar med konkreta lösningar. Min bedömning är att man till slut kommer att enas kring något som W3C rekommenderar, men exakt hur det kommer att se ut får tiden utvisa.

7.2. Ontologier och framtiden

Ett antal intressanta aspekter har kommit fram under arbetet med denna uppsats t.ex. hur liknande behov inom olika områden har givit upphov till lite olika lösningar. Att organisera är ju ett urgammalt behov hos människan och med tiden har vi fått allt mer vi anser att vi behöver organisera. De senaste decenniernas explosion i form av elektroniskt lagrad information har naturligtvis ökat på detta, och det ställs därmed allt större krav på de system vi som användare vill ha för att hitta just den information vi söker.

Vid sidan av de användningsområden för ontologier jag fokuserat på i denna uppsats, nämligen klassisk kunskapsorganisation och vidareutvecklingen av denna, samt den semantiska webben finns också andra, närliggande områden. Jag tänkte kort beröra två av dessa här, eftersom det är intressant dels för att se vilka områden som berörs, dels för att ge indikationer om andra forskningsområden som bör vara av intresse för den som blivit intresserad av ontologier (kanske till och med genom denna uppsats, vad vet jag?).

Ett av dessa områden är det som på engelska brukar kallas e-commerce, e-handel på svenska, och ibland b2b (business to business). Här handlar det till stor del om att förenkla företags inköpsprocedurer genom att samtliga leverantörer måste ansluta sig till en elektronisk handelsplats. På denna gör de sina produkter och tjänster

tillgängliga, och den ansvarige inköparen på det upphandlande företaget kan direkt jämföra utbud och pris från flera leverantörer. Beställning och fakturering sköts sedan elektroniskt, vilket minskar pappershanteringen för alla parter. Detta ställer stora krav på systemets möjlighet att strukturera information ur flera aspekter:

- En leverantör kan ha en mycket stor flora av produkter och tjänster och ibland också sådant som är specialanpassat för en viss kund.
- Olika leverantörer har med stor sannolikhet olika benämningar på liknande eller identiska produkter.
- Stora krav ställs på säkerheten. Bara den som är godkänd som kund skall ha tillgång till informationen och dessutom bara den information han eller hon behöver.

Eftersom det finns stora pengar att spara här, framförallt för stora företag, finns också en stark drivkraft vilket gör att det är väl värt att studera de ontologier som finns och utvecklas inom detta område (Fensel 2001, s. 47ff; McGuinness 2001).

Det andra område jag vill nämna här är digitala bibliotek där man har i stort sett samma problem som på webben, dvs. en konstant ökande mängd information att strukturera och söka i. Detta är också ett område där det finns flera initiativ till samverkan mellan disciplinerna för att tillvarata den kunskap som finns både inom klassisk kunskapsorganisation och inom datavetenskap. Två sådana exempel är ”Content standardisation for cultural repositories” inom OntoWeb (OntoWeb-SIG 2003) och ett förslag till EU:s sjätte ramprogram från bl.a. Lunds Universitet kallat SEMKOS (Semantic enabling by advanced Knowledge Organization Systems for largescale information integration in scientific and cultural digital libraries) (SEMKOS 2003). Man kan också notera att det av någon anledning finns fler sådana här samarbetsinitiativ i Europa än vad man hittar i övriga delar av världen.

En naturlig fråga som följd av ovanstående stycke och som jag också berört i uppsatsen, är förekomst och avsaknad av samarbete mellan flera discipliner vad gäller arbetet med ontologier. Om man generaliserar något kan man säga att intresset har varit svalt från den datavetenskapliga sidan, AI och expertsystem, att undersöka vad som finns i de mer traditionella kunskapsorganisatoriska områdena inom biblioteks- och informationsvetenskap. Till viss del kan detta kanske bero på att man har angripit problemet från olika håll och med olika teoretisk bakgrund. Från bibliotekshåll har fokus traditionellt varit på att underlätta för t.ex. låntagare att hitta de böcker de är intresserade av och därmed har kravet varit att det skall var en naturlig form av struktur. Inom den datavetenskapliga disciplinen utgår man istället från programmeringsaspekter och logiska teorier, vilket gör det enklare att implementera för datorer, men ofta svårare att förstå för människor.

Min egen ståndpunkt i denna fråga är att det visst finns områden där man hade tjänat på att samarbeta, men att det finns en viss överdrift från en del debattörer som vill hävda att ontologier bara är ett nytt namn för klassifikation. Om inte annat så tror jag att det från kunskaporganisatorisk synvinkel finns en hel del att lära sig från datavetenskapen, inte minst med tanke på den snabba utveckling vi ser inom digitala bibliotek och alternativa publiceringsmetoder.

Hur ser då framtiden ut för användandet av ontologier, framförallt inom den semantiska webben? Såsom framgått i texten tidigare ses ontologierna som en mycket viktig byggsten i sammanhanget, men det bör också påpekas att det finns problem med dem. Som redan nämnts finns det vitt skilda uppfattningar om vad en ontologi egentligen är och hur den bör konstrueras, men det finns också ett antal andra faktorer man bör ta i beaktande:

- Än så länge finns ingen standard för hur ontologier skall beskrivas, eller rättare sagt det finns ett antal olika standarder. Tiden får utvisa om det till slut blir en gemensam, överenskommen standard som gäller, eller om det blir någon form av de facto standard som kommer att dominera, eller, vilket kanske är mest troligt, att det kommer att samexistera ett antal beskrivningssätt. För att kunna kombinera ontologier krävs då att man enas om någon form av minsta gemensamma nämnare för att kunna utbyta information.
- Att skapa ontologier kräver en arbetsinsats. Även om man kan kombinera och komplettera redan existerande ontologier behövs ändå en insats av en person som behärskar området för att få ett bra resultat, eftersom det inte går att automatisera alla delar av arbetet.
- Ontologierna är beroende av att det finns existerande metadata och vi vet att det redan i dag är svårt att få in den mängd metadata man önskar för att förbättra bl.a. söksystemen. Dessutom finns stora problem med att försäkra sig att metadata som finns verkligen är relevant.
- Hur skall man hantera ”nyansskillnader”, dvs. att man i olika ontologier använder snarlika men ändå olika koncept för samma område?
- Kan man kräva att alla som skapar webbsidor också genererar den information som är nödvändig för att ett ontologibaserat system skall kunna hitta och bearbeta informationen?

Som vi såg i kapitel 4, har ontologi ursprungligen betydelsen "läran om varande", hur man skall kunna beskriva allt som finns. Detta är alltså något mycket vittomfattande och även om betydelsen i detta sammanhang är något snävare är det fortfarande en uppgift som är oerhört krävande. Alltså, att på något sätt standardisera beskrivningen av all information som finns tillgänglig på Internet och via andra källor. Jag hoppas att uppsatsen speglat lite av denna uppgifts komplexitet och både visat på möjligheter och problem som man står inför. Det känns som man idag är långt ifrån den vision Tim Berners-Lee presenterade 1994, men samtidigt: vem kunde för 10 år sedan ana att Internet och WWW skulle få den utveckling vi sett?

Jag skulle vilja avsluta med ett mer än hundra år gammalt citat från Samuel Butler som har tydliga paralleller till vad som behandlats i denna uppsats:

> ”Jag dristar mig föreslå …. att utvecklingen av mänskligheten är fullbordad när alla, överallt, utan att det tar någon tid, till ett billigt pris, äger vetskap om allt de önska veta om vad som helst, var som helst …. Detta är den stora sammansmältningen av tid och rum som vi alla strävar efter.” (Butler 1863) (författarens översättning)

8. Sammanfattning

Denna uppsats har behandlat begreppet ontologi, framförallt som det används i samband med den semantiska webben. En av de huvudsakliga frågeställningarna har varit hur ontologi ställer sig i förhållande till klassifikationsscheman och tesaurer, eftersom dessa har ett antal liknande egenskaper.

Ordet ontologi kommer ursprungligen från filosofin och har där betydelsen ”läran om varande”, men i denna uppsats används det i den något snävare betydelse som har kommit att tillämpas i informationshanteringssammanhang. Framförallt har ontologier aktualiserats de senaste åren i och med visionen om den semantiska webben, en påbyggnad till dagens World Wide Web.

Jämförelsen mellan ontologi, klassifikation och tesaurer har gjorts med den semantiska webben som referensram eftersom det är i detta sammanhang ontologierna förväntas få stor betydelse för en effektivare informationshantering och –strukturering. Det är också i ljuset av denna utveckling som det är lättast att se parallellerna med de klassiska kunskapsorganisatoriska begreppen klassifikation och tesaurer.

Resultaten av analysen av dessa begrepp visar att det finns tydliga kopplingar dem emellan, men att ontologierna tillför ett antal viktiga aspekter. Framförallt tesaurer liknar till viss grad ontologier och kan till och med betraktas som en enkel form av ontologier, beroende på hur man gör definitionen. Den förhärskande uppfattningen i litteraturen är dock att en ontologi befinner sig högre på komplexitetsskalan än klassifikation och tesaurer.

Det som skiljer ontologier är framförallt formella, alltså maskinellt bevisbara och otvetydiga, relationer och möjligheten att definiera koncept. Ontologierna blir därmed mer användbara ur ett datorperspektiv och ger datorsystemen möjlighet att göra slutledningar baserade på den information som finns. Detta innebär dels att ny information skapas tack vare att datorerna kan börja förstå sammanhang, och dels att en stor del av analysen av sökresultat som idag måste göras manuellt kan hanteras av datorer.

Man kan dock konstatera att det finns ett antal likheter mellan det arbete som utförts, framförallt inom datavetenskapen, de senaste åren för att hantera den växande mängden information som existerar både på Internet och på företagsinterna nätverk, med det klassiska informationshanteringsproblemet inom kunskapsorganisation. Tyvärr finner man också att informationsutbytet dessa discipliner emellan har varit begränsat. Det finns dock ljuspunkter, inte minst i form av ett antal EU-finansierade projekt.

9. Litteraturförteckning

AAT (2000). *AAT - Art and Architecture Thesaurus*. http://www.getty.edu/research/tools/vocabulary/aat/ [2003-07-10]

Aitchison, Jean, Gilchrist, Alan & Bawden, David (1997). *Thesaurus construction and use: a practical manual,* Aslib, London.

Arango, Guillermo (1994). A Brief Introduction to Domain Analysis. *ACM Symposium on Applied Computing, Phoenix, Arizona*, s. 42-46

Baeza-Yates, Ricardo & Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval,* ACM Press, New York.

Berners-Lee, Tim (1994). *W3 future directions*. http://www.w3.org/Talks/WWW94Tim/ [2003-05-16]

Berners-Lee, Tim, Hendler, James & Lassila, Ora (2001). The semantic web. *Scientific American*, Vol. 279, May. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 [2003-04-12]

Bray, Tim, Paoli, Jean, Sperberg-McQueen, C. M. & Maler, Eve (2000). *Extensible Markup Language (XML) 1.0*. http://www.w3c.org/TR/REC-xml [2003-04-28]

Bryan, Martin (1997). *Introduction to XML*. http://www.personal.u-net.com/~sgml/xmlintro.htm [2003-06-20]

Butler, Samuel (1863). Darwin among the Machines. *Press Newspaper,* June 13

Cailliau, Robert (1995). *A Little History of the World Wide Web*. http://www.w3.org/History.html [2003-05-16]

Chowdhury, Gobinda G. (1999). *Introduction to modern information retrieval,* Library Assoc. Publ., London.

Davies, John, Fensel, Dieter & Harmelen, Frank van (Eds.) (2003) *Towards the semantic web: ontology-driven knowledge management,* John Wiley & Sons Ltd., Chichester.

DCMI (2003). *Dublin Core Metadata Initiative.* http://www.dublincore.org [2003-05-16]

Ding, Ying (2001) A review of ontologies with the Semantic Web in view. *Journal of Information Science,* Vol. 27, No. 6, s. 377-384.

Ding, Ying & Foo, Schubert (2002) Ontology research and development. Part 1 - a review of ontology generation. *Journal of Information Science,* Vol. 28, No. 2, s. 123 - 136.

Fellbaum, Christiane (Ed.) (1998) *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge MA.

Fensel, Dieter (2001). *Ontologies: A silver bullet for knowledge management and electronic commerce*, Springer, Berlin.

Fensel, Dieter, Hendler, James, Lieberman, Henry & Wahlster, Wolfgang (Eds.) (2003) *Spinning the Semantic Web. Bringing the World Wide Web to its full potential.*, The MIT Press, Cambridge MA.

Gilchrist, Alan (2003) Thesauri, taxonomies and ontologies - an etymological note. *Journal of Documentation*, Vol. 59, No. 1, s. 7-18.

Golbeck, Jennifer, Hendler, James & Parsia, Bijan (2003). Trust Networks on the Semantic Web. *13th Annual World Wide Web Conference, Budapest*,

Gruber, Tomas R. (1993a) In *Formal Ontology in Conceptual analysis and Knowledge Representation*(Ed, Nicola Guarino, Roberto Poli) Kluwer Academic Publisher, Padova.

Gruber, Tomas R. (1993b) A translation approach to portable ontologies. *Knowledge Acquisition*, Vol. 5, No. 2, s. 199-220.

Guarino, Nicola & Giaretta, Pierdaniele (1995). Ontologies and Knowledge Bases. Towards a terminological clarification. *2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases, Amsterdam*, Enschede: IOS, s. 25-32

Hartman, Sven G. (1990). *Handledning. Liten handbok för den som arbetar med projekt, specialarbeten eller rapporter.*, Universitetet i Linköping, Linköping.

Harvey, Ross (1999). *Organising knowledge in Australia: principles and practices in libraries and information centres*, Centre for Information Studies, Wagga Wagga, NSW.

Hedberg, Sten (2001). *Metadata - Kataloginformation på Internet. Dublin Core*. http://www.kb.se/Nvb/Katalog/kat5.htm [2003-05-16]

Hjørland, Birger (2002) Domain analysis in information science. *Journal of Documentation*, Vol. 58, No. 4, s. 422-462.

Hjørland, Birger & Albrechtsen, Hanne (1995) Toward a new horizon in information studies: Domain-analysis. *Journal of the American Society for Information Science*, Vol. 46, No. 6, s. 400-425.

Iselid, Lars (2001). Den semantiska webben - en revolution på Internet? *Datormagazin*, 8, s. 96-100

Klein, Michel, Broekstra, Jeen, Fensel, Dieter, Harmelen, Frank van & Horrocks, Ian (2003) In *Spinning the Semantic Web. Bringing the World Wide Web to its full*

potential.(Eds, Fensel, Dieter, Hendler, James, Lieberman, Henry and Wahlster, Wolfgang) MIT Press, Cambridge MA.

Kronman, Ulf & Parnefjord, John (1999) Resource Description Framework Metadata för Internet. *Human IT Tidskrifter för studier av IT ur ett humanvetenskapligt perspektiv*, No. 4.

Langridge, Derek Wilton (1989). *Subject Analysis - Principles and Procedures,* Bowker - Saur, London.

Lassila, Ora & Swick, Ralph R. (1999). *Resource description framework (RDF) model and syntax specification.* http://www.w3.org/TR/REC-rdf-syntax/ [2003-03-25]

McGuinness, Deborah L. (2001) Ontologies and Online Commerce. *IEEE Intelligent Systems,* Vol. 16, No. 1, s. 8-14.

McGuinness, Deborah L. (2003) In *Spinning the Semantic Web. Bringing the World Wide Web to its full potential*.(Eds, Fensel, Dieter, Hendler, James, Lieberman, Henry and Wahlster, Wolfgang) MIT Press, Cambridge MA.

Nationalencyklopedin (2003). *Ontologi*. www.ne.se/jsp/search/article.jsp?.i_art_id=276179 [2003-03-18]

On-To-Knowledge (2002). *On-To-Knowledge*. http://www.ontoknowledge.org [2003-07-31]

Ontolingua (2002). *Ontolingua Home Page*. http://www.ksl.stanford.edu/software/ontolingua/ [2003-06-28]

OntoWeb-SIG (2003). *OntoWeb SIG on Content Management*. http://www.ontoweb.org/sig.html [2003-07-16]

Pepper, Steve (2002). *Topic Maps and RDF*. http://www.ontopia.net/topicmaps/materials/rdf.html [2003-06-06]

Poli, Roberto (1996). Ontology for knowledge organisation. *Fourth International ISKO Conference, Washington*, Frankfurt: Indeks Verlag, s. 313-319

Qin, Jian & Paling, Stephen (2001) Converting a controlled vocabulary into an ontology: the case of GEM. *Information Research,* Vol. 6, No. 2.

Ray, Erik T (2001). *Learning XML*, O'Reilly, Sebastopol.

Rowley, Jennifer & Farrow, John (1992). *Organizing knowledge,* Gower Publishing Limited, Aldershot.

SEMKOS (2003). *Semantic enabling by advanced Knowledge Organization Systems for largescale information integration in scientific and cultural digital libraries*. http://www.lub.lu.se/SEMKOS/ [2003-07-10]

Soergel, Dagobert (1996). SemWeb. Proposal for an open, multifunctional, multilingual system for integrated access to knowledge base about concepts and terminology. *Fourth International ISKO Conference, Washington*, Frankfurt: Indeks, s. 165 - 173

Soergel, Dagobert (1999) The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, Vol. 50, No. 12, s. 1119-1120.

Unicode (2003). *The Unicode Consortium*. http://www.unicode.org [2003-05-01]

Vickery, Brian C. (1997) Ontologies. *Journal of Information Science*, Vol. 23, No. 4, s. 277-286.

W3C (2002). *Web Naming and Addressing Overview*. http://www.w3c.org/Addressing/ [2003-05-20]

W3C (2003a). *RDF Primer*. http://www.w3.org/TR/rdf-primer/ [2003-10-11]

W3C (2003b). *W3C Technical Reports and Publications*. http://www.w3.org/TR/ [2003-05-28]

W3C (2003c). *W3C XML Schema*. http://www.w3.org/XML/Schema [2003-05-16]

W3C (2003d). *XML Signature WG*. http://www.w3.org/Signature/ [2003-05-18]

Webster's (1993). *Webster's Third New International Dictionary*, Könemann.

Welty, Christopher & Guarino, Nicola (2001) Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering*, Vol. 39, No. 1, s. 51-74.

Wielinga, Bob J., Schreiber, A. Th., Wielemaker, Jan & Sandberg, Jacobijn A. C. (2001). From Thesaurus to Ontology. *Proceedings of the First International Conference on Knowledge Capture, Victoria*, New York: ACM Press, s. 194-201