

Den semantiska – ny revolution på Internet?

Artificiell intelligens, peer-to-peer, natural language processing och länkvalidering i all ära, men i jakten på den perfekta söktjänsten behövs också förbättrade och pålitliga metadata. Där står sig HTML och dess metataggar rätt slätt som hjälpmedel. Lösningen heter Resource Description Framework (RDF), enligt World Wide Web Consortium (W3C) och Internets "uppfinnare" Tim Bernes-Lee.

RDF är ett ramverk som tillämpar XML. Bakom hörnet lurar dock spamarna, som kan göra att RDF aldrig kommer att kunna revolutionera sökningen på webben.

Metadata

Med metadata menas data om andra data, ett sätt att beskriva innehållet i en resurs för att göra den lättare att hitta. Som mycket annat i vårt samhälle är även metadata i behov av standarder. Tanken med standardiserade metadata är inget nytt. Redan innan det första HTML-dokumentet skapades existerade redan miljontals digitala metadata-dokument. Dessa var skapade enligt vissa regler för beskrivning av dokument och samlade i ett format kallat MARC, Machine-Readable Cataloging. Söker du fram en bok i en svensk bibliotekskatalog så är det med all säkerhet ett metadata-dokument beskrivet i formatet MARC.

Men till skillnad från en MARC-post

innehåller ett HTML-dokument allt som oftast också längre texter med tillhörande bilder – det som i biblioteket skulle motsvaras av sidorna i en bok. En MARC-post däremot innehåller bara en hänvisning till denna bok. Det som motsvarar metataggarna i ett HTML-dokument.

XML

Det finns stora brister i HTML-dokumentens beskrivning av en webbsidas innehåll. Ett verktyg som har skapats särskilt för att förbättra beskrivningen av innehållet i webbsidor är Extensible Markup Language (XML). Utarbetat av W3C nådde XML så kallad Recommendation Status i februari 1998 och plötsligt sågs XML närmast som en frälsning för utvecklingen på Internet. En andra version släpptes förra hösten.

Liksom HTML härstammar XML från SGML, ett märkspråk som skapades av IBM i slutet av sjuttio-talet och blev en ISI-standard i mitten av åttiotalet. SGML kom däremot inte att användas mer än inom förlagsindustrin och en del akademiska och statliga projekt. Med SGML som förebild skapades därför HTML av Tim Bernes-Lee som då var verksam vid CERN i Schweiz.

HTML har den nackdelen att det mer beskriver en webbsidas utseende än dess innehåll, vilket gör att sökmotorerna på Internet har stora problem att förstå innehållet på webbens miljontals

och åter miljontals sidor på ett vettigt sätt. Med XML som verktyg ges möjligheten att skapa bättre beskrivningar av webbsidornas innehåll, vilket ska underlätta arbetet för sökmotorerna. I XML har man hållit isär utseendet från innehållet genom att skapa ett specifikt språk för stilmallar som följaktligen fått namnet Extensible Stylesheet Language (XSL), och som nådde Candidate Recommendation Status i november förra året.

XML har ju den fördelen att det är utbyggbart och flexibelt, men har samtidigt den nackdelen att XML-processorn inte är lika förlåtande mot bristfälligt skriven kod som HTML. Något som kan ses som både positivt och negativt.

Egna taggar

Utbyggbarheten i XML innebär konkret att man kan skriva sina egna taggar, som inte måste men bör definieras i en så kallad DTD eller i ett XML-schema för att bli brukbart som maskinläsbara data.

Även om man kan definiera helt egna taggar i XML och det i många fall liknar HTML, så kräver validerad XML normalt mer av den som skapar webbsidor än HTML. Alla kanske inte skriver under på detta påstående, men rent generellt upplevs det så.

Det finns ett starkt behov av användarvänliga XML-editorer. Inom den akademiska världen drivs projekt där

webben

Internet håller mer eller mindre på att kollapsa under sin egen tyngd, och det finns inte längre någon möjlighet för sökmotorerna att indexera ens en bråkdel av all information. För att lösa detta problem krävs en genomgående förändring av hur innehållet på enskilda webbsidor beskrivs.

En vision är "den semantiska webben" där all information är så väl beskriven att sökmotorernas jobb blir mycket enklare.

man försöker få forskare att skriva sina texter direkt i XML-editorer i stället för Microsoft Word – som är absolut vanligast – eftersom XML är ett betydligt bättre format för elektronisk publicering i olika varianter.

Men frågan är, för att vara lite provokativ, om XML någonsin kommer att få något genomslag hos de breda folklagren av HTML-kodare. I stället kanske det förblir ett exklusivt verktyg för större företag och den akademiska världen som vill kunna utbyta maskinläsbara data. Vissa kanske vill hävda att XML redan har lämnat hypestadiet och fått ett brett genomslag. Kan så vara, men XML är ingen metadatastandard, bara ett verktyg för att skapa sådana standarder definierade genom XML-scheman och DTD:er. Problemet är inte att skapa standarder utan att komma överens om standarder. Och vad är det för mening med att skapa standarder om de sedan inte används?

RDF och den semantiska webben

XML är mycket användbart för utbyte av maskinläsbara data mellan applikationer som känner till varandras data, men inte för situationer när nya kommunikationspartners dyker upp. Av denna anledning och beroende på det starka behovet av strukturerade data på Internet där information lätt kan hittas och utbytas föddes idén om den

semantiska webben av W3C och Tim Bernes-Lee. Så här uttrycker W3C det med egna ord:

"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."

För att uppnå dessa högt uppställda mål skapade W3C ramverket RDF, som nådde Recommendation Status 22 februari 1999. RDF är inte heller en metadatastandard, utan ett ramverk för utbyggbara metadatastandarder. I RDF liksom i XML går det däremot att skapa scheman som kan bli standarder. Dessutom har RDF den öppenheten att

det tillåter att flera olika scheman används för samma dokument. Det ger RDF som ramverk en öppenhet som gör att flera olika redan existerande format för metadata kan komma att överleva, till exempel MARC eller Dublin Core (DC). RDF definierar standarder på tre nivåer:

Struktur – RDF-datamodell.

Syntax – RDF-syntax som uttrycks i XML.

Semantik – RDF-schema, till exempel DC.

RDF-datamodellen anger hur resurser struktureras och hur deras relationer sinsemellan kan beskrivas.

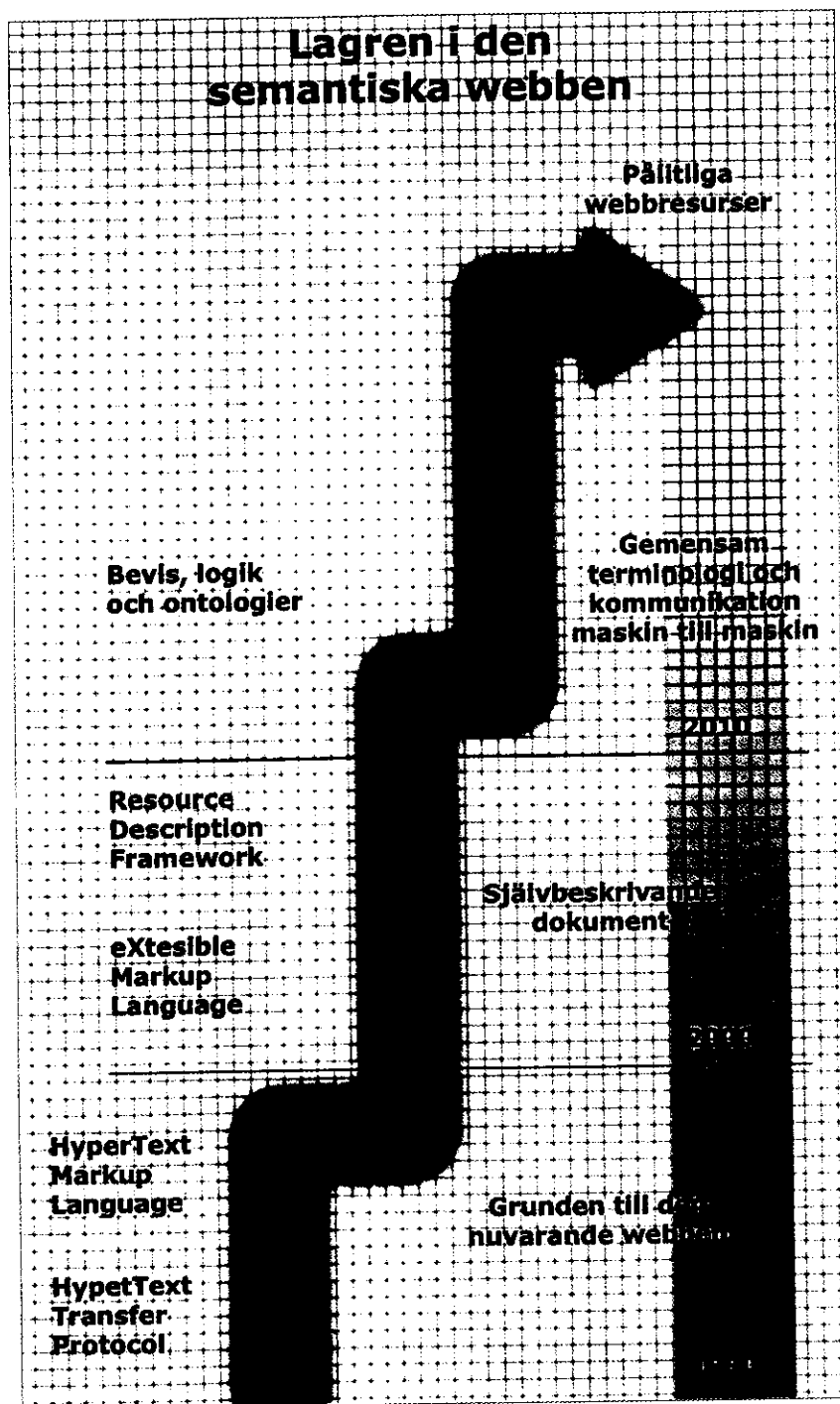
Datamodellen utgår från att det finns en unikt identifierbar resurs med ett antal egenskaper som i sin tur, var för

XHTML – utbyggbar HTML till vilken nytta?

XHTML har kallats för HTML 5.0, men är mycket mer än bara en ny version av HTML. XHTML liknar visserligen i mångt och mycket HTML 4.0 och är fortfarande kompatibelt med dagens webbläsare som klarar HTML 4.0, men det är en omformulering av HTML 4.0 som en applikation av XML. XHTML är tänkt som en övergångsstandard till XML. Liksom i XML krävs det i XHTML att koden är välformulerad, det vill säga att den följer vissa regler som finns definierade i XML 1.0-specifikationen. XHTML har också den utbyggbarhet, det vill säga att du kan definiera egna taggar, som finns i XML.

I och med den förutspådda utvecklingen inom trådlöst Internet och den mängd av handdatorer som börjar röna allt större framgång är tanken att XHTML, eftersom det finns problem med dåligt formulerad HTML, ska göra det lättare för webbsidor att fungera på den flora av olika plattformar och webbläsare som den utvecklingen medför.

XHTML 1.0 fick Recommendation Status 26 januari förra året och version 1.1 fick Proposed Recommendation Status 6 april i år.



sig, har ett värde. I Figur 1 åskådliggörs detta.

URL och URI

Resursen kan vara en samling webbsidor, en enstaka webbsida eller en del av en webbsida. Resursens unika identitet kallas Uniform Resource Identifier (URI). På Internet är det samma sak som Universe Resource Locator (URL) – resursens adress på Internet. I detta exempel: <http://www.drugnews.nu/>

Egenskaper utgör en aspekt, en relation eller ett attribut som är kännetecknande för resursen. Varje enskild egen-

skap har en specifik innebörd. I detta exempel: titel.

Det krävs också att egenskapen får ett värde. I detta exempel: Drugnews – Nyhetsbyrån som tar tempen på drogkampen!

I figur 2 kan vi se hur de tre beståndsdelarna resurs, egenskap och värde bildar ett RDF-uttryck.

Jag tänkte inte här gå djupare in på de syntaktiska delarna av RDF, utan mer resonera om vilken betydelse den semantiska webben kan få för sökning och informationsutbyte på webben.

RDF öppnar möjligheter för att

Visionen om den semantiska webben sträcker sig många år framåt. Mycket förenklat kan man säga att vi är i början av Internets andra fas (självbeskrivande dokument, XML m.m.). Nästa fas inträder när en majoritet av alla dokument och maskiner på Internet kan kommunicera med en gemensam terminologi. I visionen om den semantiska webben ligger de tidigare faserna i Internets utveckling kvar som underliggande lager.

skapa nya scheman eller använda redan förekommande. Redan nu finns möjligheten att använda DC i RDF. DC var en inspirationskälla när W3C utvecklade RDF. Men vad är då DC?

DC

DC skapades 1995 i Dublin, Ohio, USA, sätet för OCLC (www.oclc.org), vid ett möte mellan bibliotekarier och Internet-expertter. Man såg att det fanns ett behov av förbättrade metadata för webbsidor, och MARC-formatet, som är det rådande formatet på biblioteken, ansågs inte vara anpassat för det behovet. Man hade insett att på Internet skötte varje enskild webbansvarig katalogiseringen, så därför ville man skapa ett verktyg som var anpassat för dem.

I största möjliga mån ville man ändå att en överensstämmelse med bibliotekskatalogisering skulle eftersträvas. Ett metadataformat som innehåller 15 olika fält med etiketter av HTML-utseende arbetades fram.

DC är ett av de första standardiserade försöken att skapa ett metadata-schema för webben, och det stöds av svenska sökmotorer som Safari (safari.hsv.se), Svesök (www.svesok.kb.se) och Svenska Miljönätet (smn.environ.se). Särskilt väl passar DC ihop med RDF-datamodellen och XML-syntaxen. RDF kan därför innebära att DC får ny skjuts under fötterna och faktiskt överlever.

DC har fått ett erkännande, men används ytterst sparsamt på webbsidor än så länge. Till hjälp för att skapa DC-metadata för webbsidor finns en mängd så kallade DC-generatorer. En av dessa är Nordic DC metadata creator (www.lub.lu.se/c/s.dll/nmdc.pl).

Både Safari och Svenska Miljönätet har egna DC-generatorer. Dessa DC-generatorer skapar självständiga DC-metadata som inte är integrerade som ett schema i RDF.

Den tyska RDF-editorn RDFedit (www.jan-winkler.de/dev/d_rdfedit.htm) kan däremot skapa inbäddad DC.

Det går naturligtvis att använda kontrollerad vokabulär i DC, men det finns ingen möjlighet att validera den genom

en tolk. Idén med den semantiska webben är bland annat att försöka skapa kontrollerad vokabulär som går att validera. Det brukar kallas ontologier.

Tre lager

Den semantiska webbens arkitektur består av tre lager (se figur 3). Datalagret är själva datamodellen med syntax. Schemalagret ger oss möjlighet att definiera vokabulären och strukturen för att uttrycka metadata för webbresurser så att en RDF-tolk kan validera ett RDF-dokument. Men kvarstår gör ändå problemen med att tala samma språk, som ska lösas i det logiska lagret. Därför finns tanken att använda onto-

logier för att skapa kontroll över kommunikationen.

En ontologi är helt enkelt ett dokument eller en datafil som i formella termer definierar de inbördes relationerna mellan termer.

Precis som nätverksprotokoll skapats som ett språk för att datorer i nätverk ska förstå varandra är ontologier ett sätt att skapa enhetligt språk inom olika domäner som ska inte bara förbättra kommunikationen mellan olika applikationer, utan även hjälpa dem att hitta rätt information.

Ontologier skulle kunna lösa problemet med synonymer (som dikter, poesi, lyrik) och polysema ord (fil betyder

både ett verktyg, filmjolk, körfält och en datafil). Skillnaden mellan scheman och ontologier är helt enkelt att scheman beskriver dokumentstruktur och ontologier skapar enhetligt språk inom en viss domän.

Ontology Inference Layer (OIL) (www.ontoknowledge.org/oil/) är ett försök till att skapa ontologier och skulle därför kunna integreras i den semantiska webbens tredje lager, det logiska. Jag tänkte inte här gå närmare in på hur OIL är uppbyggt, men för den som vill läsa mer finns artikeln "OIL in a Nutshell" (som PDF-fil) av D. Fensel med flera på www.cs.vu.nl/~ontoknow/oil/download/oilnutshell.pdf.

Vad säger experterna om den semantiska webben?

Det finns skilda åsikter om hur framgångsrika W3C:s idéer om att förenkla sökningen på Internet kommer att bli. Vi frågade tre experter på området om vad de tror i ämnet. Frågorna var:

1. Tror du att Tim Bernes-Lees och W3C:s vision om den semantiska webben kommer att revolutionera sökning på webben?

2. Tror du att digitala signaturer kan stoppa spammare? Det vill säga de som manipulerar sökmotorer genom att skriva in populära sökord som Britney Spears och Napster i metatagarna.

Danny Sullivan

Känd söktjänstexpert med ett förflutet som journalist. Flitig skribent som driver webbtjänsten Searchenginewatch.com och deltar aktivt på konferenser, både som föreläsare och moderator.

1. Förmodligen inte. I teorin låter det fantastiskt, men i realiteten handlar det om att vi inte har pålitliga metadata och sådana system lämnar idag webbsökningen öppen för spammare.

Idén till den semantiska webben, som jag förstått den, ligger i att om vi bara förser våra sidor med passande taggar så kommer vi att hitta sidor skrivna av en viss person eller ett visst företag, om ett visst ämne, skrivet vid en viss tidpunkt och så vidare.

XML ses som nyckeln till detta. Men XML är ingen standard, bara ett ramverk. Det är detsamma som att säga: "Jag vill bygga ett hus och jag har betong och jag har stål." Du har materialen för att bygga ett hus, men inga ritningar att följa. Människorna bakom Dublin Core har föreslagit en standard för hur man taggar sidorna. OK, då har vi en standard. Problemet är bara att standarden inte används av de ledande sökmotorerna, för de vet sedan länge att metadata inte går att lita på.

2. Jag tror att digitala signaturer för metadata skulle bli väldigt svåra att genomföra, dyra att arbeta med och ändå skulle inte den stora massan av allt innehåll på webben bli katalogiserat, vilket är tanken med den semantiska webben.

Greg Notess

Söktjänstexpert och bibliotekarie vid Montana University i USA. Flitig skribent i Online och Econtent och flitig föreläsare på konferenser världen över. Driver också webbtjänsten Searchengineshowdown.com.

1. Om jag har förstått det rätt, nej. Även om det är en intressant tanke ser det ut som om det är fokuserat på forskningsresurser. Dessutom så skulle det krävas stora förändringar för att det skulle kunna ske på hela webben, som tanken är. Att få alla att ta till sig denna modell, och göra det rätt, är att kräva en hel del. Det kanske får genomslag inom vissa segment av webben, men det är alldeles för tidigt att säga något än.

2. Jag har inget stort hopp om att någonting kan förebygga spamning. I bästa fall kommer en del av dessa initiativ att tillfälligt hejda eller minska mängden av spam, men för att digitala signaturer ska få betydelse måste tillräckligt många individer börja använda dem. Och än en gång, att hoppas på att användarna globalt ska ändra sig så dramatiskt, det är att hoppas på för mycket.

Wallace Koehler

Professor och lärare vid School of library and information studies vid University of Oklahoma, USA. Har forskat om bland annat system för metadata och informationsetik.

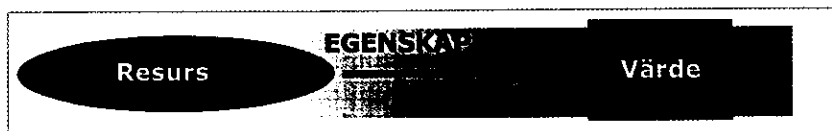
1. Revolutionera är ett starkt ord. Nej, jag tror inte det.

Problemet med detta förslag som med alla indexerings-scheman där författaren gör jobbet är just det att de är gjorda av författaren. Det har gjorts många studier som visat att för det första: nyckelord skrivna av författaren själv tenderar att bli otillräckligt standardiserade, och för det andra: även kvalificerade indexerare ofta inte kommer överens med varandra eller ens med sig själva efter en tid.

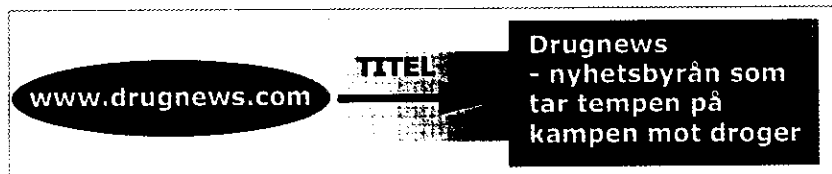
Datorer är inte mer anpassade att förstå eller tillskriva saker "mening" än människor, och i flera fall mindre än de. Den riktiga revolutionen kommer, om den kommer, när ämnesordslistor kan utarbetas med kontrollerade vokabulärer tvärs över discipliner och språkgränser.

2. Nej, men det kommer att minska deras aktivitet. För ett tag. Man måste komma ihåg att varje system som människan skapar går att kringgå. Digitala signaturer kommer att leda till ännu mer sofistikerade spänningsmetoder.

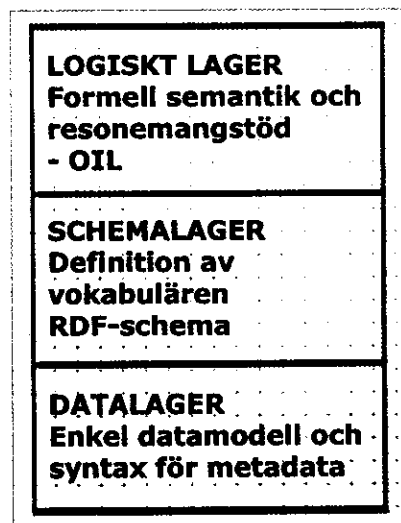




Figur 1. Datamodellen för RDF anger hur resurser struktureras och hur deras relationer sinsemellan kan beskrivas. Datamodellen utgår från att det finns en unikt identifierbar resurs med ett antal egenskaper som i sin tur, var för sig, har ett värde.



Figur 2. Här kan vi se de tre beståndsdelarna i ett RDF-uttryck: resurs, egenskap och värde.



Figur 3. Den semantiska webbens arkitektur består av tre lager. Datalagret är själva datamodellen med syntax. Schemalagret ger oss möjlighet att definiera vokabulären och strukturen för att uttrycka metadata för webbresurser så att en RDF-tolk kan validera ett RDF-dokument. Problemen med att tala samma språk ska lösas i det logiska lagret.

Övertro på RDF?

Det är inte alla som tror på den semantiska webben. XML-förespråkarna kallar ibland den semantiska webben för den pedantiska. Kritiken är bland annat den att de anser att det finns en övertro hos W3C vad gäller maskinläsbara data och förmågan att skapa mening ur dessa. Det anses också att RDF är för komplext – inte att koda, men att förstå, för att det ska få ett brett genomslag. XML-förespråkarna tycker att XML i sig själv är tillräckligt och vissa tror att artificiell intelligens (AI) och andra tekniker, inte XML just, är bättre på att skapa mening än de lösningar som förespråkarna för den semantiska webben står för.

Tim Bernes-Lee har själv, både i tal och skrift, tagit klart avstånd från AI. Semantiska webben ska på inget sätt förväxlas med AI. "I maskinläsbara RDF-dokument finns ingen inbyggd magisk AI som tillåter maskiner att återge

mänskligt mummel", som han har uttryckt det. "De visar bara på en maskins förmåga att lösa väldefinierade problem genom väldefinierade operationer på väldefinierade metadata."

Webbmasterns dilemma

Men om nu RDF och den semantiska webben verkligen kan förbättra möjligheten att hitta information, bör man då i egenskap av webbansvarig tagga sina sidor med RDF? Stöd för RDF och XML 1.0 finns ju i både Netscape 6 och Internet Explorer 5.5 – även om det inte har gått helt smärtfritt för Netscape att implementera det. Som vanligt tolkas inte heller de nya standarderna exakt likadant av webbläsarna.

Den springande punkten är ändå stödet för RDF i sökmotorerna. Varför lägga ned tid på att RDF-tagga om de stora sökmotorerna inte stödjer RDF? Och varför skulle sökmotorerna stödja

RDF (eller DC som funnits längre än RDF) när inte ens en bråkdel av alla webbsidor ute på Internet har RDF (eller DC)? Det liknar frågan om vad som kommer först – hönan eller ägget.

Några enkla svar på dessa frågor finns naturligtvis inte, men så länge det inte finns pålitliga metadata är det svårt att tro att sökmotorerna skulle börja stöda RDF.

Idag har spänningen av de metataggar som används i HTML-koden inneburit att fler och fler sökmotorer inte stöder någon annan metatagg än titletaggen. Vad skulle då kunna få sökmotorerna att tro mer på RDF? Digitala signaturer påstås vara lösningen.

Digitala signaturer

Digitala signaturer är ett sätt att genom kryptering kunna identifiera avsändare av information. Förhoppningen är att man ska kunna sortera ut dem som manipulerar metadata.

Men skeptikerna säger att tekniken bakom digitala signaturer aldrig kommer att bli 100-procentig. Inom W3C finns en arbetsgrupp för digitala signaturer (<http://www.w3.org/Signature/Activity>) som hoppas kunna presentera en föreslagen standard nu i sommar.

Oavsett om RDF kommer att innebära en revolution för sökning på Internet eller ej, så kommer det förmodligen att innebära ett steg framåt. Det brukar vara så när W3C och Tim Bernes-Lee är i farten.

Lars Iselid

World Wide Web Consortium (W3C)

W3C (www.w3.org) grundades i oktober 1994 med ambitionen att förbättra kommunikation och informationsutbyte på webben genom att utveckla gemensamma standarder. W3C leds av Tim Bernes-Lee, som brukar få äran av att ha skapat Internet. För den som vill läsa mer om honom rekommenderas boken "Weaving the web" som är en intressant läsning om hur han och andra nyckelpersoner vid forskningsinstitutet CERN i Schweiz utvecklade tanken om hypertext och www.

Medlemsorganisationen bär upp W3C ekonomiskt. I medlemsregistret finns en avsevärd del av världens ledande IT-företag som Apple, Microsoft och Oracle, för att nämna några.

Genom ett remissförfarande behandlas olika förslag inom W3C. Dessa förslag kan till slut bli standarder, som exempelvis HTML, som nu är uppe i version 4.0. Förslagen brukar kallas "note". Det finns fyra statusnivåer:

Working Draft (WD) – om W3C anser att ett förslag har värde tillsätts en arbetsgrupp som tar fram en WD.

Candidate Recommendation (CD) – här har förslaget fått betydande genomgranskning av tekniker. Implementering och teknisk feedback tar vid.

Proposed Recommendation (PR) – här markeras att man nu har tagit med hela standarden och endast mindre formella fel får förekomma.

Recommendation (R) – här råder konsensus inom W3C. Man gör kraftanstängningar för att bevara rekommendationerna och få förslaget spritt och implementerat som en standard. Alla eventuella uppkomna fel publiceras i ett "errata document". Dessa errata kan ge upphov till nya versioner av standarden.