Semanttinen Web

– kohti seuraavan polven Internet-palveluja

World Wide Web on nerokas järjestelmä tiedon ja dokumenttien välittämiseen ihmisten luettavaksi. Semanttinen web, merkitysten Internet, on visio seuraavan polven älykkäästä Internetistä, jonka käyttäjinä ovat ihmisten ohella koneet. Visiota toteutetaan XML-perustaisten "semanttisten" standardien ja välineiden avulla, joilla www:n sisältöjä voidaan esittää koneellisesti tulkittavassa muodossa ja näin toteuttaa seuraavan polven älykkäitä Internet-palveluja.

Eero Hyvönen
Tietotekniikan tutkimuslaitos (HIIT)
eero.hyvonen@cs.helsinki.fi

Nykyisen www:n tiedot on pääosin kuvattu HTML-kielellä, joka on ulkoisen muotoilun kieli. Esimerkiksi merkkaus <H1>Semantic Web </H1> ohjaa selaimen tulostamaan ylätason otsikon, muttei ei kerro mitään otsikon merkityksestä ilman ihmisen tulkintaa. Internetin käyttäjinä eivät kuitenkaan ole pelkästään ihmiset, vaan yhä enenevässä määrin koneet, kuten erilaiset sovellusohjelmistot, hakukoneet tai sähköisen kaupan agentit. Niille Internetin rakenteettomien sisältöjen, resurssien, tulkitseminen ja hyödyntäminen on vaikeaa.

Koneellisen tulkinnan ongelmaa on lähestytty yrittämällä tulkita verkon sisältöjä ihmisen tavoin. Vaikeutena on, että dokumenteissa käytetty kieli on sisällöllisesti monipuolista luonnollista kieltä, jossa voi olla myös puutteita ja virheitä. Ehkä vieläkin hankalampaa on kuvien, äänten ja musiikin, videoiden ja ohjelmistokomponenttien sisällön tulkinta. Verkon tieto on talletettu vaikeasti tulkittavissa muodoissa.

Semanttisen webin yksinkertaisena ideana on esittää tiedot siten, että niiden sisältöön päästään helpommin algoritmisesti käsiksi. Tämä mahdollistaa aiempaa olennaisesti älykkäämpien palveluiden kehittämisen ja johtaa www:n seuraavaan sukupolveen. Vastaavaan tapaan kun nykyinen XML-murros (Extensible Markup Language) erottaa tiedon rakenteen, syntaksin, sen ilmiasusta (HTML-taso), erotetaan Semanttinen web –murroksessa merkitys, semantiikka, XML-rakenteesta uusien korkammantasoisten standardien avulla.

1. POLVEN WWW (90-LUVUN ALKU)
ULKOASU EROON SIJAINTIPAIKASTA
URL, HTML, HTTP, PDF, ...

2. POLVEN WWW (90-LUVUN LOPPU)
RAKENNE EROON ULKOASUSTA
XML, XSL, ...

3. POLVEN WWW (SEMANTIC WEB)
MERKITYS EROON RAKENTEESTA
RDF(S), TOPIC MAPS, DAML+OIL, ...

Ensimmäisen polven www on dokumenttien ilmiasujen (esim. www-sivu) julkaisu- ja jakelukanava. Toisessa polvessa XML mahdollisti dokumenttien rakenteen (syntaksi) kuvaamisen ja monien ilmiasujen tuottamisen samasta rakenteesta. Kolmannen polven semanttisessa webissä domentteihin liittyy myös kuvauksi niiden sisällöstä (semantiikka), mikä mahdollistaa sisältöperustaiset älykkäät palvelut.

Semanttisen Webin teknologioita

Semanttisen webin välineet perustuvat yleensä XML:ään. XML on standardiperhe, jolla voidaan määritellä sovelluskohtaisia merkkauskieliä tietojen esittämiseksi. Esimerkiksi osoitetiedon esittämiseen voitaisiin määritellä XML-kieli ja esittää osoitteita seuraavaan tapaan:

```
<OSOITE>
<NIMI>Onni Ohjelmoija</NIMI>
<PUHELIN> 123 456 </PUHELIN>
</OSOITE>
```

XSL-transformaatioiden avulla osoitetiedoista voitaisiin automaattisesti tuottaa selaimella luettava normaali HTML-sivu, painettu osoitekirja tms. ulkoinen esitysmuoto. Ideana on dokumentin rakenteen ja ilmiasun erottaminen toisistaan.

XML-merkkaukseen ei sisälly mitään semantiikkaa koneen kannalta, vaan merkityksen kielen eri ilmauksille kuten ”OSOITE” antaa ihminen. XML kuvaa vain tiedon rakenteen, syntaksin, ei merkitystä, semantiikkaa.

Jotta kielellä olisi koneellisesti

tulkittavissa oleva merkitys, täytyy sen symbolien ja rakenteiden viitata johonkin alla olevaan malliin. Esimerkiksi logiikassa merkitys muodostuu malliteorian joukko-opillisten rakenteiden kautta, jotka kuvaavat mahdollisia maailmantiloja. Semanttisen webin konseptin keskeinen oivallus kielelliseltä kannalta on tämän käyttökelpoiseksi osoittautuneen idean tuominen www-maailmaan.

RDF(S) tunnetuin Semanttisen webin kieli

Tunnetuin Semanttisen Webin kieli on www:n kehitystä koordinoivan kansainvälisen W3C-konsortion RDF (Resource Description Framework) ja siihen liittyvä laajennus RDF Schema (RDFS). RDF:llä kuvataan www:n sisältöjen, resurssien, merkityksiä. Esimerkiksi kirjastoalan Dublin Core –standardilla voidaan kuvata viidentoista ominaisuuden avulla dokumentteihin liittyvä yleinen metatieto, kuten dokumentin laatija, laatimisaika jne. RDF on ytimeltään yksinkertainen relaatiomalli, jossa tieto koostuu joukosta (objekti, ominaisuus, arvo) –kolmikoita. RDF:n syntaksin spesifikaatio perustuu XML:ään, mutta yhtä hyvin voidaan käyttää muunkinlaisia esityskieliä – rakenne ei ole olennaista vaan alla oleva semanttinen relaatiomalli.

RDFS:n avulla määritellään RDF-sovelluksissa käytettävä sanasto. Laajennus tuo olio-ajattelun idean www:n merkkauskieliin tarjoamalla käsitteiden kuvaamiseen joukon etukäteen sovittuja perusprimitiivejä. Class, subClass ja type-ilmausten avulla voidaan määritellä käsitehierarkioita samaan tyyliin kuin olio-ohjelmoinnissa.

RDF(S) näyttäisi olevan tulossa laajempaan käyttöön. Esimerkiksi Adobe on uuden XMP-alustan myötä siirtymässä RDF-standardiin kaikissa yhtiön www-tuotteissa.

Ontologiat ovat käsitehierarkioita

Semanttisen webin konseptin keskeisimpiä käsitteitä on ontologia. Käytännössä ontologiat ovat eri sovellusalojen terminologisia, formaaleja käsitehierarkioita, joissa määritellään alalla käytettävät termit ja käsitteet ja näiden välisiä suhteita. Ontologioita ovat mm. WordNet, joka sisältää yli 100.000 englannin kielen käsitettä, IT- ja elektroniikkateollisuuden RosettaNetin sanastot, CYC ja vireillä oleva IEEE:n SUO (Standard Upper Ontology) standardointihanke ontologisista yläkäsitteistä sopimiseksi. Ontologioiden esityskielten osalta suurimman huomion kohteena ovat olleet eurooppalainen OIL ja amerikkalainen DAML. Näiden yhdistelmästä on tarkoitus luoda W3C:n standardisuositus OWL, ihmisläheinen loogiikkaperustainen käsitehierarkioiden määrittelykieli.

Ontologiatason yläpuolelle ollaan kehittämässä vielä kieliä ja standardeja loogiseen päättelyyn sekä luottamuksen lisäämiseksi www:ssä. Esimerkiksi annotointivälineillä voi kommentoida sivuja siten, että toiset käyttäjät voivat paremmin arvioida verkon resurssien luotettavuutta. Digitaalisten allekirjoitusten (digital signature) avulla voidaan vakuuttua kuvausten toimittaneen tahon identiteetistä ja näin arvioida sisältöjen arvoa. P3P on W3C:n suositus, jonka avulla käyttäjät voivat hallita yksityisyyttään suositusta tukevilla www-sivustoilla.

”Semanttisen Webin yksinkertaisena ideana on esittää tiedot siten, että niiden sisältöön päästään helpommin algoritmisesti käsiksi.”

Semanttisen Webin sovellusalueita

Metakuvausten ja ontologiatekniikoiden tärkeitä sovellusalueita ovat tiedonhaku, tietämyksen hallinta, sekä verkkokauppa ja sähköinen liiketoiminta.

Metakuvausten avulla tiedon täsmähaku helpottuu. Esimerkiksi dmoz Open Directory Project kokoaa hajautetusti open source -hengessä hakukoneiden rinnalle koko www:n indeksoivaa hierarkkista RDF(S) metakuvausta. Tietämyksen hallinnan merkitys liittyy yritysten ja muiden organisaatioiden alati lisääntyvään tarpeeseen hankkia, ylläpitää, löytää ja hyödyntää omaa tietämystään kilpailuedun saamiseksi ja toimintojen tehostamiseksi. Haasteina ovat mm. tietovarastoissa olevien dokumenttien vapaamuotoisuus ja hajautus maapalloistumisen myötä sekä heterogeenisten järjestelmien yhteiskäyttö. Semanttisen webin teknologiat tarjoavat uusia välineitä näiden ongelmien ratkaisemiseen.

Verkkokauppa tarjoaa yrityksille sähköisen jakelu- ja markkinointikanavan, mikä puolestaan mahdollistaa uudentyyppisiä liiketoimintamalleja. Sähköisessä liiketoiminnassa keskeinen kehityskohde on mm. liiketoimintaan liittyvien transaktioiden hallinta. Yksi kehityskohde on tuote- ja palvelukuvaukset ja luettelot sekä näihin liittyvät hakemistopalvelut. Semanttisen webin teknologioilla tulee olemaan yhä tärkeämpi rooli verkkokaupan ja sähköisen liiketoiminnan järjestelmissä. Semanttisen webin välineet voivat täydentää tuotteiden, palveluiden ja prosessien kuvauksessa Web Service -alueen standardeja, kuten UDDI, WSDL, SOAP ja RosettaNet. Alan taloudellinen merkitys on suuri: esimerkiksi Nokian ilmoituksen mukaan jopa 40% sen alihankintaketjusta olisi siirtymässä käyttämään RosettaNet-standardia kuluvan vuoden kuluessa.

”... kun nykyinen XML-murros erottaa rakenteen ilmiasusta, erotetaan Semanttinen web –murroksessa merkitys rakenteesta.”

Tilanne maailmalla ja Suomessa

Semanttinen web nousi lehtiotsikoihin, kun W3C käynnisti helmikuussa 2001 Semantic Web Activity –ohjelman ja tammikuussa 2002 alkoi Web Services Activity. Euroopassa toimii OntoWeb -tutkimusverkosto alan tutkimuslaitosten, yritysten ja tutkijoiden yhteistyötä edistämässä. USA:ssa puolustusministeriön DARPA rahoittaa laajaa DAML-ohjelmaa. Suomessa Tietotekniikan tutkimuslaitos HIIT ja Helsingin yliopisto yhteistyökumppaneineen järjestivät syksyllä 2002 erityisen Semantic Web Kick-Off in Finland -tilaisuuden alan kehitystyön edistämiseksi maassamme. Lokakuun lopussa 2002 Helsingissä pidetään kaksipäiväinen kansainvälinen seminaari ”XML Finland 2002: Towards the Semantic Web and Web Services”.

Lue lisää

E. Hyvönen (ed.): Semantic Web Kick-Off in Finland. Vision, Technologies, Research, and Applications. HIIT Publications 2002-01, HIIT, Helsinki, 2002, 289pp. URL: http://www.cs.helsinki.fi/u/eahyvone/stes/semanticweb/kick-off/proceedings.html