

Semanttinen Web – Mitä se on käytännössä?

Eero Hyvönen

Helsingin yliopisto, tietojenkäsittelytieteen laitos ja

Tietotekniikan tutkimuslaitos HIIT

Semanttisen laskennan tutkimusryhmä

<http://www.cs.helsinki.fi/group/seco/>

eero.hyvonen@cs.helsinki.fi

Suomalainen 1.000.000 euron Millenium-teknologiapalkinto myönnettiin huhtikuussa World Wide Webin keksijälle, professori Tim Berners-Leelle. Sir Tim johtaa nykyisin Webin kehitystä koordinoivaa kansainvälistä World Wide Web Consortium –järjestöä W3C ja kehittää itse aktiivisesti WWW:n seuraavaa sukupolvea, semanttista webiä. Aihetta tutkitaan myös Helsingin yliopiston tietojenkäsittelytieteen laitoksen ja Tietotekniikan tutkimuslaitos HIIT:n Semanttisen laskennan tutkimusryhmässä (<http://www.cs.helsinki.fi/group/seco/>). Ryhmä sai W3C:ltä kutsun tulla esittelemään Suomessa maaliskuussa 2004 koekäyttöön otettua sovellusta ”MuseoSuomi – Suomen museot semanttisessa webissä” (<http://museosuomi.cs.helsinki.fi>) World Wide Web 2004 konferenssiin New Yorkiin. Tässä artikkelissa kerrotaan MuseoSuomi esimerkkinä, mistä semanttisessa webissä on kyse käytännössä.

Kohti koneiden webiä

Semanttinen web on seuraavan polven älykäs WWW, jonka käyttäjinä ovat ihmisten ohella koneet. Ajatuksena on tallettaa tieto webiin sellaisessa hyvin määritellyssä muodossa, että koneet kykenevät ymmärtämään tiedon merkityksen eivätkä ainoastaan välittämään tietoa, kuten nykyisessä webissä. Jos tietoa voidaan koneellisesti tulkita, tulee mahdolliseksi ohjelmoida aiempaa olennaisesti älykkäämpiä sisältöperustaisia web-palveluja. Semanttisen webin toinen keskeinen sovelluskenttä on tietojärjestelmien keskinäisen yhteentoimivuuden (interoperability) parantaminen. Kun järjestelmät voivat itse ymmärtää toistensa sisältöjä, voivat ne myös vaihtaa ja yhdistellä tietoja keskenään yhteismitallisella tavalla.

Ihmiselle semanttinen web näyttäytyy ainakin alkuvaiheessaan saman oloisena kuin nykyinen web. Tiedonhakuun on tarjolla toisaalta hakukoneita, toisaalta mahdollisuus samoilla sivustoissa linkkejä seuraten. Erotuksena aiempaan on kuitenkin näiden palveluiden ”semanttisuus”. Hakupalveluiden osalta tämä merkitsee esimerkiksi sitä, että haku voidaan tehdä hakusanojen sijasta sanojen taustalla oleviin käsitteisiin perustuen, jolloin tiedonhaussa päästään parempaan saantoon ja tarkkuuteen. Samoilussa taas voidaan linkkien avulla yhdistellä tietoja näiden välisten merkityssuhteiden mukaan, jolloin navigointi webissä helpottuu ja tulee entistä mielekkäämmäksi.

XML ei riitä merkitysten kuvailuun

Webin *lingua franca* on XML. Se on standardi, jolla voidaan määritellä sovelluskohtaisia merkkäuskieliä tietojen esittämiseksi. XML:n avulla voidaan määritellä syntaksi (rakenne), mutta semantiikan (merkityksen) kuvailuun ja erottamiseen syntaksista tarvitaan uusia välineitä. Semanttisen webin peruskieli on W3C:n standardoima RDF (Resource Description Framework) ja siihen liittyvä RDF Schema (RDFS). RDF:llä kuvataan WWW:n sisältöjen, resurssien, merkityksiä eli metatietoa. RDF Scheman avulla taas määritellään RDF-sovelluksissa käytettävä sanasto. RDF:n syntaksi perustuu yleensä XML:ään, mutta yhtä hyvin voidaan käyttää muunkinlaisia esityskieliä, kuten Tim Berners-Leen esittämää N3 syntaksia. Rakenne ei ole olennaista vaan alla oleva semanttinen relaatiomalli.

RDF(S) tuo ohjelmoinnista tutun olio-ajattelun idean WWW:n merkkäuskieliin tarjoamalla käsitteiden kuvaamiseen joukon kuvausprimitiivejä. Näiden avulla voidaan määritellä eri sovellusalueiden käsitteitä, näiden muodostamia luokkahierarkioita, ominaisuuksia ja arvorajoitteita.

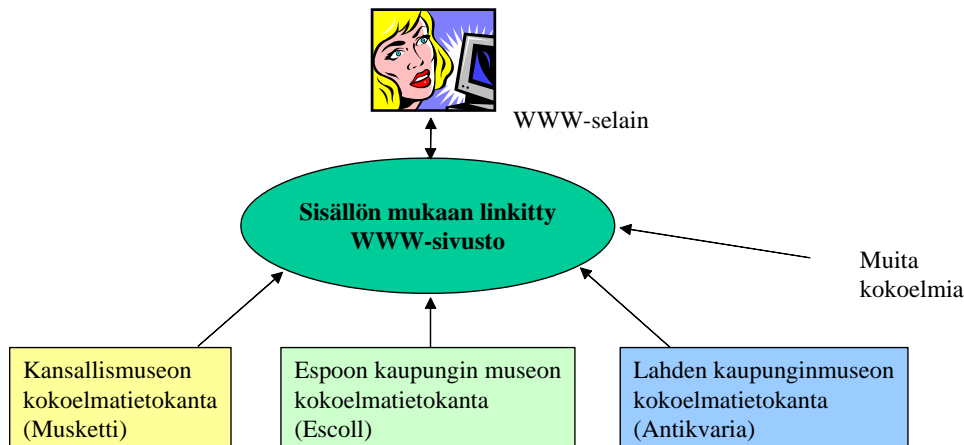
Tällaisia kuvauksia kutsutaan ontologioiksi. Yksinkertaiset ontologiat voidaan kuvata suoraan RDFS:n avulla. W3C on keväällä standardoinut (recommendation) myös kehittyneemmän RDF(S)-perustaisen ontologiakielen OWL (Web Ontology Language), joka on logiikkaperustainen käsittehierarkioiden määrittelykieli.

MuseoSuomi – Suomen museot semanttisessa webissä

MuseoSuomi-sovellus havainnollistaa monia edellä esiteltyjä semanttisen webin ydinideoita. Lähtökohtana järjestelmän kehittämiseksi oli havainto siitä, että web voi tarjota sekä kansalaisille että tutkijoille uudenlaisen, ajasta ja paikasta riippumattoman kanavan Suomen museokokoelmiin tutustumista ja tiedonhankintaa varten. Ongelmana kuitenkin oli, miten eri tietokannoissa olevat kokoelmatiedot voidaan yhdistää ja miten web-sivustoista saadaan kävijöiden kannalta aidosti mielenkiintoisia, hyödyllisiä ja helppokäyttöisiä.

MuseoSuomen tavoitteet

MuseoSuomen pilottiversioon sisältyy n. 4.000 kokoelmakohteen tiedot Kansallismuseon, Espoon kaupunginmuseon ja Lahden kaupunginmuseon kulttuurihistoriallisista esinekokoelmista. Lisäksi järjestelmään on yhdistetty Kansallismuseon muinaismuistorekisteri. Sisältöjen yhdistäminen perustuu hankkeessa kehitettyihin seitsemään käsittehierarkiaan, ontologiaan, joissa on n. 10.000 yleis- ja yksilökäsitteitä.



Kuva 1. MuseoSuomi museovieraan näkökulmasta.

MuseoSuomen keskeiset ideat järjestelmän loppukäyttäjän ja museoiden kannalta voidaan kiteyttää seuraaviin kohtiin.

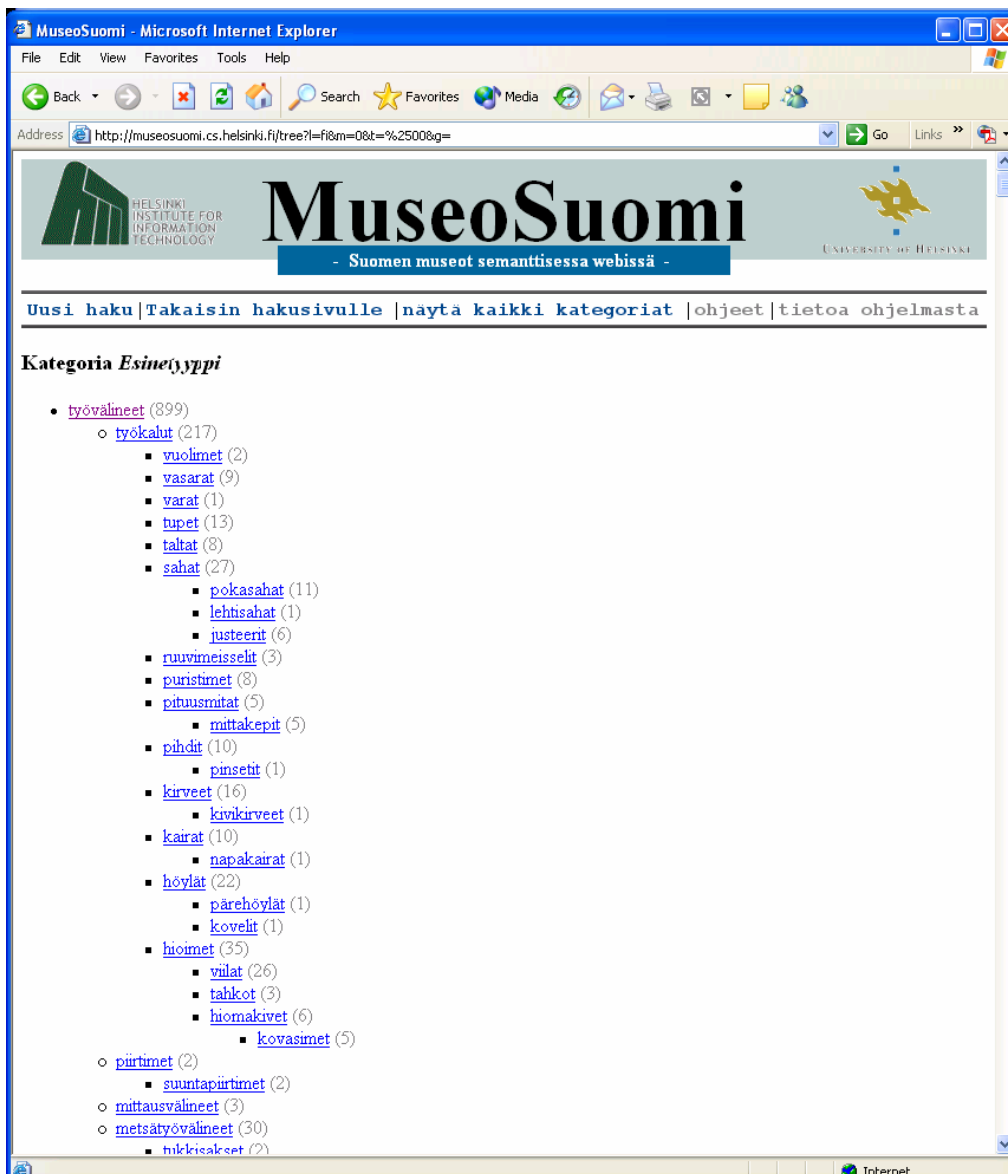
1. *Suomen museoiden saumaton yhteiskokoelma webiin.* MuseoSuomi-konseptin kunnianhimoisena tavoitteena on kaataa keinotekoiset raja-aidat sekä eri paikoissa ja erilaisissa tietojärjestelmissä sijaitsevien kokoelmien että sisällöltään erityyppisten kokoelmien välillä. Loppukäyttäjälle eri kokoelmat näkyvät yhtenäisenä semanttisesti linkitettyinä web-sivustona, johon voidaan tehdä hakuja ja jota voidaan selailta esineiden välisiä assosiaatioita seuraten (kuva 1).
2. *Käsiteperustainen tiedonhaku.* MuseoSuomessa tiedon haku ei perustu vain hakusanoihin vaan ennen kaikkea niihin liittyviin käsitteisiin. Esimerkiksi yhdellä hakusanalla ”päähine” voidaan löytää erityyppiset päähineet, kuten ”silinterit”, ”ylioppilaslakit” tai ”harakat” ilman että käyttäjän tarvitsee pystyä luettelemaan näitä sanoja erikseen.
3. *Semanttinen suosittelu linkeillä.* MuseoSuomessa tiedon haku ei kohdistu vain kokoelmaesineisiin vaan tavoitteena on samalla paljastaa käyttäjälle, miten eri kokoelmakohteet liittyvät toisiinsa. Yhteydet näkyvät käyttäjälle nimettyinä linkeinä, joita painamalla voi siirtyä kohteesta toiseen kuten webissä on tapana. Esimerkiksi rukiinlapaa kuvaavaan web-sivuun voidaan automaattisesti tuottaa linkit toisiin esinesivuihin, jotka

liittyvät kehräämiseen tai morsiuslahjoihin. Linkin luonne käy ilmi sen nimestä. Näin voidaan luoda semanttisesti linkitettyjä temaattisia kokonaisuuksia.

4. *Kansallisen julkaisukanavan luominen.* MuseoSuomi on demonstraatio julkaisukanavasta, jonka kautta maamme eri museot voisivat julkaista yhdessä kokoelmiaan webissä sisällöllisesti yhteismitallisella tavalla. Ideana on muodostaa kansalaisille webiin virtuaalisesti yksi yhtenäinen kansallinen museokokoelma, MuseoSuomi, joka sisältää eri museoista erityyppisiä aineistoja, kuten esinekokoelmia, taidetta ja maastossa olevia kulttuurikohteita.

Tiedonhaku MuseoSuomessa

MuseoSuomen käyttäjälle on tarjolla kolme tapaa hakea tietoa: *näkymähaku*, *käsitehaku* ja *semanttinen samoilu*.



Kuva 2. Esinetyyppien hierarkkinen näkymä. Näkyvissä on osa ”työvälineet” luokittelua linkkeinä.

Näkymähaku

Näkymällä tarkoitetaan kokoelma-aineiston hierarkkista luokittelua jonkin näkökulman, kuten esinetyypin, materiaalin, valmistuspaikan, käyttötilanteen tms. suhteen. MuseoSuomessa on käytössä 9 eri näkymää jaoteltuna esinetietoihin, valmistustietoihin, käyttötietoihin ja kokoelmatietoihin.

Kuvassa 2 on esitetty esimerkkinä osa MuseoSuomen esinetyypinäkymästä. Luokittelun kategoriat ovat hakukoneen käynnistäviä linkkejä ja kunkin (ali)kategorian perässä oleva luku kertoo niihin kuuluvien esineiden lukumäärän kokoelmassa. Esimerkiksi linkkiä ”taltat (8)” painamalla hakutulokseen tulee kahdeksan mahdollisesti eri kokoelmassa olevaa talttaa. Näkymähierarkian ideana on antaa käyttäjälle yleiskuva museoaineiston jakautumisesta eri luokkiin yhden näkymän suhteen ja mahdollisuus suorittaa haku minkä tahansa kategorian suhteen.

Näkymähaussa mistä tahansa näkymästä voidaan aina valita haluttu alikategoria vastaavaa linkkiä painamalla. Järjestelmä hakee silloin kaikki samanaikaisesti eri näkymien valintoja vastaavat esineet ja tulostaa tuloksen. Esimerkiksi valitsemalla esinetyypiksi ”astiat ja taloustarvikkeet” ja sitten valmistuspaikaksi ”Pohjois-Amerikka” saadaan tulokseksi kaikki Pohjois-Amerikassa valmistetut astiat ja taloustarvikkeet. Jokaisen valinnan jälkeen järjestelmä laskee odotettavissa olevien hakutulosten määrän jokaisen mahdollisen jatkovalinnan suhteen ja näyttää määrät suluissa vastaavan linkin perässä. Jos kategorian valinta johtaa tyhjiin joukkoon, estetään valinnan tekeminen kokonaan. Näin voidaan estää se, että käyttäjä joutuisi sellaisiin turhauttaviin tilanteisiin, joissa haku ei tuota lainkaan osuvia.

Näkymähaku antaa käyttäjälle sanaston, jonka avulla hän osaa järkevästi ja oikein esittää hakukyselyjä järjestelmän tuntemien käsitteiden avulla. Tämä on hyödyllistä silloin, kun käyttäjä ei osaa muotoilla täsmällisesti tiedontarvettaan hakusanoina. Tämä ongelmatilanne on tyypillinen perinteisiä kokoelmaselaimia käytettäessä.

The screenshot shows the MuseoSuomi website interface. The main content area displays a search result for a coffee mill. On the left, there is a photograph of the mill, labeled 'Kuva'. To the right of the photo is a list of 'Metatiedot' (Metadata) including name, material, production time, and manufacturer. Further right, there are sections for 'Suositukset' (Recommendations) and 'Näkymäluokitukset' (View Classifications), which are lists of related items and classification terms respectively. The website header features the MuseoSuomi logo and navigation links.

Kuva 3. Jokaiselle museokohteelle on oma web-sivunsa, jossa näkyy valitun kohteen kuva, metatiedot (keskellä), semanttiset suositukset toisiin kohteisiin (oikealla) ja ontologist näkymäluokitukset (alaosassa).

Käsithaku

MuseoSuomeen on toteutettu myös käyttäjän kannalta perinteisen sanahaun kaltaisesti toimiva hakukone. Se hakee perinteiseen tapaan kaikki kokoelmakohteet, joiden kuvailussa annettu hakusana kuten ”Helsinki” esiintyy. MuseoSuomessa tämä hakutapa hyödyntää kuitenkin myös näkymien kategorioita vastaavien ontologisten käsitteiden verkostoa. Esimerkiksi hakusana ”Helsinki” sopii valmistuspaikkanäkymän ”Helsinki” kategorian nimeen. Paikkaontologian avulla haku voidaan

kuitenkin ulottaa myös kohteisiin, joiden valmistuspaikka on jokin Helsingin osa, kuten ”Pasila” tai ”Punavuori”.

Semanttinen samoilu

Näkymä- ja käsitehaun avulla käyttäjä voi löytää itseään kiinnostavan, mielekkäästi rajatun kohdejoukon. Tämän jälkeen hän voi hiirellä valita hakutuloksesta jonkin esineen tarkempaa tutustumista varten, jolloin MuseoSuomi näyttää kohteen kuvan ja metatiedot omalla *kohdesivullaan* (kuva 3). MuseoSuomen yhtenä innovaationa on tässä yhteydessä tarjota käyttäjälle ”semanttinen selailumahdollisuus” assosiatiivisten suositteulinkkien avulla (kuvassa 4 oikeassa reunassa). MuseoSuomi voi esimerkiksi suositella linkkejä saman henkilön käyttämiin muihin esineisiin tai samaa tapahtumatyyppiin liittyvään esineistöön. Suositukset perustuvat ontologioihin ja metatietojen semanttisiin yhteyksiin. Näihin perustuen MuseoSuomen kehittäjä voi määrittellä esineitä eri tavoin yhdistäviä loogisia sääntöjä, jotka sitten näkyvät loppukäyttäjälle nimettyinä linkkeinä.

Kohteen (meta)tietojen ja linkkisuositusten lisäksi kohdesivun alareunassa näytetään kohteeseen liittyvät ontologiset luokittelut linkkeinä, joita painamalla käynnistyy uusi kategorialla vastaava näkymähaku.

Sisältöjen vieminen semanttiseen webiin asettaa uusia haasteita tiedon tallennusvaiheelle. Jotta semanttisesti yhdistettyjä sisältöjä voitaisiin jatkossa tuottaa automaattisesti tietokannoista, pitäisi tietojen tallentamisen yhteydessä ryhtyä tallentamaan asiasanojen, objektin nimien, paikkojen, toimijoiden jne. nimien ohella myös viittaus taustalla oleviin käsitteisiin. Vain näin voidaan tehdä ero Suomen eri mattivirtasten ja pyhäjärvien välillä tai erottaa toisistaan vaikkapa luettelointitieto ”harakka” lintuna, päähineenä, myllynä ja saarena Helsingin edustalla.

Jatkossa MuseoSuomeen tullaan lisäämään uudentyyppisiä ontologioita ja aineistoja, kuten Ateneumin taidemuseon ja Kiasman taidetta, Valokuvataiteen museon valokuvia, Yleisradio Oy:n videoaineistoa jne. osana Tekesin kansallista ”Suomalaiset semanttisen webin ontologiat” – tutkimusprojektia 2003–2005 (<http://www.cs.helsinki.fi/group/seco/ontologies/>). Hankkeessa on mukana peräti 19 julkisen hallinnon organisaatiota ja yritystä.

Lisätietoja

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>

MuseoSuomi-hankkeen kotisivu:

<http://www.cs.helsinki.fi/group/seco/museums/>

W3C:n Semantic Web ym. aktiviteetit

<http://www.w3.org>

Tutkimusmaailman portaali:

<http://www.semanticweb.org>

EU:n yhteistyöverkosto OntoWeb:

<http://www.ontoweb.org>

Semantic Web Suomessa:

<http://www.cs.helsinki.fi/u/eahyvone/stes/semanticweb/>

Seuraava Suomessa pidettävä laajempi konferenssi semanttiseen webiin liittyen on Suomen Tekoälypäivien 2004 yhteydessä Tiedekeskus Heurekassa 2.9.2004 järjestettävä seminaari ”Web Intelligence – Älyä verkossa” (<http://www.cs.helsinki.fi/group/seco/conference/step2004/>)