Semanttinen data verkkokaupoissa

Kristian Mattila

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Pro gradu -tutkielma
Ohjaajat: Erkki Mäkinen ja
Marko Junkkari
Kesäkuu 2015

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Kristian Mattila: Semanttinen data verkkokaupoissa
Pro gradu -tutkielma, 50 sivua
Kesäkuu 2015

Semanttisen webin tavoitteena on muuntaa dokumentteihin painottunut World Wide Web verkoksi, jossa data asetetaan etusijalle. Tällä hetkellä yli viidentoista vuoden aikana kehitetyt semanttisen webin teknologiat alkavat saada jalansijaa WWW:n valtavirrassa, esimerkiksi hakukoneyhtiöt ovat määritelleet standardin niiden tukemalle semanttiselle datalle.

Tutkielmassa tarkastellaan verkkokauppoihin soveltuvia semanttisen webin tekniikoita ja niiden käyttämisen tuomia hyötyjä. Semanttisen webin tekniikoista käsitellään lähinnä verkkokaupan tuotteiden kuvailemiseen soveltuvia määrittelyitä. Verkkokauppojen hakukonenäkyvyyden parantaminen otetaan lähtökohdaksi semanttisen datan tuottamisen käsittelyyn. Lisäksi pohditaan verkkokaupoissa tuotetun semanttisen datan käyttömahdollisuuksia verkkokaupankäyntiin liittyvissä palveluissa.

Avainsanat ja -sanonnat: semanttinen web, metadata, ontologiat, WWW, verkkokauppa, hakukoneoptimointi, Microdata, GoodRelations, schema.org.

Sisällys

1 Johdanto

Verkkokauppamyynti on kasvanut viime vuosina kiihtyvää vauhtia (SVT, 2014). Verkkokaupan osuus kaikesta käydystä kaupankäynnistä kasvaa tasaisesti vuosittain, ja samaan aikaan perinteiset myyntikanavat menettävät markkinaosuuttaan (Mäntymaa, 2015). Verkkokaupan kasvu on osa suurempaa digitalisaation ja globalisaation aiheuttamaa rakenneuudistusta, minkä seurauksena yhä useammat palvelut siirtyvät verkkoon.

World Wide Web (WWW) on maailmanlaajuinen verkko, joka sisältää valtavan kokoelman WWW-sivustoja. WWW:n sisällölle ei kuitenkaan ole olemassa yhtä kaiken sisällön kattavaa hakemistoa, jonka avulla käyttäjät voisivat löytää etsimänsä. Hakukoneet ovat ratkaisu tähän WWW:n rakenteelliseen ominaisuuteen. Hakukoneet jäsentävät WWW:ssä julkaistua informaatiota ja mahdollistavat hakujen tekemisen niiden indeksoimaan WWW:n sisältöön.

Kuluttajat käyttävät hakukoneita osana ostoprosessiin liittyvää tiedonhakua, minkä vuoksi verkkokauppojen hakukonenäkyvyys on olennainen osa toimivan markkinointistrategian kehittämistä. Hakukoneet käyttävät informaation jäsentämiseen algoritmeja, joiden toimintaan voidaan vaikuttaa tarjoamalla WWW-sivun sisältöä kuvailevaa metadataa. Yksinkertaistettuna metadatan voidaan määritellä olevan tietoa tiedosta. Tarkemmin määriteltynä metadata on tietosisältöä kuvaavaa ja sen merkitystä selittävää jäsennettyä informaatiota, jonka käyttämisellä pyritään helpottamaan tietosisällön hakua, käyttöä ja säilyttämistä (Guenther & Radebaugh, 2004). Metadata-termiä käytetään useissa eri ympäristöissä ja sen merkitys voi olla niissä erilainen; termillä voidaan esimerkiksi tarkoittaa koneellisesti luettavaa informaatiota tai aineistoa kuvaavia tietueita – määritelmistä ensimmäistä käytetään tietojenkäsittelyssä ja jälkimmäistä kirjastoissa (ibid).

Semanttinen web (Semantic Web) on Tim Berners-Leen ja kumppaneiden (2001) idea muuttaa dokumenttikeskeinen World Wide Web verkoksi, jossa data nostetaan etusijalle. Semanttisen webin teknologiat määrittelevät yhteiset merkintätavat metadatalle, jolla voidaan kuvailla WWW:ssä julkaistua informaatiota. Perinteisiä hakukoneissa käytettyjä tiedonhakumenetelmiä ovat sanojen täystäsmäytys- ja osittaistäsmäytysmenetelmät; uudempia sisällön arviointiin käytettyjä menetelmiä ovat linkkirakenteen analyysi ja metadatan hyödyntäminen. Vuonna 2009 Google, maailman käytetyin hakukone, ilmoitti tukevansa semanttista dataa (Goel et al., 2009). Hakukoneyhtiöt ovat kehittäneet metadatan merkintään omia standardeja, jotka pohjautuvat semanttisen webin teknologioihin. Näitä semanttisen

datan määrittelyjä käyttämällä verkkokauppiaat voivat parantaa WWW-sivustojensa hakukonenäkyvyyttä.

Tutkielman tavoitteena on selvittää verkkokaupan tuotteiden kuvailemiseen soveltuvia semanttisen datan määrittelyitä sekä arvioida niiden käytön hyötyjä. Tutkielma etenee niin, että toisessa luvussa tarkastellaan yleisesti semanttisen webin osa-alueita: WWW:n teknistä rakennetta, semanttisen datan merkintätapoja ja ontologioita. Kolmannessa luvussa käsitellään sähköistä kaupankäyntiä, jonka tarkastelussa pääpaino asetetaan verkossa tapahtuvaan kuluttajamyyntiin. Verkkokaupankäynnin yhteydessä käsitellään siihen liittyvää monikanavaisuutta ja hakukoneiden merkitystä verkkokaupoille. Neljännessä luvussa tarkastellaan semanttisen webin teknologioita, jotka soveltuvat verkkokauppojen yhteydessä käytettäviksi; tarkastelun lähtökohdaksi otetaan verkkokauppojen hakukonenäkyvyyden parantaminen semanttisen datan tuottamisella. Lopuksi työn havainnoista tehdään yhteenveto, esitetään aiheita jatkotutkimukselle ja pohditaan semanttisen datan tarjoamia mahdollisuuksia.

2 Semanttinen web

Tim Berners-Lee ja kumppanit (2001) esittelivät idean semanttisesta webistä, jonka keskeisin tavoite on mahdollistaa World Wide Webissä olevan informaation koneellinen tulkinta. Semanttisen webin tavoitteena on myös mahdollistaa tietokoneohjelmien tuottaminen, jotka voivat auttaa ihmisiä hyödyntämään WWW:ssä julkaistua informaatiota. Jotta informaatiota voidaan tulkita koneellisesti, tarvitaan yhteisiä, datan kuvailemista määritteleviä sopimuksia. Laajasti käytetyt yhteiset standardit johtavat myös järjestelmien välisen kommunikaation helpottumiseen; artikkelissaan Berners-Lee kollegoineen esittelevät idean mahdollisesta laitteiden välisestä integraatiosta, esimerkiksi stereolaitteisto voisi olla tietoinen samassa huoneessa olevasta puhelimesta ja osaisi puhelun tullessa säätää äänenvoimakkuutta pienemmälle. (Berners-Lee et al., 2001.)

Semanttisen webin on osaltaan tarkoitus olla ratkaisu vanhassa WWW:ssä havaittuihin puutteisiin. Ensinnäkin HTML-dokumenttien käsittely tekstinä luo ongelman, koska monitulkintaisista termeistä ei voida tietää, mitä niillä tarkoitetaan. Esimerkiksi ”Paris” voi tarkoittaa kaupunkia Ranskassa tai Kanadassa, ja se voi olla myös henkilön etunimi. Tällä hetkellä hakukoneet ratkaisevat asian jättämällä päättelyn käyttäjälle ja esittävät hakuun ”parhaimmin” sopivat tulokset. Toiseksi WWW-sivuilla esitetty informaatio on usein tallennettuna tietokantoihin ja sen pohjalta luodaan WWW-selaimelle esitetty HTML-dokumentti. Tämä toimintatapa muodostuu ongelmaksi, kun HTML-dokumentin rakenne muuttuu, esimerkiksi WWW-sivuston uudistamisen yhteydessä. Ongelma on, että HTML-dokumentin rakenteen vaihtuessa informaation esitysmuoto vaihtuu, vaikka silti kyse on sisällöllisesti samasta informaatiosta. Kolmanneksi ongelmaksi on muodostumassa WWW:ssä julkaistun informaation määrä, kun sen tulkinta jätetään käyttäjälle, jolla on työkalunaan ainoastaan WWW-selain; informaation hyödyntäminen onnistuisi tehokkaammin, jos sitä voitaisiin tarkemmin esikäsitellä koneellisesti. Semanttinen web pyrkii ratkaisemaan nämä ongelmat asettamalla datan ja datan merkityksen etusijalle, jolla se pyrkii asettamaan uuden suunnan vanhalle dokumentteihin keskittyneelle WWW:lle. (Domingue et al., 2011.)

Datan ja sen merkityksen etusijalle asettamisella saavutetaan samalla datan helpompi uudelleenkäytettävyys. Vanhassa WWW:ssä tarjolla oleva informaatio on yleensä HTML-muodossa, joka on luotu WWW-sivuston taustajärjestelmien käsittelemästä, yleensä tietokantaan tallennetusta, datasta. Informaation uudelleenkäyttäminen vanhassa WWW:ssä tarkoittaa yleensä HTML-dokumenttien jäsentämistä tai tiettyä palvelua varten määritellyn

rajapinnan käyttämistä. Informaation vaikea uudelleenkäytettävyys rajoittaa sen käyttäjäkuntaa ja informaatio jääkin usein vain suhteellisen pienen joukon käyttöön. (Adida et al., 2011.)

Semanttisen webin onnistumiselle voidaan asettaa viisi edellytystä, vaatimusta, jotka sen rakenteen pitää täyttää. Ensinnäkin järjestelmän tulee olla hajautettu, jotta uuden teknologian käyttöönotto olisi joustavaa. Hajautetulla järjestelmällä voidaan mahdollistaa sen osien asteittainen siirtyminen käyttämään uutta teknologiaa sekä järjestelmän helppo laajennettavuus, kun laajentumista ei rajoiteta keskitetyllä hallinnalla. Dataa tulee voida julkaista järjestelmässä ilman keskitettyä sisällönhallintaa ja datan tuottajilla pitää olla täysi omistajuus ja hallinta tuotettuun dataan. Toiseksi datalle pitää olla eksplisiittinen ja samalla yksinkertainen esitysmuoto, joka täyttää teknologian päävaatimukset ja samalla piilottaa teknologiaan liittyvää monimutkaisuutta. Kolmanneksi dataa tulee pystyä linkittämään, jotta järjestelmä ei ole vain kokoelma yksittäisiä datalähteitä. Datan linkittäminen mahdollistaa myös datan uudelleenkäyttämisen ja yksinkertaistaa uuden datan tuottamista. Neljänneksi datan julkaisemisen ja käytön tulee olla helppoa – esitysmuoto ei kuitenkaan saa rajoittaa datan monimutkaisuutta. Viidenneksi järjestelmässä tulee käyttää datan siirtämiseen standardeja, jotka mahdollistavat datan siirtämisen järjestelmän osien välillä, mutta eivät kuitenkaan rajoita siirretyn datan sisältöä. (Harth et al., 2011.) Seuraavissa kohdissa käydään läpi ratkaisuja edellä esitettyihin semanttisen webin vaatimuksiin.

2.1 World Wide Web

World Wide Webin kehityshistoria alkoi Euroopan hiukkasfysiikan tutkimuskeskuksessa (CERN) 1980-luvun lopulla. CERN:issä työskenteli tuolloin tuhansia tutkijoita ja siellä tehdyn tutkimuksen luonteen takia kirjoitettu tieto oli usein jo vanhaa, joten uusin tarvittava tieto jaettiin keskustelemalla. Ongelmaksi muodostui tarvittavan tiedon saatavuus: osittain suuri vaihtuvuus henkilökunnassa aikaansai sen, ettei tieto ollut enää saatavilla suullisesti, ja osittain ongelmana oli myös tiedon häviäminen kadonneiden kirjallisten dokumenttien mukana. (Berners-Lee, 1989.)

World Wide Web syntyi Tim Berners-Leen ehdotuksesta ratkaisuna CERN:in dokumentointiongelmaan. Järjestelmän oli tarkoitus mahdollistaa ajantasaisen tiedon jakaminen ja helppo päivitettävyys. Berners-Lee oli ennen CERN:iin tuloa luonut omaan käyttöönsä Enquire-nimisen järjestelmän, johon hän pystyi tallentamaan tekstiä ja linkittämään toisiinsa liittyviä tietoja. Kokemus Enquire-järjestelmästä ohjasi Berners-Leen

suunnittelemaan järjestelmän hyperteksti-ajatuksen ympärille, jossa teksti sisältää korostettuja sanoja tai alueita, joista lukija voi siirtyä viitattavaan tekstiin. Hän ajatteli, että hypertekstiä käyttämällä pystyisi toteuttamaan järjestelmän, jonka laajennusmahdollisuuksia ei olisi rajattu. Ehdotuksessaan Berners-Lee mainitsee ”hypermedian” todetakseen, ettei käyttö rajoitu vain tekstiin, vaan sisältönä voi olla myös kuvia, ääntä ja videota. (Berners-Lee, 1989.)

Berners-Lee listasi ehdotuksessaan järjestelmän vaatimukset: Ensinnäkin tallennetun tiedon oli oltava luettavissa tietoverkon yli sijainnista riippumatta, koska CERN toimi useissa eri maissa. Toiseksi järjestelmän piti olla alustariippumaton, koska tutkijat käyttivät erilaisia käyttöjärjestelmiä, ja tallennettu tieto piti pystyä lukemaan käyttäjän tietokoneen tyypistä riippumatta. Kolmanneksi järjestelmän piti olla hajautettu, koska Berners-Lee halusi välttää keskitetyn hallinnan huonot puolet ja mahdollistaa järjestelmän helpon laajentamisen. Hypertekstin kannalta tärkein vaatimus oli pystyä linkittämään tekstiä siten, että linkkiä seuraamalla saataisiin aina uusin versio linkin osoittamasta materiaalista. (Berners-Lee, 1989.)

Ehdotuksessaan Berners-Lee arvioi, että järjestelmän suunnitteluun menisi kahdelta henkilöltä kuudesta kahteentoista kuukautta ja sen jälkeen järjestelmää voitaisiin alkaa toteuttaa. Samassa hän mainitsee, että projekti olisi hyvä kohde uusien olio-pohjaisten ohjelmointitekniikoiden kokeilemiseen (Berners-Lee, 1989). CERN:in historiasta kertova sivusto mainitsee, että maailman ensimmäinen WWW-palvelin julkaistiin vuoden 1990 joulukuussa. Palvelimen jakamalla WWW-sivustolla[1] kerrottiin WWW:n perustekniikoista ja se sisälsi myös ohjeet oman palvelimen asentamiseen. (CERN, 2012)

Vuonna 2012 WWW:ssä oli 634 miljoonaa sivustoa, joista 51 miljoonaa oli lisätty vuoden aikana (Pingdom, 2013). Miljardien käyttäjien mediaksi WWW:n arkkitehtuuri on yllättävän yksinkertainen. Toisaalta, rakenteen yksinkertaisuus voi olla yksi tärkeimmistä tekijöistä, jonka vuoksi WWW on laajentunut niin suureksi. Toinen merkittävä tekijä WWW:n käytön yleistymiseen on internetin kasvaminen maailmanlaajuiseksi verkoksi.

Internetin kehitys alkoi kylmän sodan aikana ja sen alkuhetkeksi voidaan laskea Paul Baranin 1960-luvulla julkaisema tutkimus hajautetuista viestintäverkoista. Verkkojen verkon oli tarkoitus suojata Yhdysvaltojen viestintäyhteydet Neuvostoliiton mahdolliselta ydiniskulta. Vuonna 1969 ARPANET-projekti (Advanced Research Projects Agency Network) laajentui verkoksi, kun tietoverkkoon yhdistettiin kolme palvelinta, joita verkon käyttäjät pystyivät

1 http://info.cern.ch/hypertext/WWW/TheProject.html

käskyttämään – tämän toimintatavan pohjalta syntyi myöhemmin kaupallisia asiakas/palvelin -järjestelmiä. ARPANET ei kuitenkaan ollut ongelmaton, vaan palvelimien liittäminen verkkoon vaati työlästä ohjelmointia. Vuosien 1973 ja 1974 aikana syntyi nykyisen internetin toiminnan kannalta kriittinen TCP/IP-protokolla. Uuden verkkoprotokollan avulla tietoverkkoon pystyttiin liittämään koneita käyttöjärjestelmästä riippumatta – tätä innovaatiota voidaan pitää internetin perustuskivenä. (Steinbock, 1997.)

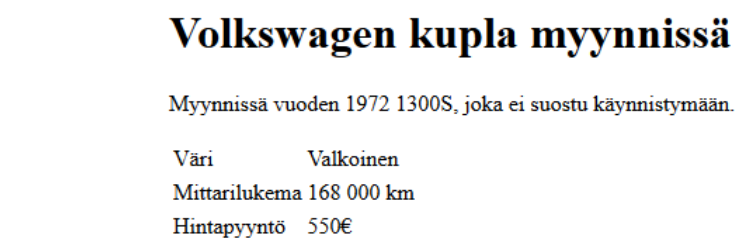
Verkkojen verkon palvelimien määrä kasvoi 80-luvun taitteessa nopeasti, kun ARPANET:iin liitettiin useita korkeakouluja, Yhdysvaltojen energiaministeriö ja NASA. Samalla ARPANET-projektin luonne alkoi muuttua, koska sen vetovastuu siirtyi Yhdysvaltojen puolustusvoimilta akateemisille yhteisöille. Projektin akateemisena aikakautena kehitettiin useita internetin peruspalveluita, esimerkiksi DNS (Domain Name System), FTP-tiedonsiirtoprotokolla ja sähköposti. DNS-järjestelmä yksinkertaisti verkon käyttöä huomattavasti, kun käyttäjän ei tarvinnut enää muistaa palvelimen IP-osoitetta vaan muunnos palvelimen nimestä (verkkotunnuksesta) IP-osoitteeseen tehtiin automaattisesti – olihan verkossa jo yli 1000 palvelinta. Vuonna 1986 Yhdysvaltojen kansallinen tiedesäätiö (National Science Foundation, NSF) lanseerasi NSFNET-verkon, jonka tarkoitus oli palvella niitä tiedeyhteisöjen jäseniä, jotka eivät olleet mukana ARPANET:issä. NSF:n verkkoprojektista muodostui myöhemmin yliopistojen ja tutkimuslaitosten alueverkkoja yhdistävä runkoverkko. Runkoverkon kasvava suosio teki nopeasti ARPANET:istä tarpeettoman ja vuonna 1989 se suljettiin. 1990-luvun alussa tietoverkkomarkkinoille lanseerattiin CIX-markkinat (Commercial Internet Exchange), joka kiihdytti tietoverkon kasvua – edessä oli internetin muodonmuutos, joka nopeutti sen yksityistymistä ja kaupallistumista. (Steinbock, 1997.)

WWW:n arkkitehtuuri on rakennettu internetin verkkotekniikoiden päälle; palvelimet tunnistetaan DNS-tietojen perusteella ja datan siirtämiseen käytetään TCP/IP-protokollaa. WWW:n tarpeisiin on suunniteltu oma sovellustason protokolla, nimeltään HTTP (Hypertext Transfer Protocol), jonka data kuljetetaan verkossa TCP/IP-protokollan avulla. Ensimmäinen HTTP:n versio, ”HTTP/0.9”, oli tarkoitettu puhtaasti datan siirtämiseen ja seuraavat versiot lisäsivät protokollaan ominaisuuksia. HTTP:n 1.0-versiossa lisättiin HTTP-kutsuihin tietueita, joilla palvelin ja asiakasohjelma pystyivät kuvailemaan lähetettyä ja vastaanotettua dataa. Palvelin pystyi esimerkiksi kertomaan WWW-selaimelle paluudatan olevan JPEG-kuva. (Fielding et al., 1999.)

```
 1 <?xml version="1.0" encoding="UTF-8"?>
 2 <!DOCTYPE html
 3   PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
 4   "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
 5 <html xmlns="http://www.w3.org/1999/xhtml">
 6 <head>
 7   <meta http-equiv="content-type"
 8         content="text/html; charset=UTF-8" />
 9   <meta name="description" content="Volkswagen myynti-ilmoitus" />
10   <meta name="keywords" content="Volkswagen,myynti-ilmoitus" />
11   <title>Volkswagen Kupla myynnissä</title>
12 </head>
13 <body>
14   <div id="content">
15     <h1>Volkswagen Kupla myynnissä</h1>
16     <p>Myynnissä vuoden 1972 1300S, joka ei suostu käynnistymään.</p>
17     <table>
18       <tr>
19         <td>Väri</td>
20         <td>Valkoinen</td>
21       </tr>
22       <tr>
23         <td>Mittarilukema</td>
24         <td>168 000 km</td>
25       </tr>
26       <tr>
27         <td>Hintapyyntö</td>
28         <td>550€</td>
29       </tr>
30     </table>
31   </div>
32 </body>
33 </html>
```

Listaus 1: Esimerkki HTML-dokumentista.

WWW-sivuja selatessa WWW-selain hakee palvelimelta HTTP-protokollan avulla sivua kuvaavan HTML-dokumentin (HyperText Markup Language) ja piirtää sen sisällön selainikkunaan. HTML-merkintäkieli rakentuu HTML-elementeistä (element), joita ei näytetä WWW-selaimessa, vaan niillä kuvaillaan HTML-dokumentin ominaisuuksia tai muotoillaan sen sisältöä. WWW-palvelimelta haettu HTML-dokumentti on tekstimuotoista dataa ja sisältää selaimessa näytetyn tekstin lisäksi sekä kuvauksen WWW-sivun rakenteesta että viittaukset sivun muihin elementteihin, esimerkiksi kuviin. (Domingue et al., 2011.)



Kuva 1: Listauksen 1 HTML-esimerkki WWW-selaimen tulkitsemana.

World Wide Web Consortium (W3C) on Tim Berners-Leen vuonna 1994 perustama yhteenliittymä, joka vastaa muun muassa WWW-standardien kehittämisestä. HTML-merkintäkielen ensimmäinen versio on vuodelta 1991 ja sittemmin sitä on kehitetty huomattavasti eteenpäin. Listauksen 1 HTML-dokumentin esimerkissä on kuvattu auton

myynti-ilmoitus XHTML-merkintäkielellä (Extensible Hypertext Markup Language), jonka määritykselle W3C asetti korkeimman recommendation-standardointitason vuonna 2000. Auton myynti-ilmoitus, johon myös jatkossa tullaan viittaamaan, koostuu otsikosta, kuvauksesta, lyhyestä ominaisuusluettelosta ja hintapyynnöstä. Kuvassa 1 on kuvakaappaus WWW-selaimen tulkitsemasta listauksen 1 HTML-dokumentista.

HTML-merkintäkielen tärkein ominaisuus on linkittäminen, jonka avulla voidaan viitata resurssiin WWW:ssä. Resurssilla tarkoitetaan WWW:ssä julkaistua dataa, esimerkiksi JPEG-kuvaa ja toista HTML-dokumenttia. Linkittämistä varten tarvitaan globaali standardi, jolla viitattava resurssi voidaan identifioida – tätä varten on määritelty URI-standardi (Uniform Resource Identifier), joka mahdollistaa resurssin identifioinnin kontekstista riippumatta. (Berners-Lee et al., 2006.) URI:n yksinkertaistettu rakenne on seuraava:

```
skeema ":" resurssin hierarkia.
```

URI-skeemoja ovat esimerkiksi http, https ja ftp. Resurssin hierarkia määrittyy käytetyn skeeman mukaan, esimerkiksi http-skeeman yhteydessä se koostuu palvelimen verkkotunnuksesta ja resurssin polusta palvelimella. Alun perin URI jaettiin kahteen luokkaan, URL (Uniform Resource Locator) ja URN (Uniform Resource Name), joilla oli tarkoitus erottaa resurssin sijainti sen nimestä. Erottelu johti juurensa ajattelutavasta, että objektilla on erikseen osoite ja nimi; samaan tapaan kuin ohjelmassa käytetyllä datalla on muuttujan nimi ja viittaus muistiosoitteeseen. Erottelu URL:n ja URN:n välillä nähtiin myöhemmin tarpeettomana ja URI-käsitettä voidaan nykyään käyttää tarkoittamaan molempia. URI siis identifioi resurssin ja samalla ilmaisee sen hakutavan (access mechanism), mutta URI:n käyttäminen ei takaa, että resurssi olisi suoraan haettavissa viittausympäristöstä. HTML-dokumentissa voidaan esimerkiksi viitata file-skeemalla tiedostoon, joka sijaitsee ainoastaan jollain tietyllä tietokoneella (”file:///C:/Windows/Help/fi-FI/credits.rtf”). WWW:n yhteydessä URI:sta voidaan käyttää URL-käsitettä, joka on vakiintunut tarkoittamaan http- ja https-skeeman URI-osoitteita. (Berners-Lee et al., 2006.)

World Wide Webin teknologioilla voidaan ratkaista osa semanttiselle webille asetetuista vaatimuksista: WWW on hajautettu järjestelmä, jonka osat voivat siirtyä itsenäisesti käyttämään uuttaa teknologiaa vaikuttamatta muiden osien toimintaan. WWW:ssä ei myöskään rajoiteta datan julkaisemista keskitetyllä hallinnalla, joten WWW täyttää semanttisen webin ensimmäisen vaatimuksen. Myös HTTP-protokolla, jota WWW:ssä käytetään, on toimiva ratkaisu datan siirtoprotokollaksi, joka ei rajoita siirretyn datan muotoa

– viides vaatimus siis täyttyy. WWW:ssä resurssien identifiointiin ja linkittämiseen käytetyt URI-osoitteet ovat osa ratkaisua semanttisen webin kolmanteen vaatimukseen. HTML-merkintäkieli soveltuu lähtökohdaksi semanttisen webin toiselle vaatimukselle yksinkertaisesta esitysmuodosta.

2.2 Semantiikan merkitseminen World Wide Webissä

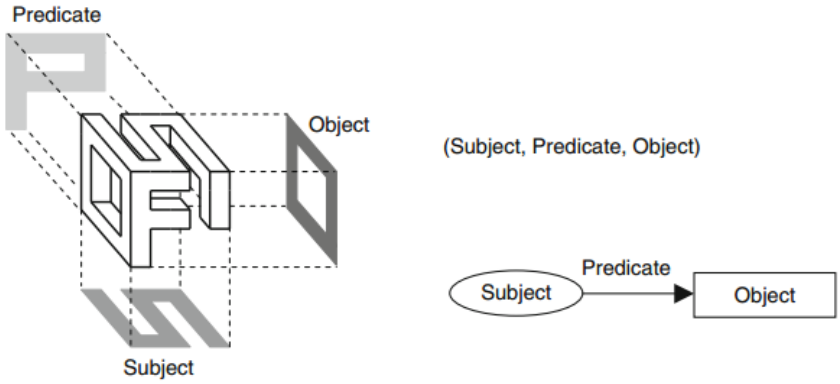
World Wide Webin alkuaikoina, ennen kuin suurin osa ihmisistä edes oli tietoisia sen olemassaolosta, useat tutkijaryhmät leikkivät idealla, että HTML-dokumentit sisältäisivät ”vinkkejä”, joiden avulla niiden sisältöä voitaisiin tulkita koneellisesti (Domingue et al., 2011). Itse asiassa osa HTML:n muotoiluun käytetyistä HTML-elementeistä on semanttisia, merkitystä määritteleviä, joten sisältöä voidaan niiden avulla tulkita jossain määrin koneellisesti (Berjon et al., 2014): Esimerkiksi *form*-elementillä luodaan lomake HTML-dokumenttiin ja samalla se määrittelee elementin sisällön. *Div*-elementti on puolestaan esimerkki yleisluontoisesta HTML-elementistä, jonka sisällöstä ei voida juurikaan tehdä päätelmiä.

Vuonna 2014 W3C julkaisi HTML5-version recommendation-standardointitasolla, tässä HTML:n uusimmassa versiossa on enemmän sisältöä kuvaavia HTML-elementtejä, esimerkiksi *main*- ja *nav*-elementit. *Main*-elementtejä tulisi olla HTML-dokumentissa ainoastaan yksi ja se on tarkoitettu dokumentin pääsisällön merkintään – sisällön, joka ei toistu WWW-sivuston muilla sivuilla. *Nav*-elementtiin tulisi sijoittaa HTML-sivuston navigaatio ja se voi toistua jokaisella WWW-sivuston sivulla. (Berjon et al., 2014.)

HTML-merkintäkielen semanttinen ilmaisuvoima rajoittuu kuitenkin edellä esiteltyyn yksinkertaiseen sisällön kuvailuun, jonka perusteella voidaan esimerkiksi todeta tekstin olevan otsikko. Jotta HTML-dokumentin sisältö voidaan kuvata tarpeeksi tarkasti, tarvitaan kuvailua varten formaali kieli. Semanttisen webin alkuaikoina tutkijat saivat vakuutettua Yhdysvaltojen asevoimien tutkimusorganisaation (DARPA, Defense Advanced Research Projects Agency) rahoittamaan semanttisen webin teknologian tutkimusta. Hanketta perusteltiin sen mahdollisuuksilla helpottaa puolustusvoimien järjestelmien välisen datan integrointia. Hankkeeseen liittyvää tutkimusta tehtiin DARPA Agent Markup Language -projektin (DAML) alaisuudessa, jonka yhteydessä aloitettiin Tim Berners-Leen johtama Semantic Web Advanced Development -projekti (SWAD). (Domingue et al., 2011.)

SWAD-kehittämisprojektin pohjana käytettiin RDF-kehyksen tutkimusta (Domingue et al., 2011). RDF-lyhenne muodostuu sanoista Resource Description Framework. Resurssi

(resource) on semanttisen webin ydinkäsitteitä ja sillä voidaan viitata esimerkiksi WWW-sivuun, videoon, laitteeseen, organisaatioon, tuotteeseen, palveluun ja henkilöön – käytännössä mihin tahansa, mikä voidaan identifioida URI:lla. Resurssien kuvaaminen (description) mahdollistaa resurssien käsittelyn – yksinkertaistettuna resurssin kuvaus on joukko attribuutteja, ominaisuuksia ja resurssiin liittyviä yhteyksiä (relations). Kehys (framework) määrittelee mallin, kielen ja syntaksin resurssien kuvauksille. (Gandon et al., 2011.)



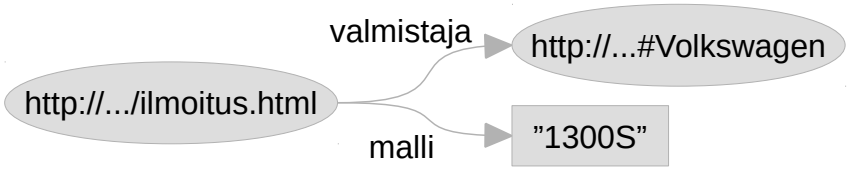
Kuva 2: RDF-triplan rakenne ja sen osatekijöistä muodostuva graafi (Gandon et al., 2011).

Lyhyemmin ilmaistuna RDF on yksinkertainen kehys, jolla voidaan semanttisesti kuvailla resursseja WWW:ssä. RDF-kehyksessä tietueet merkitään RDF-triploina, jotka koostuvat subjektista, predikaatista ja objektista (kuva 2). RDF-tripla kuvaa objektin lauseena (statement), joka muodostuu resurssista, ominaisuudesta (property) ja ominaisuuden arvosta (property value); resurssi on lauseen subjekti, ominaisuus on lauseen predikaatti ja ominaisuuden arvo on lauseen objekti. Subjektin, predikaatin ja objektin merkintään käytetään URI-osoitteita. Objektin kohdalla voidaan käyttää myös literaaliarvoa, esimerkiksi merkkijonoa, tai sen arvo voi olla tyhjä (null). (Gandon et al., 2011.)

Lauseesta, ”Autoharrastajalla on myytävänä käytetty Volkswagen 1300S”, voidaan muodostaa seuraava subjekti-predikaatti-objekti -tripla: resurssi eli subjekti on (myytävänä oleva) ”auto”, predikaatti on auton ”valmistaja” ja objekti on valmistajan nimi ”Volkswagen”. Muodostetulla RDF-triplalla kuvaillaan auton valmistajan olevan Volkswagen. Esimerkkilauseesta on tunnistettavissa myös muita RDF-triploja, esimerkiksi auton myyjä (autoharrastaja) ja malli (1300S).

RDF-triplat voidaan hahmottaa pienenä graafina (oikealla kuvassa 2), joka koostuu subjektin ja objektin muodostamista kahdesta solmusta (vertice) ja niitä yhdistävästä kaarena (arc)

kuvatusta predikaatista. Kun RDF-triplan osille käytetään URI-osoitteita, esimerkiksi sen subjekti identifioidaan URI-osoitteella, ja URI-osoitteella merkitään yhtä solmua graafissa, muodostuu RDF-triploista suurempi graafi. (Gandon et al., 2011.) Kuvassa 3 on esitetty graafina kaksi myynti-ilmoituksesta tunnistettua RDF-triplaa. Graafissa auto ja sen valmistaja on merkitty URI-osoitteilla ja auton malli merkkijonona.



Kuva 3: Myynti-ilmoituksesta tunnistetut RDF-triplat graafina.

RDF-kehys on tarkoitettu yleiseksi semanttisen webin työkaluksi, jolla voidaan kuvailla mitä tahansa resurssia. Semanttista webiä varten RDF-graafille kuitenkin tarvitaan esitysmuoto, joka voidaan käsitellä koneellisesti. Vuonna 2004 W3C julkaisi standardin, jolla RDF-graafi voidaan sarjallistaa (serialize) XML-dokumentiksi (Extensible Markup Language). (Gandon et al., 2011.) XML-standardi määrittelee muodon (format), jolla data sarjallistetaan, mutta ei määrittele sen rakennetta (structure). Muun muassa HTML-merkintäkieli pohjautuu XML:ään; HTML-standardin määrittelemät HTML-elementit ovat XML-standardin mukaisia ja niiden lisäksi HTML-standardi määrittelee HTML-elementtien järjestyksen (rakenteen) HTML-dokumentissa.

RDF-datan käyttäminen HTML-dokumenttien yhteydessä tarkoittaa käytännössä sitä, että WWW-sivustojen pitää tarjota XML-muotoon sarjallistettu RDF-data erillään HTML-dokumenteista. Jotta sisällön tuottaminen semanttiseen webiin helpottuisi, RDF/XML-muodolle on kehitetty vaihtoehtoisia tapoja RDF-datan sarjallistamiseen. Sisällön kuvailemiseen ja semanttisen datan merkintään on kehitetty RDF-kehyksen lisäksi myös muita tapoja. Yksi näistä vaihtoehtoisista tavoista on Microformats-merkinnät, josta käytetään toisinaan µF-lyhennettä. (Adida et al., 2011.)

Microformats-merkinnät on kehitetty erillään semanttisen webin tutkimuksesta mahdollistamaan semanttisen datan lisääminen suoraan HTML-dokumenttiin. Microformats-merkinnöissä käytetään HTML-elementtien attribuutteja, esimerkiksi *class*- ja *title*-attribuutteja. HTML-attribuuttien käyttö helpottaa metadatan lisäämistä ja mahdollistaa olemassa olevien HTML-tekstityökalujen käytön tähän tarkoitukseen. Microformats kehitettiin alun perin blogeja varten, joiden muokkaustyökaluissa HTML:n ominaisuuksia on rajattu esimerkiksi estämällä joidenkin HTML-elementtien käyttö kokonaan (Pohorec et al.,

2013). Microformats-merkinnöillä kuvaillun sisällön hyödyntäminen HTML-dokumenteista on myös helppoa, koska siihen voidaan käyttää HTML-merkintäkielen jäsentämiseen tarkoitettuja ohjelmakirjastoja. (Adida et al., 2011.)

Esimerkiksi yhteystietojen tai tapahtuman ajankohdan kuvaamiseen Microformats-merkinnät ovat helppo ratkaisu. Microformats-merkinnöissä HTML-dokumentin sisällön kuvailuun käytetään avainsanoja, jotka kirjoitetaan HTML-attribuuttien arvoihin. Avainsanojen käyttö HTML-attribuuteissa voi kuitenkin osoittautua ongelmallisiksi siinä tapauksessa, että HTML-dokumentissa on kahta eri tyyppistä semanttista kuvailua tarvitsevaa sisältöä. Mikäli sisältöjen kuvailuun käytettävissä termeissä on päällekkäisyyksiä, tämä voi aiheuttaa ristiriitoja semanttisissa merkinnöissä. HTML-attribuuttien käyttö voi myös olla ristiriidassa toisten standardien kanssa, esimerkiksi lukulaitteita auttaviin HTML-merkintöihin käytetään usein *title*-attribuuttia *abbr*-elementtin kanssa. (Adida et al., 2011.)

W3C on julkaissut vuonna 2008 RDF-datan sarjallistamiseen RDFa-standardin (Resource Description Framework in Attributes), jossa käytetään Microformats-merkintöjen tapaan HTML-attribuutteja. RDFa-merkinnöille tosin käytetään vain osittain HTML-standardissa määriteltyjä HTML-attribuutteja, joiden lisäksi on nimetty uusia attribuutteja semanttisia merkintöjä varten, joten päällekkäisyydet muiden standardien kanssa ovat epätodennäköisempiä. Myös ristiriidat RDFa-merkinnöillä tehdyssä kuvailussa ovat epätodennäköisiä, koska RDFa-merkinnät perustuvat RDF-kehykseen, jossa kuvaavien termien (subjekti ja predikaatti) arvoina käytetään URI-osoitteita. (Adida et al., 2011.)

Microdata on kolmas merkintätapa, jolla voidaan lisätä semanttista kuvailua suoraan HTML-dokumenttiin. Myös Microdata-formaatilla voidaan sarjallistaa RDF-dataa ja tähän käytetään HTML-attribuutteja. Tosin Microformats- ja RDFa-formaateista eroten Microdata-formaatilla tehtyihin semanttisiin merkintöihin käytetään ainoastaan niitä varten nimettyjä uusia HTML-attribuutteja. WHATWG-ryhmä (Web Hypertext Application Technology Working Group) on määritellyt Microdata-formaatin osana ryhmän julkaisemaa HTML(5)-standardia, jonka pohjalta W3C on luonut oman ehdotuksen Microdata-formaatin HTML5-merkintäkieleen lisäämiseksi. (Hickson, 2013.)

Tällä hetkellä vielä keskeneräisessä W3C:n HTML 5.1 -määrittelyssä Microdata-formaatin HTML-attribuutit on lisätty kaikille HTML-elementeille sallittuihin ”globaaleihin” HTML-attribuutteihin. HTML 5.1 -määrittely myös linkittää kyseisten HTML-attribuuttien nimet W3C:n Microdata-dokumentaatioon, joten on hyvin todennäköistä, että Microdata-formaatista

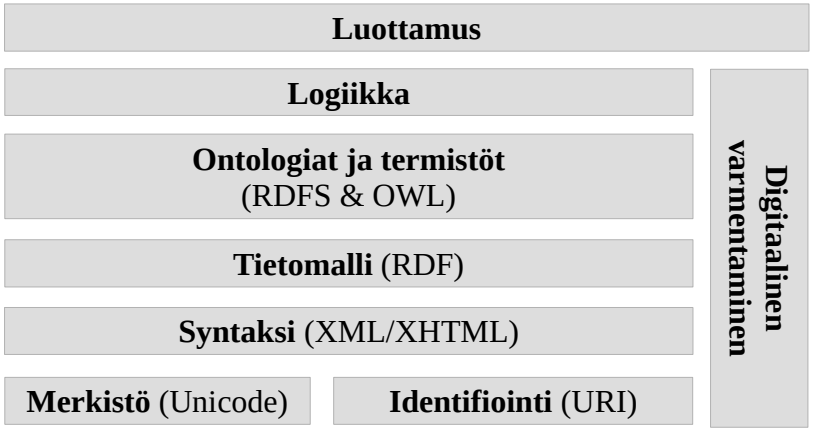
tulee W3C:n suositus HTML-dokumenttien semanttisille merkinnöille. Oletusta Microdata-formaatin suosittelusta tukee HTML 5.1-määrittelyn *data*-attribuuttien yhteydessä oleva maininta siitä, että Microdata-formaattia tulisi käyttää siinä tapauksessa, että HTML-dokumentissa halutaan julkaista dataa kolmannen osapuolen käytettäväksi. Tosin HTML+RDFa-formaatti on yksi HTML5:n laajennoksista, joten RDFa-formaattia tuskin ollaan varsinaisesti hylkäämässä. (Berjon et al., 2015.)

Yhdistämällä esitellyt semanttisen datan merkintään soveltuvat formaatit HTML-merkintäkielen kanssa saadaan ratkaistua semanttisen webin rakentamiselle asetettu vaatimus yksinkertaisesta esitysmuodosta, joka samalla piilottaa semanttiseen teknologiaan liittyvää monimutkaisuutta.

2.3 Ontologiat semanttisessa webissä

HTML-merkintäkielessä on ollut jo ensimmäisistä versioista lähtien mukana mahdollisuus liittää asiasanoja ja kuvaus HTML-dokumenttiin. HTML-dokumentissa asiasanat (keywords) ja kuvaus (description) sijoitetaan HTML-merkintäkielessä määriteltyihin *meta*-elementteihin; listauksen 1 myynti-ilmoituksessa on esimerkki asiasanojen ja kuvauksen käytöstä riveillä 9 ja 10. Asiasanojen käyttäminen HTML-dokumentin sisällön kuvailemiseen juontaa juurensa todennäköisesti WWW:n alkuperäisestä käyttötarkoituksesta – tieteellisten dokumenttien jakamisesta. Asiasanojen käyttö HTML-dokumenttien metadatana on kuitenkin ongelmallista, koska niitä varten ei ole yksittäistä kaiken WWW-sisällön kattavaa asiasanastoa, jolloin kukin WWW:ssä julkaiseva taho voi vapaasti valita asiasanat sisällölleen ja tämä johtaa siihen, ettei saman aihepiirin sivustoilla ole välttämättä samoja asiasanoja.

Sosiaalisessa mediassa on käytetty onnistuneesti tagittamista (tagging) sisällön kuvailemiseen. Tageilla tarkoitetaan käyttäjien itse sisällölle antamia kuvailutermejä, jotka ovat vapaasti määriteltävissä. Kun samoja kuvailutermejä käytetään useasti, tageista muodostuu eräänlainen luokittelujärjestelmä – käyttäjät sopivat epäsuorasti yhteisistä kuvailutermeistä. Kun kuvailutermien ja sisällön määrä kasvaa tarpeeksi suureksi, rakentuu yhteinen luokittelujärjestelmä tarkemmaksi ja siitä on mahdollista havaita selkeitä rakenteita, esimerkiksi tietyt kuvailutermit voivat esiintyä yhdessä toisia useammin. Semanttisen webin on tarkoitus ratkaista WWW:n sisällön kuvaileminen tagittamiseen verrattavalla tavalla; sisällön kuvailuun ei välttämättä tarvita yhtä kaiken kattavaa termistöä vaan riittää, että kuvaileminen toimii tietyssä kontekstissa. (Berners-Lee et al., 2006.)



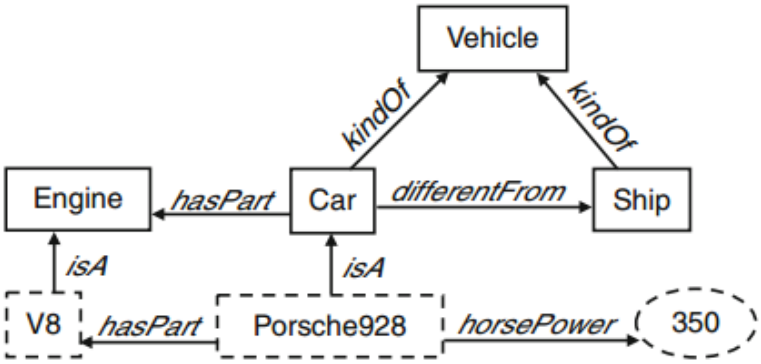
Kuva 4: Semanttisen webin teknologiapino.

Ontologioiden avulla määritellyt sanastot ovat semanttisen webin ratkaisu sisällön kuvailemiseen yhteisillä termeillä ja niiden käyttö on keskeisin osa semanttisen webin rakentumista; ontologiat sijoitetaan semanttisen webin teknologiapinossa tietomallien päälle (kuva 4). Ontologia-termi (ontology) on lähtöisin filosofian tutkimuksesta, jossa sillä tarkoitetaan olemisen ja olemassaolon käsitteiden tutkimista, kuten esimerkiksi minkä tyyppisiä asioita on olemassa ja minkälaisia suhteita olemassa olevien asioiden välillä on. Tietojenkäsittelyssä ontologisesta luokittelusta on hyötyä kohdeympäristöön liittyvien käsitteiden määrittelyssä, jolloin semanttinen sanasto voidaan rajata kuvaamaan tietyt kohteet reaalimaailmasta, esimerkiksi ”auto” ja ”moottori”. (Grimm et al., 2011.)

Tietojenkäsittelyssä ontologialle (an ontology) on useita määritelmiä, mutta semanttisen webin tutkimuksessa ontologialle on vakiintunut seuraava määritelmä:

> *An ontology is a formal explicit specification of a shared conceptualization of a domain of interest.* (Grimm et al., 2011.)

Määritelmä sisältää useita ontologioihin liittyviä keskeisiä piirteitä: Ontologia ilmaistaan formaalilla kielellä, jolloin se on koneellisesti tulkittavissa. Ontologia on eksplisiittinen kuvaus kohdeympäristöstä – mallintamattomia käsitteitä ei ole olemassa. Ontologia kuvastaa tiedeyhteisön jaettua ymmärrystä kohdeympäristön käsitteellistämisestä (conceptualization). Ontologian käsitteillä kuvaillaan rajattu kohdeympäristö (domain), ja mitä tarkemmin rajaus tehdään, sitä tarkemmin sen käsitteet voidaan määritellä. (Grimm et al., 2011.)



Kuva 5: Esimerkki ontologiasta graafina (Grimm et al., 2011).

Ontologioiden havainnollistamiseen käytetään usein graafia, joka koostuu toisiinsa yhdistetyistä käsitesolmuista (kuva 5). Ontologian esittäminen graafina kuvaa hyvin ontologian käyttöön liittyviä piirteitä: ilman solmujen välisiä yhteyksiä (interrelation), ontologia olisi kokoelma käsitteitä ja niiden välisten suhteiden antamat merkitykset puuttuisivat. Esimerkiksi auto ja moottori voisivat jäädä kahdeksi irralliseksi käsitteeksi, mutta niiden välille voidaan määritellä yhteys: autoon kuuluu moottori. Ontologiassa on tärkeää osata erottaa luokat eli abstraktit käsitteet ja oliot eli luokkien ilmentymät (instance). Kuvan 5 graafissa ”V8” ei ole oma luokkansa, vaan ”Engine”-luokan ilmentymä; tosin tämä riippuu ontologian termien tarkkuudesta, jossain ontologiassa ”V8” voisi olla ”Engine”-luokan alikäsite. Käsitteitä voidaan yhdistää toisiinsa perimällä (subsumption), jolloin perivällä luokalla on myös perityn luokan ominaisuudet (auto on [kindOf] kulkuneuvo). Ilmaisuvoimaltaan vahvemmissa ontologiakielissä voidaan sulkea pois luokkien välisiä suhteita ja estää ristiriitaiset käsitteet, esimerkiksi, että auto ei voi olla samaan aikaan laiva. Ontologiakielissä on mahdollista määritellä myös monimutkaisempia perusoletuksia (axiom) niiden perusoletuksien lisäksi, jotka määrittyvät esimerkiksi perinnän kautta. Ontologiakielissä on usein myös mahdollista määritellä ilmentymille attribuutteja, joilla voidaan ilmaista yksinkertaisia toteamuksia ilmentymistä (”Porsche928”-autossa on 350 hevosvoimaa). (Grimm et al., 2011.)

Ontologioita voidaan verrata toisiin käsitemalleihin, esimerkiksi ER-malliin (Entity Relationship). ER-mallilla kuvaillaan kohdeympäristön käsitteet suunnitteluvaiheessa ja sen pohjalta luotua tietokantaa käytetään suoritusvaiheessa datan tallentamiseen, jolloin tietokannan rakennetta ei enää muuteta (Mannila & Räihä, 1992). Ontologialla kohdeympäristön käsitteet kuvaillaan siten, että dataa voidaan käsitellä, vaikka suoritusvaiheessa annettu data ei välttämättä sisällä kaikkea mahdollista informaatiota. Ontologioiden avulla datasta on mahdollista johtaa implisiittistä tietoa aksioomien avulla,

esimerkiksi ”Volkswagen Kupla on auto” -lauseesta voidaan ontologian avulla johtaa tieto, että Volkswagen Kupla on kulkuneuvo. ER-mallilla luodusta tietokannasta voitaisiin todennäköisesti johtaa sama tieto, mutta ER-mallilla olisi mahdotonta todeta, ettei Volkswagen Kupla ole laiva. (Grimm et al., 2011.) Tosin EER-mallilla (Enhanced Entity Relationship), ER-mallin laajennoksella, tämä olisi mahdollista (Elmasri & Navathe, 2014).

Kuten ontologioiden määritelmässä todettiin, niiden ilmaisemiseen tarvitaan formaaleja kieliä. Ontologiakielet ovat ontologioiden formaaleja määrittelytapoja, jotka perustuvat logiikkaan. Aiemmin esitelty RDF-malli on ontologiakieli ja sen tärkein ominaisuus on ”rdf:type”-predikaatti[2], jolla voidaan määritellä resurssin luokka sen kuvailuun käytetystä ontologiasta – RDF-mallia ei ole siis sidottu mihinkään tiettyyn ontologiaan. RDF-malli on yleisluonteinen standardi metadatan kuvaamiseen ja se ei sisällä juurikaan rajoituksia, RDF:llä voidaan esimerkiksi määritellä täysin hyödytön ”rdf:type, rdf:type, rdf:Property”-tripla. Pelkän RDF-mallin käyttäminen ontologian määrittelyyn on työlästä, koska sillä ei ole mahdollista määritellä luokkien välisiä yhteyksiä (interrelation), minkä vuoksi luokille yhteisiä määrityksiä jouduttaisiin kopioimaan luokkien välillä. (Domingue et al., 2011.)

RDFS (RDF Schema) on RDF-mallin päälle määritelty ontologiakieli, joka mahdollistaa luokkien ja niiden välisten yhteyksien määrittelemisen. RDFS-skeema kärsii kuitenkin RDF-mallin puutteista: Ensinnäkään RDFS:llä ei voi tehdä poissulkevia määrityksiä. Toiseksi RDFS ei salli lukumäärien asettamista ominaisuuksille, esimerkiksi että autolla on oltava neljä pyörää. Kolmas rajoitus on, että RDFS:ssä ei voi rajoittaa ominaisuuksien arvoja, esimerkiksi ettei tuotteen myyjä ja ostaja ole samat. (Pohorec et al., 2013.)

OWL (Ontology Web Language) paikkaa RDFS:n puutteita, tosin niiden korjaamiseksi OWL on käytännössä toteutus kuvauslogiikasta ja sen vuoksi ehkä turhankin ilmaisuvoimainen. OWL on jaettu kolmeen alikieleen: OWL Lite, OWL DL ja OWL Full. OWL Litellä voidaan määritellä luokkarakenteita ja yksinkertaisia rajoituksia. OWL DL lisää kielen ilmaisuvoimaa ja takaa laskennallisen täydellisyyden (computational completeness). OWL Full yhdistää OWL-kielen RDF- ja RDFS-määrityksiin, ja on siten yhteensopiva RDF:n kanssa. OWL Full on kuitenkin niin ilmaisuvoimainen, ettei sillä voida taata laskennallista täydellisyyttä. (Pohorec et al., 2013.)

Ontologiakielillä määritellyt ontologiat muodostavat termistön, jota voidaan käyttää

---

2 CURIE-syntaksin (Compact URI) mukainen lyhennös ”http://www.w3.org/1999/02/22-rdf-syntax-ns#type”-osoitteesta.

kohdeympäristön kuvaamiseen. FOAF-ontologia (Friend of a friend) on esimerkki semanttisen webin kevyestä (lightweight) ontologiasta, jolla voidaan kuvailla ihmisiä, ihmisten aktiviteetteja ja ihmisten välisiä suhteita. FOAF-ontologian termien määrittely pohjautuu RDF-, RDFS- ja OWL-kielien termeille. Kuka tahansa voi käyttää FOAF-ontologiaa itsensä kuvailuun WWW:ssä ja lisäämällä esimerkiksi kaverisuhteet kuvailuun, voidaan muodostaa sosiaalisia verkostoja (social network). (Grimm et al., 2011.)

```
 1 <!DOCTYPE html>
 2 <html>
 3 <head>
 4   <meta charset="UTF-8">
 5   <title>Eero Esimerkki</title>
 6 </head>
 7 <body>
 8   <main itemscope itemtype="http://xmlns.com/foaf/0.1/Person">
 9     <h1>
10       <span itemprop="firstName">Eero</span>
11       <span itemprop="lastName">Esimerkki</span>
12     </h1>
13     <p>Kollegani W3C:ssä:</p>
14     <ul>
15       <li>
16         <a itemprop="knows"
17            href="http://www.w3.org/People/Berners-Lee/"
18            >Tim Berners-Lee</a>
19       </li>
20       <li>
21         <a itemprop="knows"
22           href="http://www.w3.org/People/Ivan/"
23           >Ivan Herman</a>
24       </li>
25     </ul>
26   </main>
27 </body>
28 </html>
```

Listaus 2: Esimerkki FOAF-ontologian käytöstä.

FOAF-ontologian käyttöä on havainnollistettu kuvitteellisen Eero Esimerkin kotisivulla (listaus 2), johon on lisätty sen sisältöä kuvailevaa semanttista dataa. FOAF-esimerkki on HTML5-määrityksen mukainen ja siinä on käytetty Microdata-formaattia semanttisen datan HTML-dokumenttiin upottamiseksi. FOAF-ontologiassa on määritelty Person-luokka (*itemtype*-attribuutti rivillä 8), joka sisältää henkilön kuvailemiseen käytettäviä ominaisuuksia, esimerkiksi etu- ja sukunimen (firstName, lastName), joita on käytetty esimerkin riveillä 10 ja 11. Ontologiassa on myös *name*-ominaisuus, mutta se on määritelty Thing-luokalle ja Person-luokka ei määrityksen mukaan peri Thing-luokkaa (Brickley & Miller, 2014), joten henkilön nimeä ei voida ilmaista seuraavalla HTML-tekstillä:

```
<h1 itemprop="name">Eero Esimerkki</h1>.
```

Henkilöiden välisten suhteiden kuvailuun FOAF-ontologiassa on määritelty Person-luokalle *knows*-ominaisuus, jonka arvoksi annetaan toinen Person-luokan ilmentymä, esimerkiksi URI-osoitteena. Microdata-standardin mukaan *a*-elementtiä käytettäessä ominaisuuden

(*itemprop*) arvoksi (property value) määräytyy *href*-attribuutin URI-osoite, joten FOAF-esimerkin riveillä 16 ja 21 *knows*-ominaisuuden arvoksi määräytyy URI-osoite (Hickson, 2013). Viitattava henkilö identifioidaan URI-osoitteen avulla ja sitä seuraamalla voidaan hakea viitatun henkilön profiilitiedot.

URI-osoitteiden käyttö RDF-mallissa mahdollistaa viitattavien kohteiden identifioinnin ja samalla URI-osoitteiden seuraaminen mahdollistaa tiedon hakemisen viitattavista kohteista. Ongelmaksi kuitenkin muodostuu URI-osoitteen valinta tietoa julkaistaessa – miten viitattavan kohteen identifioiva URI-osoite löydetään ja voidaanko varmistua, että löydetty URI-osoite on oikea? FOAF-esimerkissä viitatulle Ivan Hermanille löytyi esimerkkiä kirjoittaessa kaksi URI-osoitetta: ”http://www.w3.org/People/Ivan/” ja ”http://www.ivan-herman.net/foaf#me”.

Löytyneistä URI-osoitteista esimerkkiin valittiin W3C-sivuston osoite. Käytännössä WWW-sisällön tuottajat joutuvat tekemään samanlaisia valintoja URI-osoitteiden kohdalla. Joten on hyvin todennäköistä, että saman kohteen identifiointiin käytetään useita URI-osoitteita. OWL-kielessä on osittainen ratkaisu ongelmaan, sillä ilmentymien samuus voidaan määritellä ”owl:sameAs”-ominaisuudella. Ongelman ratkaisemiseksi tarvitaan vielä järjestelmä, josta voidaan hakea samaa tarkoittavia URI-osoitteita – sameAs.org[3] on luotu tätä varten. Ivan Hermanin henkilöprofiiliin viittaavalle URI-osoitteelle sameAs-palvelu löysi 99 vastaavuutta.

Useiden URI-osoitteiden identifioidessa samaa kohdetta herää kysymys niiden takana olevasta informaatiosta. Esimerkiksi Ivan Hermanin tapauksessa kaikista URI-osoitteista tuskin löytyy toistettuna kaikkia henkilötietoja tai henkilöiden välisiä suhteita. Käytännössä tämä johtaa siihen, että semanttisen webin dataa indeksoiva – tai muuten käsittelevä – järjestelmä joutuu jäsentämään informaation jokaisesta löydetystä URI-osoitteesta. Toisaalta informaation pirstaloituminen ja toistuminen mahdollistaa tarkemman kuvan muodostamisen kohteesta, kun eri tahot voivat kuvailla samaa kohdetta jopa eri ontologioilla.

Ontologioilla saadaan täydennettyä aiemmin esitettyjä semanttiseen webiin liittyviä teknologioita. Ontologiat ovat eksplisiittinen esitysmuoto monimutkaiselle datalle ja niiden käytöllä mahdollistetaan datan julkaiseminen, uudelleenkäyttö ja linkittäminen. Yhdistämällä ontologiat ja muut semanttisen webin teknologiat, saadaan muodostettua perusta semanttiselle webille.

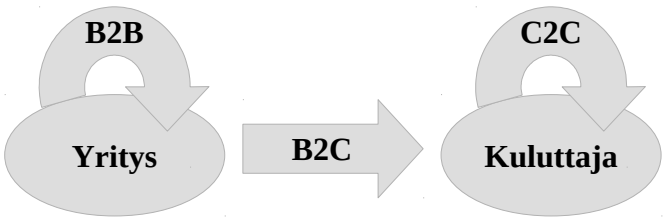
---

3 http://sameas.org/

3 Sähköinen kaupankäynti

Sähköinen kaupankäynti (electronic commerce, E-Commerce) on yleisnimitys WWW:ssä tapahtuvalle kaupalle (Havumäki & Jaranka, 2014). Käsite kattaa sekä palveluiden että tuotteiden esittelemisen, markkinoinnin, myymisen, ostamisen, maksamisen ja jakelun eri muodoissaan (Tinnilä et al., 2008). Kuten Havumäki ja Jaranka (2014) esittävät, on WWW nykypäivänä itsestään selvä osa niin kuluttajien kuin yritystenkin arkea: kuluttajat hakevat tietoja yritysten tuotteista ja palveluista, ostavat ja tilaavat niitä, kuten myös luovat ja jakavat tietoa eteenpäin muille, ja yritykset puolestaan käyttävät WWW:tä muun muassa myyntiin, markkinointiin kuten myös yhteydenpitoon niin asiakkaiden kuin muiden yhteistyökumppaniensa kanssa.

Käytettävien tekniikoiden monipuolistumisen myötä E-Commerce -termin rinnalle on syntynyt myös muita sähköisen kaupankäynnin määritteitä. M-Commercella tarkoitetaan mobiiliverkkokauppaa, jossa verkkokaupassa asioidaan mobiililaitteen avulla. F-Commerce, eli Facebook-Commerce, on sosiaalisessa mediassa käytävää kaupankäyntiä. Shop-in-shop -nimike on vakiintunut tarkoittamaan virtuaalista kauppapaikkaa, jonka tarjonta koostuu useiden verkkokauppiaiden tuotteista. (Havumäki & Jaranka, 2014.)



Kuva 6: Sähköisen kaupankäynnin lyhenteet.

Sähköisen kaupankäynnin yhteydessä käytetään usein englannin kielestä peräisin olevia lyhenteitä, jotka on nimetty myynnin osapuolten mukaan: B2B, B2C ja C2C (kuva 6). Yritysten välisestä kaupankäynnistä käytetään B2B-lyhennettä (business-to-business), kuluttajakaupasta B2C-lyhennettä (business-to-consumer) ja kuluttajien välisestä kaupasta C2C-lyhenettä (consumer-to-consumer). (Tinnilä et al., 2008.)

Yritysten välisessä sähköisessä kaupankäynnissä (B2B) ovat kyseessä yleensä suljetut järjestelmät, joilla yritykset tekevät keskenään kauppaa, esimerkiksi tukkukauppiaat myyvät tuotteitaan vähittäismyyjille tätä varten rakennetun verkkokaupan kautta. Tähän sähköisen kaupankäynnin muotoon liittyy myös e-hankinta -termi (e-procurement), jolla tarkoitetaan tarjousten tekemistä sähköisesti ja sekä tuotteiden että palveluiden ostamista sähköisiä

kanavia pitkin. (Havumäki & Jaranka, 2014.)

Suoraan kuluttajien välillä tehtävä myynti (C2C) tapahtuu esimerkiksi erilaisissa verkkohuutokaupoissa sekä vapaamuotoisesti esimerkiksi sosiaalisen median osto ja myynti -sivuilla. Kansainvälisesti tunnetuin verkkohuutokauppa on Ebay. Koska kuluttajien välisessä kaupassa ostajan oikeudet ovat ainakin toistaiseksi melko heikot ja mahdollisuudet tuotteeseen tutustuminen elektronisessa ympäristössä puutteelliset, on C2C-kaupan onnistumisen keskeisin vaatimus riittävät luottamusmekanismit. Tyypillisin näistä mekanismeista on käyttäjien toisilleen antama palaute. (Tinnilä et al., 2008.)

Sähköinen kuluttajakauppa (B2C) on itsessään laaja käsite, koska sillä tarkoitetaan kaikkea liiketoimintaa, jossa kuluttaja ostaa yritykseltä tuotteen, palvelun tai sisältöä tietoverkon välityksellä (Tinnilä et al., 2008). Luottamus on tärkeässä osassa sähköisessä kuluttajakaupassa: kuluttajan tulee luottaa niin verkkokauppaan kauppakumppanina kuin verkkotekniikkaan ostamisvälineenä (Havumäki & Jaranka, 2014). Suomessa matkailu- ja majoituspalvelut, vaatteet, pääsyliput sekä rahapelit ovat pysyneet vuosia verkkokaupan suosituimpina tuoteryhminä (SVT, 2014). Vuonna 2013 verkkokaupan osuus Suomessa koko vähittäiskaupasta oli melko vähäinen, 8 %, mutta sen volyymi oli kuitenkin neljässä vuodessa kasvanut viidenneksen ja kasvun varaa on (Anders Innovations, 2014).

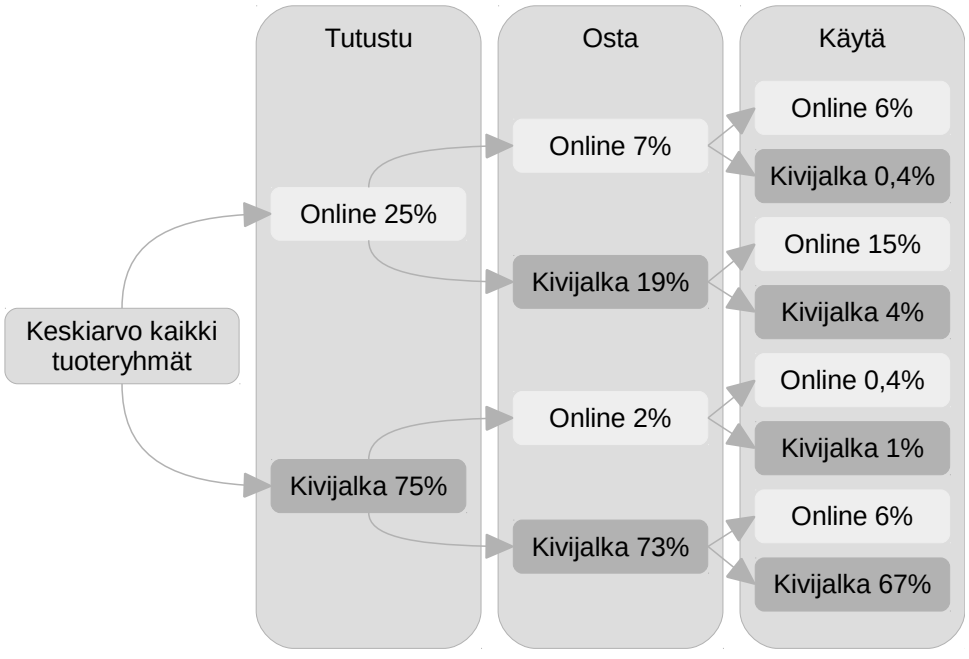
3.1 Verkkokaupat ja monikanavaisuus

”Verkkokaupalla tarkoitetaan yksittäisen yrityksen internetsivustoa, jossa myydään tuotteita tai palveluja” (Havumäki & Jaranka, 2014). Verkkokaupankäynnillä voi nähdä monenlaisia etuja sekä yritykselle että kuluttajalle. Yrityksen näkökulmasta säästöjä syntyy muun muassa esimerkiksi tila- ja henkilöstökustannuksien pienentyessä, sillä verkkokaupassa tilausten vastaanottaminen ja maksaminen ovat automatisoitavissa eikä kivijalkakaupan hyllyjä täydy olla jatkuvasti järjestämässä vaan tuote-esittelyt pysyvät verkkosivulla järjestyksessä. Lisäksi verkkokauppa on tavoitettavissa eri puolilta maailmaa kaikkina vuorokauden aikoina, mikä on etu sekä kuluttajan että yrityksen kannalta. (Havumäki & Jaranka, 2014.)

Viimeisen vuosikymmenen aikana on uutisoitu paljon kivijalkamyymälöiden haasteista verkkokaupan kasvaessa (ks. esim. Boxberg, 2013). Kivijalka- ja verkkokaupan suhde ei ole kuitenkaan niin yksinkertainen kuin monet tiedotusvälineet antavat ymmärtää. Tulevaisuudenkuva, jossa kuluiltaan ja hinnoittelultaan edullisemmat verkkokaupat ovat tyystin korvanneet kivijalkamyymälät, ei ole realistinen. Myöskään mobiililaitteiden aiheuttaman kaupallisen vallankumouksen nimeen vannovan Ebayn toimitusjohtaja John

Danahoen mukaan fyysiset myymälät eivät tule katoamaan, sen sijaan niiden luonne tulee muuttumaan saumattomaksi online-toimintojen kanssa; jo tällä hetkellä esimerkiksi tuotteiden tilaaminen verkon kautta kaupan omaan myymälään on yleistä (Veverka, 2013; Havumäki & Jaranka, 2014). Tutkijat puhuvatkin ”ubi-ajasta” liiketoiminnan tulevaisuuskuvana. Ubi viittaa englannin kielen käsitteeseen *ubiquitous technology* eli sulautettu teknologia. Ubi-aika tarkoittaa sitä, että tulevaisuudessa teknologia tullee olemaan niin itsestään selvä osa kaikkea liiketoimintaa, että elektronisesta liiketoiminnasta (tai kivijalkakaupasta) on turha enää puhua omana käsitteenä. (Tinnilä et al., 2008.)

Verkko- ja kivijalkakaupan suhdetta parhaiten selittää asiakkaan monikanavaisen ostokäyttäytymisen tarkastelu. Asiakkaan ostoprosessin voi jakaa esimerkiksi kolmeen vaiheeseen: tutustuminen, osto ja käyttö. Näistä tutustuminen tarkoittaa tuotteiden vertailua ja tuotteen ominaisuuksia koskevan tiedon etsimistä. Ostovaiheessa kuluttaja taas ostaa ja maksaa tuotteen. Käyttövaihe puolestaan sisältää erilaisia palveluita, joita tarvitaan tuotteen oston jälkeen kuten neuvonta tuotteen käytössä, huollossa tai tuotteen palauttamisessa. (Havumäki & Jaranka, 2014.)



Kuva 7: Kuluttajien monikanavainen ostokäyttäytyminen mieluisimman kanavan mukaan. (Wikström, 2012)

TNS Gallupin tekemä tutkimus (Wikström, 2012) osoittaa kivijalkakaupan olevan monille kuluttajille lähes aina mieluisin kanava käyttövaihetta lukuun ottamatta (kuva 7). Käytännössä kanavan valinta riippuu usein paljon palvelun tai tuotteen laadusta sekä kuluttajan iästä. Esimerkiksi moni sovittanee mieluusti lenkkarit liikkeessä, mutta ostaa ne mieluiten sieltä,

mistä ne halvimmalla saa (vrt. Boxberg, 2013). Lisäksi TNS Gallupin tutkimus käsittelee kuluttajien mieluisinta kanavaa, ei todellista käyttäytymistä. Entisestään yleistyneet älypuhelimet ovat lisänneet verkkokaupan arkipäiväisyyttä TNS Gallupin tutkimuksen jälkeen ja kansainväliset tutkimukset ovatkin antaneet viitteitä, että yhä useammat sekä hakevat tietoa mobiililaitteilla että tekevät ostoksen verkossa, vaikka kävisivät myös tutustumassa tuotteeseen myymälässä (ks. Havumäki & Jaranka, 2014). Yksi trendi onkin, että monet liikkeet ovat muuttuneet näyteikkunoiksi, esittelytiloiksi, joissa voi käydä kokeilemassa tuotteita ja sitten tilata ne verkosta (Anders Innovations, 2014).

Kivijalkamyymälä voi antaa pääosin verkossakin toimivalle yritykselle uskottavuutta ja luotettavuutta sekä mahdollistaa usein joustavamman ostoprosessin, kun tuotteeseen voi halutessaan käydä tutustumassa myymälässä ja kun sen voi mahdollisesti tilata noudettavaksi myymälään. Tämä selittää sitä, miksi esimerkiksi verkkokauppoina aloittaneet Verkkokauppa.com ja Konebox ovat perustaneet myös myymälöitä (Lahtinen, 2013). Verkkokauppa.comin menestystä perinteisiin kodinelektroniikkaketjuihin verrattuna selittää kuitenkin varmasti sekin, että Verkkokauppa.comilla on hyvin pieni, kolmelle suurelle paikkakunnalle sijoitettu myymäläverkosto, eikä myymälää joka kaupungissa.

Monikanavainen palveluntarjonta on yritykselle tärkeää. Ihanteellista olisi, jos samat palvelut voisi saada kanavasta riippumatta, mutta resurssien niukkuuden vuoksi yritykset joutuvat käytännössä kuitenkin priorisoimaan palveluntarjontaansa eri kanavien välillä (Havumäki & Jaranka, 2014). Pohtiessa verkkokaupan perustamista yrityksen tuleekin miettiä, mikä on verkkokaupan rooli suhteessa yrityksen liiketoiminnan kokonaisuuteen. Kahtena eri ääripäänä on verkkokauppa itsenäisenä, tuottoa tavoittelevana liiketoimintana ja toisessa ääripäässä verkkokauppa kivijalkakaupan myyntiä tukevana ja lisäävänä, hintojen ja tuotteisiin tutustumisen mahdollistavana kanavana, joka ei pyri merkittävään itsenäiseen myyntiin. Tosiasia on, että monilla aloilla tyypillisin ostoprosessi alkaa tarjonnan tutkimisella verkossa, joten verkkokaupan roolia kivijalkakaupan tukena ei ole syytä väheksyä. (Lahtinen, 2013.)

3.2 Hakukoneet osana verkkokaupan markkinointia

Markkinointiviestinnällä on tärkeä rooli yrityksen ja sen tuotteiden näkyväksi tekemisessä niin verkkokaupankäynnissä kuin perinteisessä kaupankäynnissä. Markkinointi luo mielikuvia ja sen kautta pyritään luomaan kysyntää ja aikaansaamaan ostopäätöksiä. Se myös tarjoaa tietoa yrityksestä ja sen tuotteista sekä luo yrityksen brändiä. (Havumäki & Jaranka, 2014.)

Markkinointiviestinnän mediat jaetaan tyypillisesti kolmeen ryhmään: omistettu media,

ostettu media ja ansaittu media. Omistettu media viittaa esimerkiksi yrityksen omaan sivustoon, blogiin tai viestintään sosiaalisessa mediassa. Ostettu media puolestaan luonnollisesti tarkoittaa hankittuja ilmoituksia ja markkinointia esimerkiksi sanomalehdissä, radiossa, verkossa, messuilla tai sponsoroinnin kautta. Ansaittu media taas merkitsee kuluttajien ja asiantuntijoiden arvioita ja kommentteja esimerkiksi blogeissa, hintavertailusivustoja sekä kaikkea suullisesti liikkuvaa tietoa. WWW:n myötä erityisesti ansaitun median rooli on kasvanut. WWW:tä käytetään tietolähteenä, kun kuluttaja yrittää perehtyä niin yrityksen luotettavuuteen kuin itse tuotteen ominaisuuksiin. Tämän vuoksi yrityksen tulisikin panostaa hakukonenäkyvyyteen. (Havumäki & Jaranka, 2014.) Jos kuluttaja ei tiedä verkkokaupan verkkotunnusta ei hän välttämättä tiedä kaupan olevan edes olemassa, sillä ei ole olemassa puhelinluettelon tyyppistä hakemistoa kaikista maailman verkkosivuista. Tämän ongelman hakukoneet ovat luotu ratkaisemaan, joten verkkokaupalle on tärkeää näkyä mahdollisimman hyvin potentiaalisten asiakkaiden tekemissä hauissa. (Anders Innovations, 2014.)

Lahtinen (2013) tarkastelee kuluttajan ostoprosessia hieman Havumäestä ja Jarangasta (2014) poiketen. Lahtinen tarkastelee ostoprosessia ennen kaikkea markkinoinnin kohdentamisen näkökulmasta ja jakaa ostoprosessin viiteen vaiheeseen: ongelman tunnistamiseen, tiedon etsimiseen, vaihtoehtojen vertailuun, ostopäätökseen ja hankinnan jälkeiseen arviointiin. Tiedon etsintävaiheessa kuluttaja etsii keinoa tyydyttää hankintatarve, joka on syntynyt ostoprosessin alussa tunnistetusta ongelmasta. Tiedon etsintä voi olla sisäistä tai ulkoista, eli kokemukseen perustuvaa ja muistin varaista tai uuden tiedon etsimistä ympäristöstä. Lahtinen (2013) myös mainitsee, että tiedon etsintä ei välttämättä liity tunnistettuun ongelmaan, vaan etsintä voi olla jatkuvaa, esimerkiksi markkinoiden tapahtumien seuraamista. Tiedon etsinnän jälkeen kuluttaja vertailee löytämiään vaihtoehtoja ja valitsee sopivimman vaihtoehdon asettamiensa kriteerien perusteella.

Tiedon etsimisvaiheessa ja vaihtoehtojen vertailuvaiheessa markkinoinnin tavoitteena on saada ohjattua liikennettä verkkokauppaan. Kuluttajat käyttävät hakukoneita tiedon hakemiseen, joten hakukoneoptimointi on yksi keinoista lisätä kävijöiden lukumäärää verkkokaupassa. Hakukoneoptimoinnilla verkkokauppa pyritään nostamaan mahdollisimman korkealle hakutuloksissa verkkokauppaan liittyvillä hakutermeillä haettaessa, koska mitä kauempana kärjestä verkkokauppa on, sitä epätodennäköisemmin sinne siirrytään hakutuloksista – kolmannelle tulossivulle joutuminen vastaa sitä, että verkkokauppaa ei olisi olemassa. (Lahtinen, 2013.)

Havumäki ja Jaranka (2014) käsittelevät hakukoneoptimointia lyhyesti ja keskittyvät pääasiassa ostettuun mediaan. Lahtinen puolestaan käsittelee hakukoneoptimointia laajemmin ja jakaa sen kahteen: sisäiseen ja ulkoiseen hakukoneoptimointiin. Ulkoisella hakukoneoptimoinnilla tarkoitetaan verkkokauppaan viittaamista muista WWW-sivustoista. Samassa Lahtinen (2013) varoittaa linkkifarmeista, ainoastaan muihin sivustoihin linkkaavista WWW-sivustoista, joita käyttämällä voi aiheuttaa pikemminkin haittaa hakukonenäkyvyydelle.

Sisäisen hakukoneoptimoinnin yhteydessä Lahtinen (2013) käy kattavasti läpi verkkokaupan rakennetta ja sisältöä parantavia tekniikoita. Hän aloittaa sisällön optimoinnin verkkokaupan verkkotunnuksesta. Jos verkkotunnus täsmää kuluttajan käyttämiin hakusanoihin, on verkkokauppa hyvin todennäköisesti ensimmäisten hakutulosten joukossa. Toiseksi Lahtinen käsittelee URL-osoitteiden rakennetta: jos ne sisältävät selkeitä asiasanoja, esimerkiksi tuotekategorian ja tuotenimikkeen, parantuu sivujen hakukonelistautuminen merkittävästi esimerkiksi pelkkiä numeroita sisältäviin URL-osoitteisiin verrattuna. Kolmanneksi Lahtinen käsittelee HTML-dokumentin rakenteeseen liittyviä optimointikeinoja: WWW-sivun otsikon tulee kuvata sen sisältöä, esimerkiksi tuotesivun otsikon tulisi sisältää tuotteen nimi. Tuotekuville tulee lisätä *alt*-attribuutti, jolla kuvaillaan kuvan sisältöä. Hakukoneet eivät osaa käsitellä kuvien sisältöä tarkasti, joten esimerkiksi tuotenimen käyttäminen *alt*-attribuuttina auttaa hakukonetta WWW-sivun indeksoinnissa. HTML-merkintäkielen otsikko-elementtejä (esimerkiksi h1 ja h2) tulee käyttää siten, että myös ne sisältävät WWW-sivun sisältöä kuvaavia asiasanoja. Lahtinen (2013) käsittelee myös metadatan lisäämistä HTML-dokumenttiin; tosin hän keskittyy *meta*-elementin käyttämiseen ja siihen miten sillä saadaan lisättyä WWW-sivulle kuvausteksti.

Hakukoneoptimointiin liittyvän teknisen osuuden Lahtinen päättää robots.txt- ja sitemap.xml-tiedostojen esittelyyn. Tiedostoista ensimmäisellä hakukoneita voidaan ohjeistaa jättämään verkkokaupan osiota indeksoimatta. Tosin robots.txt-tiedostolla voidaan vaikuttaa ainoastaan sen sisältöä kunnioittaviin hakukoneisiin. Toisella tiedostolla kerrotaan hakukoneille URL-osoitteet, jotka ne voivat käsitellä ja samalla URL-osoitteiden osoittamalle sisällölle voidaan määritellä painoarvo desimaaliluvulla nollan ja yhden väliltä. Sivustokartta, eli sitemap.xml-tiedosto, lähetetään Googlelle verkkovastaavan työkalulla (Googlen työkalu WWW-sivuston ylläpitäjälle). (Lahtinen, 2013.) Tämä ei kuitenkaan ole ainut keino. Hakukoneita voidaan myös kehottaa hakemaan sivustokartta WWW-sivustolta tai sen sijainti voidaan lisätä robots.txt-tiedostoon, jolloin hakukoneet hakevat tiedoston omatoimisesti (sitemaps.org,

2008).

Kuluttajien tekemän tiedonhaun lisäksi sijainti hakutuloksissa on tärkeää myös siksi, että kaikki asiakkaat eivät välttämättä siirry WWW-selaimella suoraan verkkokauppaan. Lahtinen (2013) esittää hyvän huomion WWW-selaimen käytöstä: kaikki käyttäjät eivät välttämättä kirjoita WWW-sivuston verkkotunnusta sille varattuun tekstikenttään, vaan käyttävät WWW-selaimen hakukenttää tai hakukoneeseen osoittavaa aloitussivua siirtyäkseen WWW-sivustolle. Tätä kutsutaan niin sanotun siirtymähakusanan käytöksi, jolloin käyttäjä ei pyri varsinaisesti löytämään mitään, vaan käyttää hakukonetta tietämälleen WWW-sivustolle siirtymiseen. Tämän vuoksi on syytä seurata verkkokaupan verkkotunnuksella tehdyn haun tuloksia ja pitää huolta, että verkkokauppa pysyy ensimmäisenä tai ensimmäisen sivun tulosten joukossa.

4 Semanttinen web ja verkkokaupat

Semanttisen webin tavoite datan muodostamasta verkosta on jossain määrin verrattavissa yritysten sähköisen liiketoiminnan järjestelmiin, jotka voivat koostua rajapintojen kautta toisiinsa integroiduista eri toimijoiden tuottamista palveluista. Suurten yritysten tietojärjestelmät voivat olla jopa suurempia kuin koko internet viisitoista vuotta sitten, koska järjestelmiin liittyviä tahoja on lukuisia, esimerkiksi tavarantoimittajat, pankit ja asiakaspalvelujärjestelmät. Semanttiseen webiin liittyvien teknologioiden käyttö ei rajoitu ainoastaan WWW:hen, vaan niitä voidaan käyttää myös sähköiseen liiketoimintaan liittyvissä järjestelmissä. Standardoituja metadatan merkintätapoja voidaan käyttää järjestelmien välillä vaihdettavan datan kuvailuun, mikä voi merkittävästi vähentää järjestelmien integrointiin vaadittavaa työmäärää ja siten myös integroinnista koituvia kustannuksia. Käytännössä semanttisen webin teknologioiden laajempi käyttö voi johtaa siihen, että liiketoiminnan järjestelmät toteutetaan SOA-arkkitehtuurin (Service Oriented Architecture) mukaisesti ja järjestelmiin liittyvät toiminnot ostetaan yksittäisiltä palveluntuottajilta. (Benjamins et al., 2011.)

Edellä kuvattu koskee lähinnä yritysten välistä kaupankäyntiä, mutta nopeasti yleistynyt informaatioteknologian käyttö on muuttanut myös kuluttajakaupankäyntiä ja hiljalleen ohjannut ostotottumuksia verkkokauppojen suuntaan. Matkustaminen ja turismi ovat hyviä esimerkkejä muuttuneista kuluttajakaupan osa-alueista. SVT:n (2014) tekemässä tutkimuksessa väestön tieto- ja viestintätekniikan käytöstä matkaliput ja majoitus ovat kaksi viidestä suosituimmasta tuoteryhmästä, joita hankitaan verkkokauppojen kautta. Lentokoneyhtiöiden ja hotellien siirryttyä tarjoamaan palveluitaan WWW:ssä suoraan kuluttajille on kysyntä perinteisten matkatoimistojen palveluille vähentynyt selvästi, kun kuluttajat voivat etsiä matkaan liittyvää informaatiota suoraan verkosta ja tehdä hankinnat suoraan palveluiden tuottajilta tai välittäjäpalveluiden kautta (Grün et al., 2011).

Toisaalta itse järjestettyjen matkojen yhteydessä kuluttaja joutuu käyttämään paljon aikaa matkakohteeseen ja matkustamiseen liittyvien yksityiskohtien selvittämiseen: kuluttaja siirtyy matkustaessaan pois tutusta ympäristöstä, ja hankittuja palveluita, esimerkiksi hotelliyötä, on mahdotonta arvioida ennen niiden käyttämistä – tämän vuoksi ennen matkaa etsityn tiedon merkitys kasvaa entisestään. Semanttisen webin teknologioiden avulla voidaan tuottaa palveluita, jotka auttavat kuluttajaa arvioimaan eri vaihtoehtoja ja näin tukevat häntä

päätöksenteossa. Esimerkiksi Reisewissen-projektissa[4] kehitettiin palvelu, joka käyttää semanttisen webin teknologioita hotellin etsinnän helpottamiseksi. Palvelu muuntaa käyttäjän antamat hakukriteerit semanttiseksi hauksi, jonka tuloksena käyttäjälle esitetään lista sopivista hotelleista. Toinen samankaltainen projekti on TrustYou-sivusto[5], joka tarjoaa myös hotellien hakutoiminnon. TrustYou:n toiminta perustuu useista lähteistä (Trip Advisor, Expedia, Qype) koottuihin hotellien käyttäjäarvioihin, joista palvelu muodostaa semanttista dataa NLP-tekniikoilla (Natural Language Processing). (Grün et al., 2011.)

Verkkokauppojen yhteydessä semanttisen webin teknologioita voidaan hyödyntää vastaavalla tavalla. Kuvailemalla verkkokauppojen sisältöä ontologioilla mahdollistetaan kuvaillun datan avulla tuotettujen palveluiden kehittäminen ja olemassa olevien palveluiden parantaminen; esimerkiksi auttamalla hakukoneita verkkokaupan sisällön jäsentämisessä mahdollistetaan tarkemman informaation tarjoaminen tiedon hakua suorittaville kuluttajille.

4.1 GoodRelations – ontologia tuotteille verkossa

Tuotteiden luokitteluun ja kuvailuun on kehitetty useita standardeja, esimerkiksi yhdysvaltalainen UNSPSC (United Nations Standard Products and Services Code) ja saksalainen eCl@ss. UNSPSC on yhdysvaltalainen luokittelujärjestelmä tuotteiden ja palveluiden kuvailemiseen sähköisen kaupankäynnin järjestelmissä. Eurooppalainen kilpailija UNSPSC-standardille on eCl@ss-standardi, jonka ensimmäinen versio on julkaistu vuonna 2000 ja jota on sen jälkeen päivitetty useaan otteeseen vastaamaan toimialojen vaatimuksia. Standardin 9.0-versiossa on 40 800 tuoteluokkaa ja 16 800 tuoteominaisuutta. eCl@ss-standardi on osa ETIM-hanketta (Electro-Technical Information Model), jonka tavoitteena on ottaa käyttöön kansainvälinen ETIM-standardi sähkötuotteiden kuvailemiseen (ETIM, 2015). ETIM-standardin käyttöönottamiseksi on myös Suomessa aloitettu projekti STK-liiton (Sähköteknisen Kaupan Liitto) toimesta (STK, 2015).

ECl@ss- ja UNSPSC-standardien ilmaisuvoima ei kuitenkaan riitä täyttämään sähköisen kaupankäynnin vaatimuksia semanttisessa webissä, vaikka niille on semanttisen webin tutkimuksen yhteydessä määritelty vastaavat ontologiat OWL-kielellä: eClassOWL ja unspscOWL. Luokitusjärjestelmien ongelmana on, että semanttisessa webissä ontologialta vaaditaan enemmän kuin mahdollisuutta kuvailla ”resurssin X olevan tuoteluokan Y ilmentymä”. GoodRelations-ontologia on Martin Heppin ja hänen kollegoidensa (2008)

4 http://reisewissen.ag-nbi.de/

5 http://www.trustyou.com/

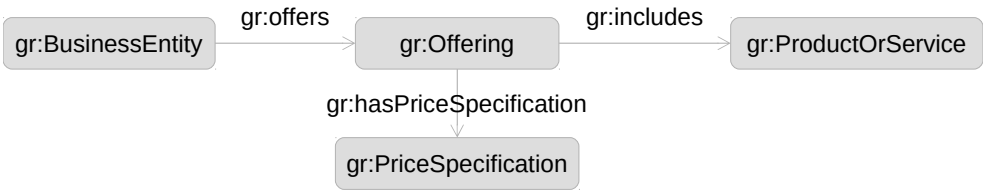
ratkaisu tuotteiden ja palveluiden kuvailemiseen semanttisessa webissä.

GoodRelations-ontologiaa on kehitetty vuodesta 2005 lähtien ja kehitystyössä on otettu huomioon eri toimialojen asettamia vaatimuksia. Ontologiaa on myös testattu käyttämällä sitä suuren tuotemäärän kuvailuun. GoodRelations pohjautuu RDF-, RDFS- ja OWL DL-kielien termeille, ja sen rakenne on pidetty mahdollisimman kevyenä. Ontologian kevyen rakenteen vuoksi sillä kuvaillun datan jäsentämiseen riittää RDFS-yhteensopivuus. Ratkaisu GoodRelations-ontologian rakenteeseen johtuu osittain tutkimuksen ajankohtana käytössä olleista tekniikoista ja siitä, että tutkijat halusivat säilyttää mahdollisuuden laajentaa eClassOWL-ontologiaa GoodRelations-ontologiassa määritellyillä termeillä. (Hepp, 2008.)

Kehitystyön lähtökohdaksi GoodRelations-ontologialle asetettiin neljä käyttötapaa: (1) Ontologialla kuvaillaan asiakkaalle tai yritykselle myytävää tuotetta, joka voi olla suoraan ostettavissa tai siitä esitellään ainoastaan tuotetiedot. (2) Ontologialla kuvaillaan tuotteen merkki, malli ja tekniset tiedot. Tuotetiedot voivat olla valmistajan määrittelemiä ja myytävien tuotteet tiedot sisältävät, mahdollisesti toisaalla, määritellyt tuotekuvaukset. Myytävillä tuotteilla voi olla myös lisäominaisuuksia kuten valmistuspäivämäärä tai sarjanumero. (3) Ontologialla kuvaillaan tuotevalikoimaa, josta asiakas voi vuokrata tuotteen. Tarjottava tuotevalikoima määräytyy tuoteluokan, merkin, mallin ja ominaisuuksien arvoalueen (range, esim. vaatteen koko) perusteella. (4) Ontologialla kuvaillaan tarjottavaa palvelua. Tarjottava palvelu (esim. huoltaminen, korjaaminen tai hävittäminen) kohdistuu tuotevalikoimaan, joka määräytyy tuoteluokan, merkin, mallin ja ominaisuuksien arvoalueen perusteella. (Hepp, 2008.)

GoodRelations-ontologian tavoitteena on määritellä tietorakenne, jota voidaan käyttää kaikilla toimialoilla. Tietorakennetta pitää pystyä käyttämään tuotteen koko toimitusketjussa, raaka-aineista jälleenmyyntiin ja aina tuotetukeen asti. Tietorakenne ei myöskään saa olla syntaksiin sidottu, eli sen pitää toimia kaikkien olemassa olevien ja mahdollisesti tulevien semanttisen webin merkintäkielien kanssa, esimerkiksi Microdata, RDFa, RDF/XML. (Hepp, 2011)

Asetetut tavoitteet saavutetaan GoodRelations-ontologiassa neljän pääkäsitteen ympärille rakennetun logiikan avulla. Agentilla (agent) määritellään kaupankäynnin osapuoli, eli asiakas tai yritys. Objektilla (object) määritellään tuote tai palvelu. Lupauksella (promise, offer) määritellään tuotteeseen tai palveluun liittyvän oikeuden (omistaminen, väliaikainen käyttö, lisenssi) luovuttaminen hyvitystä (compensation) vastaan. (Hepp, 2011.)

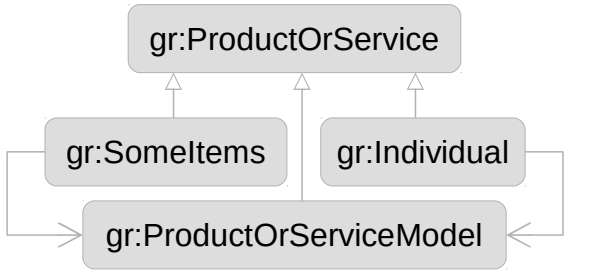


Kuva 8: Luokkakaavio GoodRelations-ontologian pääluokista.

Ontologian pääkäsitteitä vastaavat luokat on nimetty seuraavasti[6]:

- ”gr:BusinessEntity”-luokka vastaa agenttia.
- ”gr:Offering”-luokka vastaa lupausta (tarjousta) myydä, korjata tai vuokrata jotain.
- ”gr:ProductOrService”-luokka vastaa objektia, eli lupauksen kohteena olevaa tuotetta tai palvelua.
- ”gr:PriceSpecification”-luokka vastaa lupauksen täyttämisestä saatavaa hyvitystä. (Hepp, 2011.)

Kuvan 8 luokkakaavioon merkitty ontologian pääkäsitteet ja niitä yhdistävät termit.



Kuva 9: Luokkakaavio GoodRelations-ontologian tuoteluokista.

Verkkokauppojen kannalta GoodRelations-ontologian mielenkiintoisin kohde on ”gr:ProductOrService”-luokka, koska sen aliluokilla kuvaillaan tarjottavat tuotteet tai palvelut. Kuvassa 9 on GoodRelations-dokumentaation pohjalta luotu luokkakaavio, joka havainnollistaa tuoteluokkien perintää ja niiden välisiä suhteita.

Ontologian luokkarakenteen suunnittelussa on käytetty prototyyppi-suunnittelumallia. Tuoteluokista ”gr:ProductOrService” on tarkoitettu ainoastaan abstraktiksi luokaksi, josta perityillä luokilla tehdään varsinaiset tuotekuvailut – tosin dokumentaatiossa todetaan, että mikään ei varsinaisesti estä abstraktin luokan käyttämistä. Tuotteen kuvailuun on määritelty kaksi luokkaa, ”gr:Individual” ja ”gr:SomeItems”, joista ensimmäistä käytetään tuoteyksilön

6 CURIE-osoitteiden ”gr:”-etuliite viittaa ”http://purl.org/goodrelations/v1#”-merkkijonoon.

kuvailuun ja toista yleiseen tuotteiden kuvailuun. Tuoteyksilöllä tarkoitetaan yhtä yksittäistä tuotetta, jolle voidaan määritellä identifioiva tunniste, esimerkiksi sarjanumero. Esimerkiksi käytettyä autoa ostettaessa kyse on tietystä tuoteyksilöstä. Yleisellä tuotteiden kuvailemisella tarkoitetaan myytävien tuotteiden kuvailemista esimerkiksi verkkokaupassa, jolloin ei eritellä mitä tuoteyksilöä ollaan myymässä.

Ontologiassa on lisäksi määritelty ”gr:ProductOrServiceModel”-luokka, jota voidaan käyttää tuotetietojen kuvailuun. Luokan avulla määriteltyjä tuotetietoja on tarkoitus käyttää useiden tuotteiden ja tuoteyksilöiden kuvailemisessa, jolloin toistuvia tietoja ei tarvitse määritellä uudelleen – ajatus mukailee ideaa olioiden kloonaamisesta prototyyppi-suunnittelumallissa (Gamma et al., 1994). Tuotetiedot voidaan esimerkiksi julkaista valmistajan sivuilla, jolloin niillä voidaan täydentää myynnissä olevasta tuotteesta ilmoitettuja tietoja. ”gr:ProductOrServiceModel”-luokalla määriteltyjä tuotetietoja käytetään viittaamalla luokan ilmentymään ”gr:Individual”- tai ”gr:SomeItems”-luokan ilmentymästä (ks. kuva 9).

gr:ProductOrService

Jäsen	Tietotyyppi
gr:color	rdfs:Literal
gr:condition	rdfs:Literal
gr:hasBrand	gr:Brand
gr:hasGTIN-8	xsd:string
gr:hasManufacturer	gr:BusinessEntity
gr:isSimilarTo	gr:ProductOrService
gr:name	rdfs:Literal
gr:weight	gr:QuantitativeValue

Kuva 10: Osa "gr:ProductOrService"-luokan jäsenistä.

Ontologian yksi keskeisimpiä termejä on ”gr:ProductOrService”-luokka, jonka jäsenistä on poimittu muutamia kuvaan 10. Osa luokan jäsenistä viittaa ontologiassa määriteltyihin termeihin, esimerkiksi ”gr:hasManufacturer”-jäsenen arvona käytetään ”gr:BussinessEntity”-luokan ilmentymää. GoodRelations-ontologiassa monien luokkien jäsenten tietotyyppinä on merkkijono tai literaaliarvo, joista literaaliarvo useimmissa tapauksissa käytännössä tarkoittaa myös merkkijonoa. Usein ontologia määrittelee kuitenkin merkkijonon arvolle käytettävän standardin; esimerkiksi ”gr:hasGTIN-8”-jäsenen arvon pitää olla validi GTIN-numerosarja

(Global Trade Item Number), jota käytetään kauppanimikkeen yksilöimiseen – GTIN-koodi on aiemmin käytetyn EAN-koodin korvaaja (GS1, 2015).

”Gr:ProductOrService”-luokan jäsenistä on valittu kuvaan 10 ”gr:color”-jäsen, koska ontologia ei määrittele sen arvolle käytettävää standardia. Värijärjestelmän määrittelemättä jättäminen on jo sinänsä mielenkiintoista, koska ontologia kuitenkin käyttää joidenkin luokkien jäsenten kohdalla standardeja niiden arvoalueiden määrittelyyn, esimerkiksi aiemmin mainittua GTIN-standardia ja maakoodeille ISO 3166-1 -standardia. Värijärjestelmän puuttuminen on mielenkiintoinen myös siinä suhteessa, että ontologia ottaa huomioon erilaiset tuoteversiot tuoteluokkien rakenteen kautta; ”gr:ProductOrServiceModel”-luokan avulla voidaan esimerkiksi määritellä tuotetiedoiksi materiaali ja koko, jolloin ainoaksi erottavaksi tekijäksi voisi jäädä myytävien tuotteiden väri.

GoodRelations-ontologian suunnitelijoista Hepp on ollut määrittelemässä eClassOWL-ontologiaa (Hepp & Radinger). eCl@ss-standardi nimeää 285 väriä, joita voidaan käyttää väriarvoina, vaihtoehtoisesti väri on mahdollista määritellä värikoodin (color code) ja värijärjestelmän (color code system) yhdistelmänä (eCl@ss, 2014). Ontologioiden välisestä yhteydestä herää kysymys, miksi kumpaakaan värin määrittelytapaa ei ole otettu mukaan GoodRelations-ontologiassa.

GoodRelations-ontologian yhtenä tavoitteena on mahdollistaa tuotteiden semanttinen kuvaileminen HTML-dokumenteissa. Värin määrittelyyn olisi voitu ontologiassa ottaa mallia myös WWW-standardeista, koska HTML5-merkintäkieleen liittyvässä CSS3-merkintäkielessä (Cascading Style Sheets) värin arvo voidaan määritellä usealla eri tavalla ja yksi niistä on ennalta nimettyjen värien käyttäminen (Çelik et al., 2011). Ennalta määritellyt väriarvot olisi voitu lisätä myös GoodRelations-ontologiaan, koska siinä käytetään muutenkin enumeraatioita. Enumeraatiot ovat lueteltuja tietotyyppejä, jotka määritellevät valmiiksi kaikki mahdolliset luokan arvot (ilmentymät). Ontologiassa esimerkiksi tarjoukselle (”gr:Offer”) voidaan määritellä siihen oikeutettu kohderyhmä ”gr:BusinessEntityType”-enumeraatiolla (Hepp, 2011). Samaan tapaan ontologiassa olisi voitu määritellä sallitut väriarvot ja käyttää niitä tuotteen kuvailussa, ja mikäli ennalta määritellyt värit eivät olisi toimineet joissain tapauksissa, niin nykyistä vapaaseen tekstiin perustuvaa kuvailua olisi voitu käyttää vaihtoehtoisena tapana.

Toisaalta ontologian rakenteen kevyenä pitämistä voidaan pitää perusteluna värien suhteen tehdylle ratkaisulle – vapaan tekstin käyttö värin arvona on helppoa. Ontologian käyttäjien

pitää tosin olla tarkkoja, mikäli haluavat tuottaa tarkkaa semanttista dataa. Tuotetietojen kuvailussa pitää ottaa huomioon tilanne, jossa sivuston kieli eroaa tuotetiedoista. Tuotetietojen kielen eroaminen sivuston kielestä olisi mahdollista esimerkiksi siinä tapauksessa, että tuotetiedot tulevat suoraan tavarantoimittajalta ja niitä käytetään sivustolla ilman välissä tehtävää tarkistusta ja muunnosta. Mikäli tuotetietojen kieli eroaa sivuston kielestä eikä tuotetietoja voida kääntää automaattisesti, voidaan tällaisessa tilanteessa käyttää HTML-merkintäkielen *lang*-attribuuttia. *Lang*-attribuutilla voidaan ilmaista HTML-elementin kieli ja auttaa dataa jäsentävää järjestelmää ymmärtämään sen sisältö.

Sähköiseen liiketoimintaan kehitettyjen ontologioiden käyttöä on selvitetty OUSAF-kehyksen (Ontology Usage Analysis Framework) tutkimuksen yhteydessä. OUSAF-kehyksen tarkoitus on määritellä standardoitu tapa ontologioiden käytön mittaamiseen ja analysoimiseen. Kehyksen toimintaa testattiin tutkimusta varten kootulla datalla, joka koostui 211 eri WWW-sivustolta kerätystä 22,3 miljoonasta RDF-triplasta. Datalähteiksi valittiin WWW-sivustoja, joilla käytettiin sähköiseen liiketoimintaan suunniteltua ontologiaa. (Ashraf et al., 2014.)

Semanttisen datan tuottamistavan, ontologioiden käytön ja datan määrän perusteella tutkijat tunnistivat kolme semanttisen datan julkaisijaryhmää: suuret jälleenmyyjät, WWW-kaupat (web shops) ja data-palveluntarjoajat (data service providers). Suurilla jälleenmyyjillä tarkoitettiin toimijoita, joiden liiketoiminta perustuu lähinnä myymälöihin ja jotka ovat uusia verkkokauppojen hyödyntäjiä, esimerkiksi BestBuy. WWW-kaupoilla tarkoitettiin pieniä ja keskisuuria verkkokauppiaita, joiden liiketoiminta tapahtuu pääsääntöisesti verkossa. Data-palveluntarjoajilla tarkoitettiin toimijoita, jotka muuntavat yritysten omien järjestelmien datan semanttiseksi dataksi, esimerkiksi Linked Open Commerce[7]. (Ashraf et al., 2014.)

Tutkimuksen datalähteistä 97 % käytti GoodRelations-ontologiaa ja eCl@ss-ontologian käyttäjiä oli 18 % (Ashraf et al., 2014). Datalähteiden pienen määrän vuoksi tuloksesta ei voida tehdä vielä johtopäätöksiä siitä, miten paljon GoodRelations-ontologiaa käytetään koko WWW:ssä. Tulos kuvaa sitä, miten hyvin GoodRelations-ontologia on otettu käyttöön tutkimusdataan valituissa WWW-sivustoissa – tosin tutkimuksessa oli mukana isoja toimijoita, joten tuloksella voidaan perustella GoodRelations-ontologian sopivuutta verkkokauppojen tuotedatan semanttiseen kuvailemiseen.

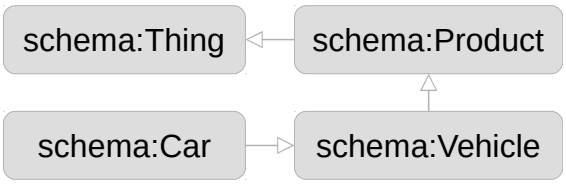
7 http://linkedopencommerce.com/

4.2 Schema.org – ontologia hakukoneille

World Wide Webin hajautettu rakenne ja informaation julkaisutapa muodostavat haasteen datan käyttäjille. Datan käsittely vaikeutuu, koska sisällön julkaisija voi vapaasti valita tavan kuvailla tuottamaansa sisältöä. Tämä hankaloittaa varsinkin datan automaattista indeksointia tekeviä hakukoneyhtiöitä, joiden on vaikea tulkita monin eri tavoin kuvailtua sisältöä.

Vuonna 2011 Google, Bing ja Yahoo ilmoittivat yhteisestä schema.org-hankkeesta, jonka tavoitteena on määritellä yhteinen termistö rakenteelliselle datalle verkossa. Schema.org-sivuston on tarkoitus toimia tietolähteenä kehittäjille, WWW-sivuston ylläpitäjille ja omistajille. (Ramanathan, 2011.) Vielä saman vuoden aikana myös venäläinen hakukoneyhtiö Yandex liittyi mukaan schema.org-hankkeeseen (TechCrunch, 2011).

Schema.org-tietomallissa kuvailtavia kohteita kutsutaan asioiksi (item), joiden rakenne on määritelty schema.org-dokumentaatiossa, esimerkiksi auto on määritelty ”http://www.schema.org/Car”-osoitteessa. Asiat rakentuvat niihin kuuluvista ominaisuuksista (property), jotka määritellään myös schema.org-sivustoon viittaavilla URI-osoitteilla, esimerkiksi ”http://www.schema.org/fuelType”[8]. Ominaisuuksien arvot voidaan antaa muun muassa merkkijonona ja muina literaaliarvoina, esimerkiksi kokonaislukuna. (Patel-Schneider, 2014.)

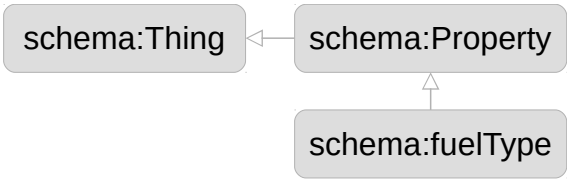


Kuva 11: Luokkakaavio ”schema:Car”-asian yläkäsitteistä.

Schema.org-tietomallin asiat vastaavat käytännössä ontologioiden luokkia. Asioiden kuvailemiseen käytetyt käsitteet on järjestetty tietomallissa hierarkiaksi, jossa varsinaiseen kuvailuun käytetyt käsitteet ovat yleensä alikäsitteitä, jotka on peritty niitä laveammista yläkäsitteistä. Schema.org-tietomallissa perityt käsitteet sisältävät niiden yläkäsitteiden ominaisuudet eli ne toimivat samaan tapaan kuin luokat ontologioissa. (Patel-Schneider, 2014.) Jokainen hierarkian käsitteistä tunnistetaan omalla URI-osoitteella – ne eivät tosin heijasta tyypin sijaintia hierarkiassa. Esimerkiksi ”http://www.schema.org/Car”-asia on hierarkiassa neljännellä tasolla, mutta URI osoittaa suoraan käsitteeseen. (schema.org, 2015b.)

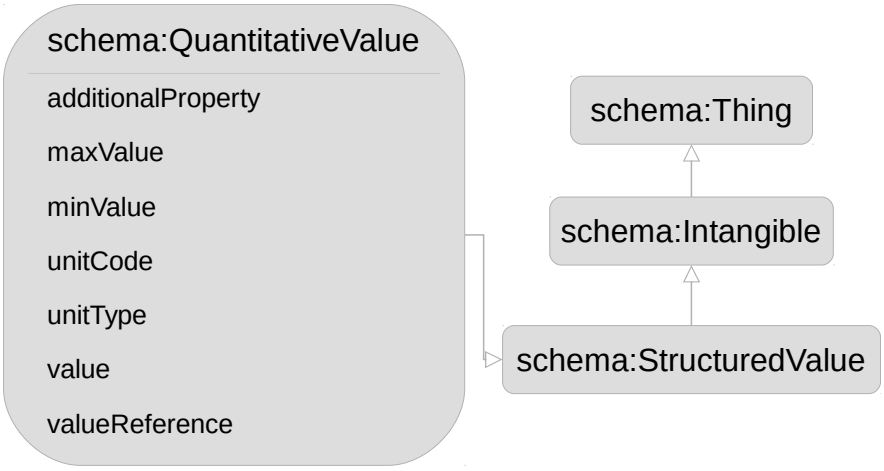
8 Jatkossa ”http://www.schema.org/”-osoitteille käytetään ”schema:”-etuliitettä.

Kuvassa 11 on esitetty luokkakaaviona ”schema:Car”-asian sijainti tietomallin käsitehierarkiassa.



Kuva 12: Luokkakaavio ”schema:fuelType”-ominaisuuden perimistä käsitteistä.

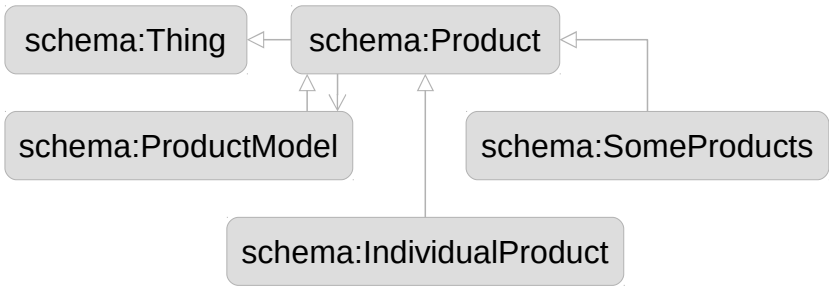
Tietomallin asiat koostuvat ominaisuuksista, joiden arvot kuvailevat kohdetta. Asioiden ominaisuudet ovat myös osa schema.org-tietomallin käsitehierarkiaa. Tosin ominaisuuksille ei muodostu syvää hierarkkista puurakennetta, vaan ne kaikki ovat ”schema:Property”-käsitteen alikäsitteitä. Kuvan 12 luokkakaavioon on otettu ”schema:fuelType”-ominaisuus esimerkkinä ”schema:Property”-käsitteen alikäsitteistä. Asioiden ominaisuudet eroavat asioista myös siinä suhteessa, että ne eivät peri ”schema:Property”-käsitteen ominaisuuksia. (schema.org, 2015b.)



Kuva 13: Luokkakaavio "schema:QuantitativeValue"-tietotyypistä ja sen yläkäsitteistä.

Aiemmin mainittujen literaaliarvojen lisäksi asioiden ominaisuuksien arvoille on määritelty myös tarkempia tietotyyppejä, jotka koostuvat useista ominaisuuksista. Ominaisuuksien arvojen tietotyypit sijaitsevat samassa hierarkiassa asioiden kanssa. Esimerkiksi ”schema:Car”-asian ”fuelConsumption”-ominaisuus annetaan ”schema:QuantitativeValue”-tietotyypin ilmentymänä, joka puolestaan voidaan määritellä esimerkiksi mittayksikön (unitCode tai unitType) ja arvon (value) yhdistelmällä. (schema.org, 2015b.) Kuvan 13 luokkakaaviossa on esitetty ”schema:QuantitativeValue”-tietotyypin jäsenet ja yläkäsitteet, joista tietotyyppi on peritty.

Schema.org-tietomallin dokumentaatio ja rakenne on saanut osakseen kritiikkiä. Dokumentaatiossa ei esimerkiksi selkeästi erotella asioita (item), ominaisuuksia (property) ja ominaisuuksien arvojen tietotyyppejä. Vaikka ominaisuudet ovat osa samaa hierarkiaa, niillä selvästi tarkoitetaan eri asiaa kuin asioilla, koska niiden nimet alkavat pienellä kirjaimella ja eroavat siten asioista ja ominaisuuksien arvojen tietotyypeistä. Asioiden ja ominaisuuksien arvojen tietotyyppien välinen ero on hieman epäselvä, dokumentaatiosta ei esimerkiksi selviä, voiko näitä tietotyyppejä käyttää asioina. Esimerkiksi ”schema:Brand”-tietotyyppi on aineeton (intangible) käsite, mutta joissain yhteyksissä sitä voitaisiin tarvita erillään (tuotemerkkilistaus verkkokaupassa). (Patel-Schneider, 2014.)



Kuva 14: Luokkakaavio schema.org-tietomallin tuotteisiin liittyvistä käsitteistä.

Verkkokauppojen tarpeisiin schema.org-tietomalli sisältää samankaltaisen rakenteen kuin GoodRelations-ontologia. Samankaltaisuus GoodRelations-ontologian kanssa johtuu siitä, että sen kehittäjä Martin Hepp on ollut aktiivisesti kehittämässä schema.org-tietomallia (schema.org, 2015a). Merkittävä ero ontologiaan on, että Schema.org-tietomallissa tuotteet ja palvelut on eriytetty. Myös tuotteisiin liittyvä käsitehierarkia on muuttunut: tietomallissa ”gr:SomeItems”-luokkaa vastaa ”schema:Product”-asia ja ”schema:ProductModel”-luokan ilmentymällä kuvaillut tuotetiedot liitetään suoraan siihen (kuva 14). Lisäksi tietomalliin on lisätty uusi käsite, ”schema:SomeProducts”, jota käytetään tuote-erän kuvailuun. Myös tuotekäsitteiden ominaisuuksia on nimetty schema.org-tietomallissa uudelleen ja niiden tietotyyppejä on muutettu; esimerkiksi tuotteen ”gr:condition”-ominaisuus on schema.org-tietomallissa ”schema:itemCondition” ja tuotteen kunto annetaan enumeraation arvona aikaisemman vapaan merkkijonoarvon sijaan. (schema.org, 2015b.) Tehdyt muutokset vaikuttavat hyviltä, ainakin tuotteisiin liittyvien käsitteiden hierarkia on selkeämpi kuin GoodRelations-ontologiassa, kun rakennetta hieman häirinnyt abstrakti luokka on poistettu.

```
 1 <!DOCTYPE html>
 2 <html>
 3 <head>
 4   <meta charset="UTF-8">
 5   <meta name="description" content="Volkswagen myynti-ilmoitus">
 6   <meta name="keywords" content="Volkswagen,myynti-ilmoitus">
 7   <title>Volkswagen Kupla myynnissä</title>
 8 </head>
 9 <body>
10   <main itemscope itemtype="http://schema.org/Car">
11     <meta itemprop="name" content="Volkswagen 1300S">
12     <h1>Volkswagen Kupla myynnissä</h1>
13     <p>Myynnissä vuoden 1972 1300S, joka ei suostu käynnistymään.</p>
14     <table>
15       <tr>
16         <td>Väri</td>
17         <td itemprop="color">Valkoinen</td>
18       </tr>
19       <tr itemprop="mileageFromOdometer" itemscope
20           itemtype="http://schema.org/QuantitativeValue">
21         <td>Mittarilukema</td>
22         <td>
23           <meta itemprop="unitCode" content="KTM">
24           <meta itemprop="value" content="168000.00">
25           168 000 km
26         </td>
27       </tr>
28       <tr itemprop="offers" itemscope
29           itemtype="http://schema.org/Offer">
30         <td>Hintapyyntö</td>
31         <td>
32           <meta itemprop="priceCurrency" content="EUR">
33           <meta itemprop="price" content="550.00">
34           550€
35         </td>
36       </tr>
37     </table>
38   </main>
39 </body>
40 </html>
```

Listaus 3: Myynti-ilmoitus semanttisella datalla.

Listauksessa 3 on HTML5-muodossa aiemmin esimerkkinä käytetty auton myynti-ilmoitus, johon on lisätty schema.org-dokumentaation mukaista semanttista dataa. Listauksen HTML-rakenne eroaa jonkun verran XHTML-versiosta (listaus 1, sivulla 7), mutta sisältö on pysynyt samana. Suurin ero esimerkkien välille tulee lisätyistä *meta*-elementeistä, joilla semanttinen data on saatu upotettua HTML-dokumenttiin. Ilman *meta*-elementtien käyttämistä dokumentin rakennetta olisi pitänyt muuttaa tai ympäröidä leipätekstin sanoja *span*-elementeillä, jotta tarvittavat Microdata-attribuutit olisi saatu lisättyä dokumenttiin. Kun esimerkin HTML-dokumentista jäsennetään sen sisältämä semanttinen data, voidaan sen perusteella todeta, että dokumentti sisältää myynti-ilmoituksen 168 000 kilometriä (”mileageFromOdometer”) ajetusta valkoisesta ”Volkswagen 1300S”-autosta (”schema:Car”), josta pyydetään 550 euroa (”price” ja ”priceCurrency”).

Schema.org-dokumentaatio määrittelee joidenkin ominaisuuksien kohdalla niiden arvoille käytettävän standardin; esimerkiksi listauksessa 3 käytetty ”schema:mileageFromOdometer”-ominaisuus määritellään ”schema:QuantitativeValue”-tietotyypillä, joka voidaan antaa

mittayksikön (unitCode tai unitType) ja arvon (value) yhdistelmänä. Mittayksikölle käytetään UN/CEFACT-organisaation[9] ”Codes for Units of Measure Used in International Trade”-suosituksen mukaisia koodeja, esimerkin ”KTM”-koodi vastaa kilometri-mittayksikköä (UN/CEFACT, 2015).

Tosin schema.org jättää, GoodRelations-ontologian tapaan, määrittelemättä väreille käytettävän standardin. Väristandardin puuttuminen tietomallin määrittelystä on sinänsä outoa, koska voisi olettaa, että hakukoneyhtiöillä olisi tavoitteena saada sisällöntuottajat kuvailemaan WWW-sivustojen sisältö mahdollisimman tarkasti. Tarkasti kuvailtu sisältö mahdollistaisi sisällön tarkemman jäsentämisen, ja indeksoitua informaatiota voitaisiin hyödyntää hakukoneissa monipuolisemmin.

Semanttisten merkintäkielien (Microdata, Microformats ja RDFa) ja schema.org-tietomallin käyttöä tutkittiin Common Crawl -organisaation keräämään dataan pohjautuneessa tutkimuksessa (Meusel et al., 2014). Tutkimuksessa käytettiin vuosien 2010, 2012 ja 2013 aineistoja, joista vuoden 2013 aineistossa oli 12,8 miljoonalta WWW-sivustolta kerättyä 2,2 miljardia HTML-dokumenttia. Vuoden 2013 aineiston WWW-sivustoista 13,87 % ja HTML-dokumenteista 26,33 % sisälsi vähintään yhtä semanttista datatyyppiä, esimerkiksi schema.org mukaista semanttista kuvailua. (Meusel et al., 2014.)

Semanttinen data jäsennettiin tutkimuksessa RDF-nelikoiksi (RDF-quad), joissa neljäs tekijä oli datan lähde, eli WWW-sivusto. Vuoden 2013 aineistosta tunnistetuista RDF-nelikoista suurin osa oli merkitty Microdata-formaatilla, jonka käyttö olikin vähintään nelinkertaistunut vuoden 2012 lukuihin verrattuna. Tuotetietueita vuoden 2013 aineistossa oli yhteensä 202 miljoonaa, kerättynä 71000 WWW-sivustolta, ja niiden kuvailuun oli käytetty ylivoimaisesti eniten (80 %) schema.org-tietomallin mukaista kuvailua. (Meusel et al., 2014.) Tutkimuksen perusteella Microdata-formaatin käyttö on suositeltavaa, varsinkin tuotetietojen kuvailussa, jolloin voidaan käyttää schema.org-määrityksen mukaista rakennetta semanttiselle datalle. Tutkimustuloksia voidaan pitää luotettavina, koska aineistosta oli karsittu tuloksia vääristäviä sivustoja, esimerkiksi aikuisviihdesivustoja, joten tutkimuksen löydökset perustuvat normaaleihin WWW-sivustoihin.

4.3 Semanttinen data verkkokaupoissa

Schema.org-tietomallin mukaisen semanttisen datan lisääminen verkkokauppaan on suositeltavaa, koska se auttaa hakukonetta tiedon jäsentämisessä ja siten parantaa

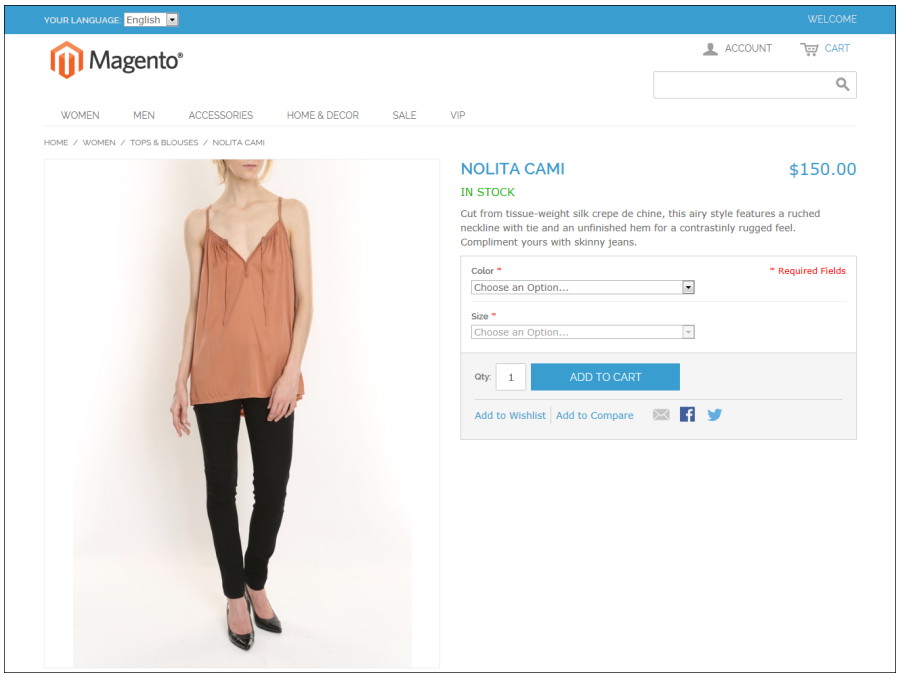
9 United Nations Centre for Trade Facilitation and Electronic Business

verkkokaupan mahdollisuuksia saada parempi sijoitus hakutuloksissa. Hakukoneyhtiöt suosittelevat schema.org-datan lisäämiseen Microdata-formaattia ja varoittavat muiden semanttisen datan formaattien samanaikaisesta käytöstä, koska useiden formaattien käyttäminen voi aiheuttaa virheitä tiedon jäsentämisessä. (Fishkin & Høgenhaven, 2013.) Suomalaisissa verkkokaupan ja sähköisen kaupankäynnin käsikirjoissa käsitellään hakukoneoptimointia osana verkkokauppojen markkinointia, mutta esimerkiksi Lahtinen (2013) ei mainitse schema.org:in mukaisen metadatan käyttöä hakukoneoptimoinnin laajasta käsittelystä huolimatta. Kuitenkin vähintään isoimmat suomalaiset verkkokaupat, kuten Verkkokauppa.com, tuottavat schema.org-dataa, joten ohjeistamisen puuttuminen schema.org:in sekä muun semanttisen datan tuottamiseen on selvä parannusta vaativa kohta kotimaiseen sähköisen kaupankäynnin kirjallisuuteen (vrt. Lahtinen, 2013; Anders Innovations, 2014; Havumäki & Jaranka, 2014).

Kuva 15: Esimerkki Googlen hakutulosriveistä, joista ensimmäisen muotoiluun on käytetty semanttista dataa.

Hakukonelistauksessa semanttista dataa tuottavat WWW-sivustot erottuvat yleensä selvästi WWW-sivustoista, jotka eivät sitä tuota. Semanttista dataa tuottavista WWW-sivuista hakukoneet osaavat näyttää perustietojen lisäksi semanttisessa datassa annettuja tietoja, esimerkiksi tuotteen hinnan tai tuotteen saamien arvioiden keskiarvon. Kuvassa 15 on esimerkki Googlella tehdystä hausta, jonka ensimmäisen hakutuloksen osoittama WWW-sivustolta on otettu huomioon schema.org:in mukainen semanttinen data. Muotoiltuja hakutuloksia, jotka sisältävät hakukoneen indeksoimaa semanttista dataa, kutsutaan eri nimillä hakukoneyhtiöstä riippuen, esimerkiksi Google kutsuu niitä ”Rich Snippet”-termillä. Hakukoneyhtiöt eivät kuitenkaan ota huomioon mitä tahansa semanttista dataa hakutuloksia muotoillessaan, vaikka semanttinen data olisi schema.org-tietomallin mukaista, vaan tuki on rajoitettu muutamaan tietomallissa määriteltyyn luokkaan. (Bing; Google, 2015a.)

Verkkokaupan tuotteiden kuvailemiseen voidaan käyttää kaikille tuotteille sopivaa ”schema:Product”-luokkaa. Yhden luokan käyttäminen yksinkertaistaa semanttisen datan tuottamista, sillä järjestelmältä ei vaadita tietoa tuotteen tyypistä. Mikäli verkkokauppa myy useita erilaisia tuotteita – esimerkiksi kodinkoneita, kirjoja ja cd-levyjä – niin schema.org-tietomallista löytyy myös tarkempia luokkia niiden kuvailemiseen. Kaikki verkkokaupan tuotteiden kuvailuun sopivat schema.org-tietomallin luokat eivät ole yllättäen ”schema:Product”-luokasta johdettuja. Ainoastaan ”schema:Vehicle” ja ”schema:Car” löytyvät ”schema:Product”-luokan alta, ja esimerkiksi kirjakauppoihin soveltuva ”schema:Book”-luokka on peritty ”schema:CreativeWork”-luokasta. Semanttisen datan verkkokauppaan lisäämistä suunniteltaessa onkin syytä käydä läpi schema.org-dokumentaatio ja tarkistaa, onko myytäville tuotenimikkeille mahdollisesti määritelty tarkempia luokkia. Jos tarkemman luokan käytön esteeksi muodostuu tuotetyypin puuttuminen järjestelmään tallennetuista tuotetiedoista, niin ratkaisua voidaan hakea esimerkiksi kehittämällä tuotekategorioihin pohjautuva algoritmi. Tarkempia schema.org-luokkia käytettäessä kannattaa ottaa huomioon hakukonekohtaiset rajoitukset siinä, miten semanttista dataa näytetään hakutuloksissa; mikäli semanttisella datalla muotoillut hakutulokset ovat tavoitteena, kannattaa mahdollisesti käyttää ainoastaan yleistä ”schema:Product”-luokkaa.



Kuva 16: Esimerkki tuotesivusta, jolla voidaan valita tuotteen ominaisuudet (Magento, 2015).

Riippuen siitä, miten tuotetiedot esitetään verkkokaupassa, tuotteiden semanttisessa kuvailussa voi ilmetä ongelmia. Mikäli verkkokaupassa myydään tuotteita, joista on tarjolla

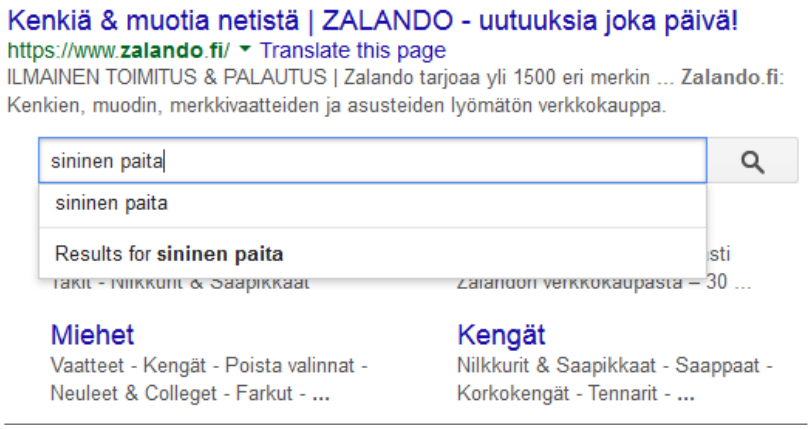
useita eri malleja, voidaan tuotteiden kohdalla tehdä kaksi ratkaisua: jokaiselle tuotemallille näytetään oma tuotesivu tai käytetään yhtä kaikille tuotemalleille yhteistä tuotesivua, jolla kuluttaja valitsee ostettavan tuotteen ominaisuudet (ks. kuva 16). Kaikille tuotemalleille yhteisen tuotesivun semanttinen merkitseminen on ongelmallista, koska teknisesti tarkasteltuna WWW-sivu sisältää vain yhden tuotteen ja tuotemalleihin liittyvät ominaisuudet ovat listattuna sen yhteydessä, esimerkiksi pudotusvalikoissa. Verkkokauppa kuitenkin myy kaikkia tuotemalleja erikseen ja todennäköisesti hyötyisi niiden kaikkien hakukonelistautumisesta, esimerkiksi kuluttajan hakiessa tietyn väristä mallia paidasta.

Yksi ratkaisu tuotevalintojen semanttiseen merkintään on lisätä semanttinen data tuotesivulle *meta*-elementeillä, jolloin hakukone voi jäsentää sivulta kaikki tuotemallit ja samalla verkkokaupan tuotekatalogi pysyy siistinä, kun sitä ei täytetä saman tuotteen eri malleilla. Vaikka yhden tuotesivun käyttäminen kaikille tuotemalleille on semanttisen datan tuottamisen kannalta hieman vaikeampaa, on sen käyttäminen verkkokaupan hakukonelistautumisen kannalta järkevämpää kuin tuotemallikohtaisten sivujen käyttäminen. Tuotemallikohtaisissa sivuissa suurin osa sivun sisällöstä pysyy samana, joten on vaarana, että hakukone vähentää sivun arvoa tulkitsemalla sen kopioiduksi sisällöksi, jolla yritetään parantaa verkkokaupan sijoitusta hakutuloksissa (Fishkin & Høgenhaven, 2013).

Verkkokaupan semanttisia merkintöjä suunniteltaessa on hyvä huomioida, ettei schema.org-dokumentaatio määrittele, mitkä luokan ominaisuuksista ovat pakollisia. Käytännössä schema.org-tietomallin käyttäjillä on täysi vapaus käyttää mitä tahansa luokan ominaisuuksista, tosin ei välttämättä ole hyödyllistä lisätä semanttista dataa esimerkiksi ainoastaan tuotteen hinnalle. Semanttisen datan käytöstä tehdyn tutkimuksen (Meusel et al., 2014) mukaan tuotteita kuvailtaessa tuotteille annettiin yleensä ainoastaan vain muutama semanttisen datan tietue, noin 50 % tuotetiedoista sisälsi ainoastaan tuotteen nimen, hinnan ja tuotekuvan. Ainoastaan 15 %:lle tuotteista oli määritelty ”productID”-ominaisuus (Meusel et al., 2014), joten pelkän semanttisen datan pohjalta on vaikea luoda palvelua, joka indeksoisi tuotteita ja mahdollistaisi esimerkiksi niiden hinnan vertailemisen. Mikäli semanttisen datan tarkkuus paranee verkkokaupoissa, on hyvin todennäköistä, että tarjolla olevan datan avulla tuotetaan palveluita ja palvelukohtaiset integraatiot vähentyvät. Esimerkiksi hintavertailua tarjoavaan vertaa.fi-palveluun verkkokaupat liittyvät palvelua varten tuotetun XML-tiedoston avulla[10].

10 http://www.vertaa.fi/info/verkkokauppanne/

Semanttista dataa voitaisiin tälläkin hetkellä hyödyntää verkkokaupoissa kilpailijoiden seurantaan, esimerkiksi Suomessa tämä olisi melko helppoa toteuttaa, koska tietyillä aloilla ei välttämättä ole montaakaan kilpailijaa. Automaattista valvontaa voitaisiin hyödyntää esimerkiksi tuotteen hinnan seuraamiseen kilpailijan verkkokaupassa. Tosin reagointi hinnan muuttumiseen ei kannata olla täysin automaattista, koska hinnan muutos voi välillisesti maksaa paljon enemmän kauppiaalle, jos kappalemyynti ei muutu hinnanmuutokseen suhteutettuna tarpeeksi. Toinen ongelma automaattisessa hinnan muuttamisessa muodostuu siinä tapauksessa, että myös kilpailija käyttää vastaavaa järjestelmää ja järjestelmien reagoidessa toistensa tekemiin hinnanmuutoksiin lopputulos voi olla erittäin yllättävä. Yksi esimerkki toisiaan vastaan kilpailleista järjestelmistä on biologiantutkimuksessa käytetyn kirjan hinnan nousu 23 miljoonaan dollariin: Michael Eisen (2011) seurasi kahden Amazonissa myytävän ”The Making of a Fly”-kirjan kopion hinnanmuutoksia. Kymmenen päivän aikana noin sadan dollarin hintaisen kirjan hintapyynti nousi moninkertaiseksi, koska kahden kauppiaan järjestelmät käyttivät kolmannen osapuolen toimittamia algoritmeja tuotteiden hintojen automaattiseen muutokseen.



Kuva 17: Esimerkki verkkokaupan hakutoiminnon näkymisestä Googlen hakutuloksissa.

Verkkokauppojen semanttisen datan tuottaminen ei rajoitu ainoastaan tuotetietoihin, vaan schema.org-tietomallissa on myös muita verkkokauppaympäristöön sopivia luokkia. Esimerkiksi ”schema:SearchAction”-luokalla voidaan kertoa verkkokaupan hakutoiminnon URL-osoite. Kun hakutoiminnon URL-osoite on kuvailtu semanttisella datalla, niin Google näyttää verkkokaupan verkkotunnukseen osoittavan hakutulosrivin yhteydessä tekstikentän, jota käyttämällä voidaan siirtyä suoraan verkkokaupan hakutoiminnon tuottamalle hakutulossivulle. (Google, 2015b.) Kuvassa 17 on esimerkki hakutoiminnon näkymisestä Googlen hakutuloksissa. Jos hakukentässä painettaisiin rivinvaihto-näppäintä tai haku-

painiketta, siirtyisi WWW-selain zalando.fi-verkkokauppaan ja näyttäisi verkkokaupan suorittaman sanahaun tulokset.

Schema.org-tietomallin lisäksi verkkokaupoissa kannattaa ottaa huomioon myös muut semanttista dataa hyödyntävät tahot, esimerkiksi sosiaalinen media. Esimerkkinä muista semanttisen datan määrittelyistä voidaan käyttää Open Graph -protokollaa, jota tukevat muun muassa Facebook ja Google+ (Facebook, 2014). Kun Facebookin julkaisuun lisätään URL-osoite, hakee Facebook URL-osoitteen takana olevan HTML-dokumentin ja poimii siitä olennaisen osan julkaisuun. Ongelmaksi voi muodostua Facebookin algoritmi, joka päättelee HTML-dokumentin sisällöstä olennaisimman osan – algoritmin toimintaan voidaan onneksi vaikuttaa Open Graph -merkinnöillä. Open Graph -data lisätään HTML-dokumenttiin RDFa-merkinnöillä ja niillä voidaan määritellä muun muassa dokumentin otsikko, kuvaus ja Facebook julkaisuissa käytettävä WWW-sivuston nimi. (Facebook.) Semanttisen datan tutkimuksessa käsitellystä RDFa-datasta suurin osa oli Open Graph -määrittelyn mukaista (Meusel et al., 2014).

5 Lopuksi

Tutkielmassa esiteltyjen semanttisen webin teknologioiden avulla voidaan monipuolisesti kuvailla verkkokaupan sisältöä ja mahdollistaa kuvaillun informaation käyttäminen kolmansien osapuolien järjestelmissä. Semanttisen datan käyttäjistä tutkielmassa keskityttiin lähinnä hakukoneisiin, koska hakukonenäkyvyys on verkkokauppojen toiminnan kannalta tärkeää. Hakukonenäkyvyyden lisäksi semanttisella datalla voidaan myös muuten vaikuttaa verkkokaupan näkyvyyteen WWW:ssä, esimerkiksi sosiaalisessa mediassa. Käyttämällä sosiaalisen median sivustojen tukemia semanttisen datan formaatteja voidaan vaikuttaa siihen, miten julkaisuissa jaettu verkkokauppojen sisältö näytetään näissä palveluissa.

Tällä hetkellä verkkokauppojen tuottama semanttinen kuvailu ei ole kovin tarkkaa, eikä se esimerkiksi sisällä tuotetunnisteita. Mikäli verkkokauppojen tuottama semanttinen data jatkossa tarkentuu, voidaan semanttisen datan avulla mahdollisesti tuottaa uusia palveluita. Kuluttajien kirjoittamat tuotearvostelut voisivat olla yksi uuden palvelun lähtökohta. Jos tuotearvosteluita voitaisiin jäsentää tarkasti, esimerkiksi tunnistamaan luotettavasti arvioitu tuote, niin indeksoitujen arvostelujen pohjalta voitaisiin luoda palvelu, josta kuluttaja voisi lukea mitä muut tuotteen hankkineet ovat kirjoittaneet ostoksestaan. Palvelun tuottaminen ei olisi välttämättä aivan ongelmatonta, vaikka datan jäsentäminen onnistuisikin virheettömästi; ainakin tuotearvosteluiden luotettavuus nousee yhdeksi tekijäksi, joka tulee ratkaista. Esimerkiksi Amazon kärsii automatisoiduista tuotearvosteluista ja on aloittanut toimenpiteet neljää tuotearvosteluita tuottavaa palvelua vastaan (BBC News, 2015).

GoodRelations-ontologian käsittelyn yhteydessä huomattu puute tuotteen värin määrittelyssä voisi olla hyvä jatkotutkimuksen kohde. Yksi mahdollinen ratkaisu voisi olla eCl@ss-standardissakin käytetty värikoodin ja värijärjestelmän yhdistelmä, joka sallii värin määrittelyn HKS-, PANTONE- ja RAL-värijärjestelmillä (eCl@ss, 2014). Tarkemman määrittelyn lisääminen väreille olisi perusteltua, koska GoodRelations-ontologia tukee esimerkiksi tuotekoodin antamista useassa eri muodossa (GTIN-8-, GTIN-13- ja GTIN-14) – värien kohdalla tulisi tarjota sama mahdollisuus tarkan semanttisen datan tuottamiseen.

Myös Schema.org-tietomallin määrittely vaatisi tarkennusta tuotteen värin määrittelyn kohdalla. Tuotteen värin kuvaileminen luonnollisella kielellä voi muodostua ongelmaksi moniväristen tuotteiden kohdalla, koska kuvailuun voidaan käyttää monia erilaisia ilmaisuja, esimerkiksi ”punainen/musta”, ”punamusta” ja ”pun./must.”. Väriarvon kohdalla käytetty

merkkijono riippuu verkkokaupan kielestä ja siitä, miten verkkokaupan tuotteiden tiedot on tallennettu sen taustajärjestelmään. Schema.org-tietomallilla on tosin mahdollista kuvailla erikseen kaikki tuotteeseen kuuluvat värit, koska tietomalli ei rajoita semanttisessa datassa annettujen väriominaisuuksien lukumäärää. Schema.org-tietomallin määrittelystä puuttuu kuitenkin määrittely värien kuvailuun käytetylle standardille; mikäli schema.org-tietomallia tarkennettaisiin väriominaisuuden osalta, hakukoneyhtiöt voisivat tarkemman semanttisen datan pohjalta rakentaa esimerkiksi tuotehaun, jossa käyttäjä voisi selata tuotteita valitsemansa värin tai väriyhdistelmän perusteella.

Mavridis ja Symeonidis (2015) ovat tutkineet hakukonenäkyvyyteen vaikuttavia tekijöitä tätä varten kehitetyllä kehyksellä. Kehyksen toimintaa testattiin suorittamalla hakuja kahteen ennalta määriteltyyn aihepiiriin: ohjelmistotuotanto ja urheilu. Haut suoritettiin kolmessa hakukoneessa: Bing, Google ja Yahoo. Suoritetuille hauille tuloksena saatuja WWW-sivuja analysoitiin tarkemmin ja niistä etsittiin hakutulokseen vaikuttavia tekijöitä, esimerkiksi WWW-sivuun osoittavien sivuston sisäisten ja ulkoisten linkkien lukumäärää. Tämän jälkeen tutkimuksessa arvioitiin löydettyjen hakutulokseen vaikuttavien tekijöiden merkitystä. Vaikka kehyksen testien perusteella semanttisen datan merkitys hakukonenäkyvyyteen ei ollut yhtä merkittävä kuin esimerkiksi WWW-sivuun osoittavien linkkien lukumäärä, pystyivät tutkijat toteamaan, että hakukoneet ottavat semanttisen datan huomioon sisältöä arvioidessaan. Lisäksi testiin valituista hakukoneista Google ja Bing vaikuttivat antavan semanttiselle datalle enemmän painoarvoa kuin Yahoo.

Hakukonenäkyvyyteen liittyvän tutkimuksen perusteella ei voi vielä todeta, että semanttinen data toisi merkittävää hyötyä verkkokaupalle jo tutkimuksen kapean kohdealueen vuoksi. Merkittävää tutkimuksen tuloksessa on se, että hakukoneyhtiöt ottavat huomioon semanttisen datan WWW-sivuja indeksoidessaan – joka toisaalta on melko itsestään selvää, koska hakukoneyhtiöt ovat yhdessä tuottaneet tietomallin WWW-sivuille lisättävälle metadatalle. Vaikka semanttisen datan tuottamisen merkitys ei tällä hetkellä olisi suurta, hakukoneyhtiöt voivat kuitenkin muuttaa toimintaansa siihen lupaa pyytämättä. Esimerkiksi vuoden 2015 alussa Google muutti algoritmiaan suosimaan mobiililaitteissa toimivia WWW-sivustoja, kun haut tehdään mobiililaitteesta (Makino & Phan, 2015). Tulevaisuudessa semanttisen datan tuottaminen voi parantaa verkkokaupan hakukonenäkyvyyttä jopa huomattavasti, ja tällä hetkellä schema.org-tietomallin mukaista semanttista dataa tuottamalla voi erottua kilpailijoista, jotka eivät sitä vielä tee.

Viiteluettelo

Adida, B., Birbeck, M. & Herman, I. (2011). Semantic Annotation and Retrieval: Web of Hypertext – RDFa and Microformats. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 157–190. Springer.

Anders Innovations (2014). *Verkkokauppaopas 2015*. Tietoyhteiskunnan kehittämiskeskus.

Ashraf, J., Hussain, O. K. & Hussain, F. K. (2014). Empirical analysis of domain ontology usage on the Web: eCommerce domain in focus. *Concurrency and Computation: Practice and Experience*, 26(5), pp 1157–1184.

BBC News. *Amazon seeks to shut down paid review sites*. [online] (2015-10-04) (BBC). Available from: http://www.bbc.com/news/technology-32251698. [Accessed 2015-05-16].

Benjamins, V. R., Radoff, M., Davis, M., Greaves, M., Lockwood, R. & Contreras, J. (2011). Semantic Technology Adoption: A Business Perspective. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 619–657. Springer.

Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E., Hickson, I., Atkins, T., Pieters, S., Weiss, Y., Cáceres, M. & Marquis, M. *HTML 5.1: W3C Working Draft*. [online] (2015) (World Wide Web Consortium). Available from: http://www.w3.org/TR/2015/WD-html51-20150506/. [Accessed 2015-06-02].

Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E., Pfeiffer, S. & Hickson, I. *HTML5: W3C Recommendation*. [online] (2014) (World Wide Web Consortium). Available from: http://www.w3.org/TR/2014/REC-html5-20141028/. [Accessed 2015-06-02].

Berners-Lee, T. *Information Management: A Proposal*. [online] (1989) (World Wide Web Consortium). Available from: http://www.w3.org/History/1989/proposal.html. [Accessed 2015-05-08].

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N. & Weitzner, D. J. (2006). A framework for web science. *Foundations and Trends in Web Science*, 1(1), pp 1–130.

Berners-Lee, T., Hendler, J., Lassila, O. & others (2001). The semantic web. *Scientific American*, 284(5), pp 28–37.

Bing. *Marking Up Your Site with Structured Data*. [online]. Available from: http://www.bing.com/webmaster/help/marking-up-your-site-with-structured-data-3a93e731. [Accessed 2015-06-19].

Boxberg, K. *Myymälät kuihtuvat kun verkkokauppa kasvaa*. [online] (2013-07-19) (Helsingin Sanomat). Saatavissa: http://www.hs.fi/talous/a1374117779238. [Haettu 2015-06-08].

Brickley, D. & Miller, L. *FOAF Vocabulary Specification 0.99*. [online] (2014-01-14). Available from: http://xmlns.com/foaf/spec/20140114.html. [Accessed 2015-06-08].

Çelik, T., Lilley, C., Baron, D., Pemberton, S. & Pettit, B. *CSS Color Module Level 3: W3C Recommendation*. [online] (2011-07-07) (World Wide Web Consortium). Available from: http://www.w3.org/TR/css3-color/. [Accessed 2015-06-12].

CERN. *The history of CERN*. [online] (2012). Available from: http://timeline.web.cern.ch/timelines/The-history-of-CERN. [Accessed 2015-05-08].

Domingue, J., Fensel, D. & Hendler, J. (2011). Introduction to the Semantic Web Technologies. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 1–41. Springer.

eCl@ss. *eCl@ss Version 9.0*. [online] (2014-08-12). Available from: http://www.eclasscontent.com/. [Accessed 2015-06-12].

Eisen, M. *Amazon's $23,698,655.93 book about flies*. [online] (2011-04-22). Available from: http://www.michaeleisen.org/blog/?p=358. [Accessed 2015-05-16].

Elmasri, R. & Navathe, S. B. (2014). *Fundamentals of Database Systems*. Pearson.

ETIM. *About ETIM International*. [online] (2015). Available from: http://www.etim-international.com/about-us. [Accessed 2015-06-10].

Facebook. *The Open Graph protocol*. [online] (2014-10-20). Available from: http://ogp.me/. [Accessed 2015-06-20].

Facebook. *Sharing Best Practices for Websites & Mobile Apps*. [online]. Available from: https://developers.facebook.com/docs/sharing/best-practices. [Accessed 2015-06-20].

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. & Berners-Lee, T. *Hypertext transfer protocol–HTTP/1.1*. [online] (1999) (Internet Engineering Task Force (IETF)). Available from: http://www.ietf.org/rfc/rfc2616.txt. [Accessed 2015-

05-20].

Fishkin, R. & Høgenhaven, T. (2013). *Inbound Marketing and SEO: Insights from the Moz Blog*. John Wiley & Sons.

Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education.

Gandon, F., Krummenacher, R., Han, S.-K. & Toma, I. (2011). Semantic Annotation and Retrieval: RDF. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 117–155. Springer.

Goel, K., Guha, R. V. & Hansson, O. *Introducing rich snippets*. [online] (2009-12-05). Available from: http://googlewebmastercentral.blogspot.de/2009/05/introducing-rich-snippets.html. [Accessed 2015-06-01].

Google. *Rich Snippets*. [online] (2015a-12-02). Available from: https://developers.google.com/structured-data/rich-snippets/. [Accessed 2015-06-19].

Google. *Sitelinks Search Box*. [online] (2015b-07-05). Available from: https://developers.google.com/structured-data/slsb-overview. [Accessed 2015-06-20].

Grimm, S., Abecker, A., Völker, J. & Studer, R. (2011). Ontologies and the Semantic Web. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 507–579. Springer.

Grün, C., Huemer, C., Liegl, P., Mayrhofer, D., Motal, T., Schuster, R., Werthner, H. & Zapletal, M. (2011). eBusiness. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 787–847. Springer.

GS1. *Version 15, GS1 General Specifications*. [online] (2015). Available from: http://www.gs1.org/sites/default/files/docs/barcodes/GS1_General_Specifications.pdf. [Accessed 2015-06-12].

Guenther, R. & Radebaugh, J. (2004). Understanding metadata. *National Information Standard Organization (NISO) Press*.

Harth, A., Janik, M. & Staab, S. (2011). Semantic Web Architecture. In: Domingue, J., Fensel, D., & Hendler, J. (Eds) *Handbook of Semantic Web Technologies*. pp 43–75. Springer.

Havumäki, H. & Jaranka, E. (2014). *Sähköinen kaupankäynti*. Sanoma Pro.

Hepp, M. (2008). GoodRelations: An Ontology for Describing Products and Services Offers on the Web. In: Gangemi, A. & Euzenat, J. (Eds) *Knowledge Engineering: Practice and Patterns*. pp 329–346. Springer.

Hepp, M. *GoodRelations Language Reference V 1.0*. [online] (2011-01-10) (GoodRelations). Available from: http://www.heppnetz.de/ontologies/goodrelations/v1.html. [Accessed 2015-05-14].

Hepp, M. & Radinger, A. *eClassOWL - The Web Ontology for Products and Services*. [online]. Available from: http://www.heppnetz.de/projects/eclassowl/. [Accessed 2015-06-12].

Hickson, I. *HTML Microdata: W3C Working Group Note*. [online] (2013-10-29) (World Wide Web Consortium). Available from: http://www.w3.org/TR/2013/NOTE-microdata-20131029/. [Accessed 2015-06-07].

Lahtinen, T. (2013). *Verkkokaupan käsikirja*. Yrityskirjat.

Magento. *Magento Community Edition User Guide*. [online] (2015). Available from: http://merch.docs.magento.com/ce/user_guide/Magento_Community_Edition_User_Guide.html. [Accessed 2015-06-19].

Makino, T. & Phan, D. *Rolling out the mobile-friendly update*. [online] (2015-04-21). Available from: http://googlewebmastercentral.blogspot.fi/2015/04/rolling-out-mobile-friendly-update.html. [Accessed 2015-06-22].

Mannila, H. & Räihä, K.-J. (1992). *The design of relational databases*. Addison-Wesley.

Mavridis, T. & Symeonidis, A. L. (2015). Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms. *Engineering Applications of Artificial Intelligence*, 41, pp 75–91.

Meusel, R., Petrovski, P. & Bizer, C. (2014). The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., & Goble, C. (Eds) *The Semantic Web – ISWC 2014*. pp 277–292. Springer.

Mäntymaa, E. *Verkkokirjakauppa kasvaa – samaan aikaan kivijalkakaupat kituuttavat*. [online] (2015-03-29) (Yle uutiset). Saatavissa: http://yle.fi/uutiset/verkkokirjakauppa_kasvaa__samaan_aikaan_kivijalkakaupat_kituu

ttavat/7895981. [Haettu 2015-05-15].

Patel-Schneider, P. (2014). Analyzing Schema.org. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., & Goble, C. (Eds) *The Semantic Web – ISWC 2014*. pp 261–276. Springer.

Pingdom. *Internet 2012 in numbers*. [online] (2013-01-16) (Pingdom). Available from: http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/. [Accessed 2015-05-19].

Pohorec, S., Zorman, M. & Kokol, P. (2013). Analysis of approaches to structured data on the web. *Computer Standards & Interfaces*, 36(1), pp 256–262.

Ramanathan, G. *Introducing schema.org: Search engines come together for a richer web*. [online] (2011-02-06). Available from: http://googleblog.blogspot.fi/2011/06/introducing-schemaorg-search-engines.html. [Accessed 2015-05-14].

schema.org. *Schema.org: About*. [online] (2015a-05-13). Available from: http://www.schema.org/docs/about.html. [Accessed 2015-06-10].

schema.org. *Schema.org version 2.0*. [online] (2015b-05-13). Available from: http://www.schema.org/version/2.0/. [Accessed 2015-06-10].

sitemaps.org. *sitemaps.org*. [online] (2008-02-27). Available from: http://www.sitemaps.org/. [Accessed 2015-06-17].

Steinbock, D. (1997). *Verkkobisnes: Internetin kehityskaari, kaupallistuminen ja verkottuminen*. Uniacta.

STK. *ETIM tulossa - Mitä se tarkoittaa?*. [online] (2015-04-29). Available from: http://www.stkliitto.fi/ETIM. [Accessed 2015-06-10].

SVT. *5. Verkkokauppa*. [online] (2014-06-11) (Väestön tieto- ja viestintätekniikan käyttö). Saatavissa: http://www.stat.fi/til/sutivi/2014/sutivi_2014_2014-11-06_kat_005_fi.html. [Haettu 2015-05-15].

TechCrunch. *Yandex joins Google, Yahoo! and Bing to collaborate on Schema.org*. [online] (2011-01-11). Available from: http://techcrunch.com/2011/11/01/yandex-joins-google-yahoo-and-bing-to-collaborate-on-schema-org/. [Accessed 2015-05-14].

Tinnilä, M., Vihervaara, T., Klimscheffskij, J. & Laurila, A. (2008). *Elektroninen liiketoiminta 2.0: Avainkäsitteistä ansaintamalleihin*. Teknologiainfo Teknova.

UN/CEFACT. *Codes for Units of Measure Used in International Trade*. [online] (2015). Available from: http://www.unece.org/fileadmin/DAM/cefact/recommendations/rec20/rec20_Rev11e_2015.xls. [Accessed 2015-06-14].

Veverka, M. *Unplugged: Got cloud?*. [online] (2013-02-18) (USA TODAY). Available from: http://www.usatoday.com/story/tech/columnist/veverka/2013/02/18/rackspace-intel-amazon-gogrid/1924261/. [Accessed 2015-08-06].

Wikström, V. *Minkä kanavan kuluttaja valitsisi: Internet muuttaa ostokäyttäytymistä*. [online] (2012-03-20) (TNS Gallup). Saatavissa: http://www.tutkimusseura.org/tiedostot/200312/tns.pdf. [Haettu 2014-06-08].