Yhteisöllinen semanttinen web 2.0 -seminaari, Aalto-yliopisto, 15.1.2010
Tiedote vapaasti käytettäväksi (31.12.2009)

FinnONTO-hanke loi perustan suomalaiselle semanttiselle web 2.0:lle

Eero Hyvönen
Aalto-yliopisto, mediatekniikan laitos ja
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Semanttisen laskennan tutkimusryhmä SeCo
http://www.seco.tkk.fi/

Perinteinen web koostuu joukosta toisiinsa linkeillä yhdistettyjä verkkosivuja; siitä voidaan käyttää nimitystä Web of Pages. Semanttinen web puolestaan yhdistää verkkosivuihin ja palveluihin liittyvät käsitteet assosiatiiviseksi semanttiseksi käsiteverkoksi, josta käytetään nimitystä Web of Data tai Linked (Open) Data. Avoimen tiedon keruuta koordinoivan Linked Data -yhteisön julkaisema käsiteverkko sisältää jo miljardeja solmuja ja kaaria kymmenistä eri tietokannoista, kuten koko Wikipedian semanttisen version. Työssä käytettävien semanttisen webin standardien kehitystyötä koordinoidaan Sir Tim Berners-Leen johtaman Word Wide Web konsortion toimesta.

Semanttisen webin käsiteverkon ideana on toimia eräänlaisena tietokoneiden ymmärtämänä Wikipediana. Sen avulla verkkopalvelut voivat esimerkiksi tietää, että Suomi on Euroopassa, mikrotietokoneet ovat laitteita tai että alpakka on joko metalliseos tai luonnonkuitu. Tällainen ns. ontologinen tieto mahdollistaa verkkosisältöjen ja -palveluiden sisällöllisen yhteentoimivuuden, yhdistämisen, sekä älykkäiden verkkopalveluiden toteuttamisen. Ontologinen käsiteverkko muodostaa webissä eräänlaisen sisällöllisen tieverkoston, infrastruktuurin, jonka avulla tietosisältöjä voidaan hakea niiden sisällön perusteella, yhdistää toisiinsa automaattisesti, ja jonka avulla tietojärjestelmät voivat kommunikoida keskenään. Suomessa semanttisen webin kansallinen hanke FinnONTO 2003-2007 ja FinnONTO 2.0 2008-2010 on rakentanut perustan suomalaisen semanttisen webin ontologiselle sisältöinfrastruktuurille. Sen sanasto- ja ontologiapalvelut ovat yritysten ja organisaatioiden vapaasti käytettävissä Kansallisena ontologiakirjastopalveluna ONKI. Kansallisen semanttisen webin sisältöinfrastruktuurin mahdollisuuksia on pilotoitu laajoissa kansallisissa portaalihankkeissa kuten Kulttuurisampo.fi ja TerveSuomi.fi.

Tarkastellaan esimerkkinä semanttisen webin mahdollisuuksista kulttuurisisältöjen, kuten museoiden, kirjastojen ja arkistoiden kokoelmien julkaisemista kansallisella tasolla webissä. Asiakkaan kannalta on toivottavaa, että eri organisaatioiden ja kansalaisten tuottamat kulttuurisisällöt löytyisivät keskitetystä hakupalvelusta verkossa eikä ainoastaan eri organisaatioiden omista verkkopalveluista etsien. Suomessa on esimerkiksi n. 1000 eri museota, joista pienimpien ei ehkä edes kannatta perustaa omaa erillistä hakupalvelua. Sekä asiakkaan että tiedon tuottajien avuksi tarvitaan siksi yhteisöllinen, eri organisaatioiden ja kansalaisten julkaisujärjestelmä. FinnONTO-hankkeessa toteutettu prototyyppi “Kulttuurisampo – suomalainen kulttuuri semanttisessa web 2.0:ssa” (http://www.kulttuurisampo.fi/) osoittaa, miten tällainen kansallisen tasonjärjestelmä voidaan luoda semanttisen webin ja Web 2.0:n perustalle.

Haasteina semanttinen yhteentoimivuus ja sisällöntuotanto

Kulttuurisammon kaltaisen kansallisen tason julkaisujärjestelmän toteuttamisessa on kaksi periaatteellista haastetta: semanttinen ja organisatorinen. Semanttisena haasteena on sisältöjen yhteentoimivuus (interoperability) kuvan 1 mukaisesti. Kulttuurisisällöt ovat hyvin monimuotoisia sekä sisällöltään että tallennusmuodoiltaan ja linkittyvät toisiinsa monilla eri

tavoilla. Esimerkiksi Akseli Gallen-Kallelan elämänkerta (Kansallisbiografiassa) liittyy semanttisesti hänen teoksiinsa (Ateneumin ym. taidemuseoissa), Kalevalaan (Suomalaisen Kirjallisuuden Seura), paikkoihin ja rakennuksiin luonnossa (esim. Tarvaspää Espoo kaupunginmuseon GIS-järjestelmässä), hänen kollegoittensa elämänkertoihin, historiallisiin tapahtumiin (Agricola – Suomen historiaverkossa), videoihin (Yleisradion Elävässä arkistossa) jne. Kun suomalaisiin kokoelmiin kuuluu miljoonia eri kohteita, ei aineistojen muokkausta yhteismitalliseksi ja sisällöllistä linkitystä toisiinsa voida tehdä käsin, vaan sen on oltava automaattista.

Kuva 1. Kulttuurisampo yhdistää monimuotoisia kulttuuriaineistoja sisällöllisesti toisiinsa.

Palvelun organisatorisena haasteena on alan toimijoiden riippumattomuus ja toisistaan poikkeavat käytännöt sisällöntuotannossa kuvan 2 mukaisesti. Aineistoja tuottavat mm. erilaiset muistiorganisaatiot kokoelmatietokantoihinsa, kansalaiset valokuvaamalla kohteita, Wikipedian ja Linked Open Data -hankkeiden kaltaiset Web 2.0 -yhteisöt, karttapalveluiden, ilmavalokuvien tuottajat, Maanmittauslaitos jne. Ilman yhteisiä pelisääntöjä sisällönkuvailussa tietojen yhdistäminen ei ole suoraviivaista ja siinä tarvitaan kustannuksiltaan kallista käsityötä.

Kuva 2. Heterogeenisia kulttuurisisältöjä tuottavat toisistaan riippumattomat organisaatiot ja kansalaiset.

Organisaatiolähtöisyyden ja loppukäyttäjän näkökulman yhdistäminen

Semanttisen webin ja Web 2.0:n ideoiden avulla molempia ongelmia voidaan lähestyä samanaikaisesti uudella tavalla kuvan 3 mukaisesti. Järjestelmän ytimessä on kansallisista käsitteistä muodostuvat ontologiat eli sanastot, jotka muodostavat Web of Data -konseptin

mukaisen semanttisen webin pysyvän ydinverkon. Esimerkiksi Kulttuurisampo.fi-palvelussa käytetyt ontologiat muodostavat satojen tuhansien käsitteiden verkoston, josta mm. käy ilmi että akvarellit ovat eräänlaisia maalauksia, jotka puolestaan ovat taideteoksia, että Otaniemi on osa Espoota, joka on kunta Suomessa, ja että Akseli Gallen-Kallela oli tiettyyn aikaan elänyt suomalainen taiteilija, joka oli Hugo Simbergin oppilas, tunsi C. G. Mannerheimin jne. Eri sisällöntuottajien tuottama kuvailutieto (metatieto) kokoelmistaan ja muusta tiedosta saadaan yhteismitalliseksi sopimalla yhteisesti metatietoformaateista ja peilaamalla metatietojen arvot yhteiselle kansalliselle FinnONTO-ontologiaverkostolle. Yhteentoimivuuden kannalta keskeistä on, että semanttisen webin ontologiat määritellään tavalla, jota tietokoneet osaavat tulkita (”ymmärtää”) yksinkertaisten loogisten periaatteiden mukaisesti. Nämä on määritelty webin kehitystä kansainvälisesti koordinoivan W3C-järjestön standardisuosituksina (RDF(S), OWL, SPARQL, SWRL).

FinnONTO-infrastruktuurin varaan voidaan kehittää aiempaa älykkäämpiä haku-, linkitys-, suosittelu- ym. palveluita. Esimerkiksi hakusanalla “Espoo” löytyvät “Otaniemessä” olevat kohteet, “taideteoksia” etsittäessä eivät akvarellit putoa tuloksista pois, ”Gallen-Kallelan” töitä haettaessa voidaan suositella katsottavaksi myös hänen opettajansa maalauksia jne. Järjestelmän toinen etu on, että eri tahojen tuottamat tiedon muruset yhdistyvät ontologioiden kautta laajaksi semanttiseksi verkoksi, jolloin jokaisen tiedon tuottajan omaa sisältöä voidaan rikastaa siihen liittyvien muiden sisältöjen kautta. Syntyy aito win-win-tilanne, jossa voittavat sekä kaikki sisällöntuottajat että loppukäyttäjät.

Kuva 3. Kansallisen ontologiainfrastruktuurin hyödyntäminen Kulttuurisampo-järjestelmässä.

Esimerkiksi Kulttuurisampo yhdistelee sisällöllisesti toisiinsa mm. Suomen kansallismuseossa olevia esineitä, Valtion taidemuseon maalauksia, Yleisradion Elävän arkiston ja Opinportin videoita, Helsingin kaupunginkirjaston romaaneja ja novelleja, Maamittauslaitoksen vanhoja karttoja sekä Suomalaisen Kirjallisuuden Seuran kansanrunoja ja -sävelmiä, Kalevalan runoja jne. Järjestelmässä on tällä hetkellä 134 000 kohdetta, jotka edustavat 67 eri sisältötyyppiä 22 kotimaisesta organisaatiosta. Lisäksi palveluun on liitetty kuuden ulkomaisen lähteen tietokantoja, kuten Wikipedia sekä Googlen yhteisöllinen Panoramio-kuvapalvelu. Näiden kautta ja kokoelmakohteita kommentoimalla tarjoutuu myös kansalaisille mahdollisuus toimia Kulttuurisammon sisällöntuottajina Web 2.0 -hengessä. Sisällöntuotannon rajapintoja ollaan avaamassa eri tahoille Web 2.0 -hengessä. Esimerkiksi kirjastoalalla on menossa kansallinen

hanke Kirjasampo, jossa kaikki suomalainen kaunokirjallisuus (yli 55 000 teosta) sisällönkuvaillaan Kulttuurisampoon.

Kulttuurisampo- ja ONKI-järjestelmässä on kehitetty semanttisen julkaisujärjestelmän malli, innovaatio, jossa hajautetusti ja organisaatiolähtöisesti tuotettu tieto voidaan yhdistää ja tarjota loppukäyttäjille näiden omista tiedonhakutarpeista lähtien. Mallia voidaan soveltaa kulttuurin ohella monilla muillakin aloilla FinnONTO-infrastruktuurin ja siihen sisältyvien ontologiapalveluiden kautta.

Kansallinen ontologiapalvelu ONKI käytettävissänne

Kulttuurisampo perustuu FinnONTO-hankkeessa valmistuneeseen prototyyppiin kansallisesta semanttisen webin sisältöinfrastruktuurista. Sen ytimenä on joukko laajoja kansallisia semanttisen webin ontologioita, kuten Kansalliskirjaston YSA-asiasanastoon perustuva Yleinen suomalainen ontologia YSO (yli 20 000 käsitettä), museoalan ontologia MAO (n. 6700 käsitettä), Agriforest-sanastoon perustuva Maa- ja metsätalousalan ontologia AFO (n. 6000 käsitettä), taideteollisuusalan TAO (n. 2600) ja valokuvausalan ontologia VALO (n. 1900 käsitettä). Uusia ontologioita ollaan liittämässä järjestelmään, esimerkiki Merenkulkualan ontologia MERO ja Kaunokirjallisuuden ontologia KAUNO. Maanmittauslaitoksen paikannimirekisteristä on muodostettu Suomen paikkaontologia SUO (800 000 paikkaa ja käsitettä Suomesta ja yli 4 miljoonaa ulkomailta). Henkilöitä ja organisaatioita varten kehitetään toimijaontologiaa TOIMO, jonka ytimenä on Getty-säätiön kansainvälinen n. 120 000 kulttuurihenkilöä sisältävä ULAN rekisteri. Peilaamalla eri ontologioiden käsitteet toisilleen syntyy tuloksena kaiken kattava KOKO-ontologia, johon Kulttuurisampo.fi-palvelu perustuu.

Käytännön sisällönkuvailutyötä varten ontologiat ja sanastot on julkaistu Kansallisena ontologiapalveluna ONKI (http://www.yso.fi/). Sen palvelimien avulla kansalliset ja kansainväliset sanastot ja ontologiat voidaan ottaa kustannustehokkaasti käyttöön yritysten ja julkisten laitosten omissa tietojärjestelmissä AJAX-leijukkeina (widget) tai Web Service -rajapintojen kautta. ONKI-palvelussa on julkaistu lähes 70 eri ontologiaa ja sanastoa kolmen eri palvelinratkaisun avulla. Esimerkiksi paikkaontologiaa voi käyttää ONKI Geo -palvelun kautta, joka puolestaan hyödyntää Google Maps -palvelua.

Yhteisten käytänteiden tukemiseksi ontologiat ja sanastot julkaistaan lähtökohtaisesti open source periaatteella (Creative Commons Attribution -lisenssi) ja ONKI-palveluiden käyttö on maksutonta. Oma ONKI -palvelun kautta asiakkaat voivat myös ladata palveluun omia sanastojaan ja ontologioitaan ja julkaista niitä toiminnallisina ONKI-palveluina.

Järjestelmä on Living Laboratory -käytössä useissa eri organisaatioissa yliopistomaailman ulkopuolella, myös ulkomailla. Esimerkiksi maamme käytetyin sanastopalvelu, Kansalliskirjaston VESA:sta (n. 12 miljoonaa hakua vuodessa) on rinnakkaiskäytössä ja uusi VESA ONKI, joka suunnitelman mukaan korvaa jatkossa nykyisen VESA-palvelun. Sydäntautiliiton ja UKK-instituutin sisällönhallintajärjestelmät on kytketty ONKI palvelimiin liittyen TerveSuomi.fi portaalin aineistojen tuotantoon, Kirjastot.fi-portaalin Kysy kirjastonhoitajalta -palvelu. Julkisella sektorilla käytössä olevan IPSV-Integrated Public Sector Vocabulary -sanaston ONKI-versiota koekäytetään Englannissa ja yhdysvaltalaisen Getty-säätiön AAT- ja TGN-sanastoja Italiassa, Maltalla ja Eestissä ONKI-palvelun kautta.

Semanttinen TerveSuomi.fi-järjestelmä tuotantoon

FinnONTO-hankkeen toinen laaja sovelluskohde on Terveyden ja hyvinvoinnin laitoksen (THL) koordinoima kansallinen TerveSuomi.fi-hanke, jossa maamme eri terveysorganisaatiossa tuotettu terveyden edistämiseen liittyvä tieto saatetaan kansalaisten käytettäväksi semanttisen portaalin avulla. FinnONTO-hankkeessa lehitetty TerveSuomi-palvelun prototyyppi otettiin koekäyttöön julkisessa verkossa syksyllä 2008 Teknillisessä korkeakoulussa ja sille myönnettiin pari kuukautta myöhemmin kansainvälinen Semantic Web Challenge Award -teknologiapalkinto

semanttisen webin kansainvälisen tutkijayhteisön toimesta. Protyypistä tuotteistettu tuotantoversio (http://www.tervesuomi.fi) julkistettiin THL:n toimesta 13.5. TERVE-SOS-tapahtumassa Helsingissä. Kulttuurisampoa ja TerveSuomea vastaava prototyyppi semanttisesta portaalista on kehitteillä myös liiketoiminta-alalle liittyen Teollisuus- ja elinkeinoministeriön ylläpitämän YritysSuomi.fi-portaalin sisältöihin ja palveluihin. Jo aiemmin v. 2004 julkaistu ja TerveSuomen tavoin Semantic Web Challenge Award -palkinnon saanut FinnONTO-sovellus MuseoSuomi.fi on edelleen aktiivisessa käytössä verkossa (http://www.museosuomi.fi/).

Kustannusten säästö keskitetyillä avoimilla ratkaisuilla

Semanttisen webin ehkä keskeisin haaste on kustannustehokas metatiedon ja ontologioiden tuotanto. FinnONTO-hankkeen ideana on luoda maahamme avoin ontologiainfrastruktuuri, jolloin uusia sovelluksia voidaan kehittää kustannustehokkaasti hyödyntämälle aiemmin kehitettyjä ontologioita. ONKI-palveluiden kautta ontologioiden käyttöön liittyviä toiminnallisuuksia voidaan toteuttaa keskitetysti ja hyödyntää kustannustehokkaasti asiakasorganisaatioissa hieman saamaan tapaa kuin esimerkiksi Google Maps -palvelua. Tutkimushankkeen tuloksena on syntynyt joukko ontologioihin, web-tekniikoihin ja kieliteknologiaan perustuvia työkaluja sisältöjen ja palveluiden (puoli)automaattista semanttista kuvailua varten, kuten metadataeditori SAHA, tiedoneristin POKA ja sisällön tarkistin VERA. Näitä on sovellettu Kulttuurisammon ja TerveSuomen ohella esimerkiksi SanomaData Oy:n lehtiaineistoille, kirjastoalan Kirjastot.fi-palvelun help-desk-palvelussa ”Kysy kirjastonhoitajalta” sekä Sininen meteoriitti Oy:n ohjelmistotuotteissa.

FinnONTO 2.0 -hankkeen yrityskonsortioon kuuluu kaudella 2008-2010 38 yritystä ja julkista organisaatiota, mikä tekee siitä tiettävästi Tekesin historian laajimman hankkeen yrityskonsortion koolla mitattuna. Pääosa kehitystyöstä on tehty Teknillisen korkeakoulun mediatekniikan ja Helsingin yliopiston tietojenkäsittelytieteen laitoksen Semanttisen laskennan tutkimusryhmässä SeCo. Tutkimuspartnereina ovat olleet myös TKK:n geoinformaatio- ja paikannustekniikan laitos, Helsingin yliopiston yleisen kielitieteen laitos sekä Tampereen yliopiston informaatiotutkimuksen laitos. FinnONTO-hanke on kerännyt ympärilleen laajan kotimaisen yhteistyöverkoston ja järjestänyt vuosittain semanttisen webin konferensseja sekä kotimaassa ja ulkomailla, mm. Pariisin Louvressa ja Korean Pusassa. Hanke on verkottunut ulkomaille tuloksena mm. tutkijavaihtoa ja kansainvälisiä EU-projekteja. Teknisten ideoiden ohella myös FinnONTO-konseptia kokonaisuutena on otettu käyttöön ulkomailla, mm. Eestissä käynnistyvässä EstONTO-hankkeessa. FinnONTO palkittiin Tekesin Fenix-ohjelman ”parhaana projektina”.

Lisätietoja, verkkopalveluita, kirjallisuutta, aineistoja

Kulttuurisampo	http://www.seco.tkk.fi/applications/kulttuurisampo/
TerveSuomi	http://www.seco.tkk.fi/applications/tervesuomi/
ONKI-palvelu	http://www.seco.tkk.fi/services/onki/
FinnONTO-hanke	http://www.seco.tkk.fi/projects/finnonto/

Kirjoittaja on professori Aalto-yliopiston mediatekniikan laitoksella ja tutkimusjohtaja Helsingin yliopiston tietojenkäsittelytieteen laitoksella. Hän johtaa FinnONTO-hanketta ja sen tutkimustyöstä päävastuuta kantavaa Semanttisen laskennan tutkimusryhmää.