Aalto University

UNIVERSITY OF HELSINKI

19.12. 2011

Linked Data Finland:
Metatiedon tuotanto, yhdistäminen ja sovellukset

Hankekuvaus

Eero Hyvönen

Aalto-yliopisto
Mediatekniikan laitos
PL 15500, 00076 AALTO
puh. 050 3841618, fax. 451 3356
eero.hyvonen@tkk.fi
http://www.seco.tkk.fi/

SeCo
semantic computing

SEMANTTISEN LASKENNAN
TUTKIMUSRYHMÄ

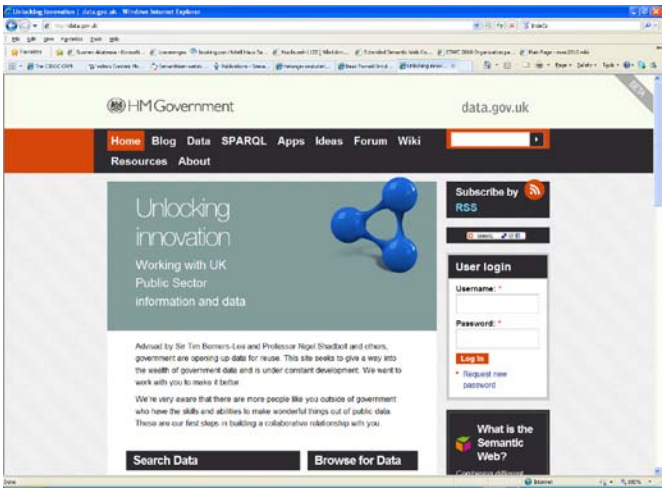
SISÄLLYSLUETTELO

TAVOITE

Linked Data Finland (LDF) on monialainen kansallinen hanke, jonka tavoitteena on käynnistää ja edistää tietovarantojen hyödyntämistä maassamme uusimpien semanttisten web-teknologioiden avulla. Hankkeen lähtökohtana on kansainvälinen Linked Data -ajattelu ja –liike (LD) sekä siihen liittyvät teknologiat, standardit ja parhaat käytännöt. Nämä mahdollistavat uudella tavalla semanttisen tietosisällön automaattisen tuotannon, sisältöjen yhdistämisen (linked data) toisiin tietosisältöihin yhteentoimivasti (semantic interopability), julkaisemisen käyttöön otettavina verkkopalveluina (mashup, web service) sekä uudet, aiempaa älykkäämmät ja kontekstiherkät sovellukset. Teknologiaa ja innovaatioita kehitetään yritysmaailman pilottisovellusten yhteydessä, joiden kautta voidaan arvioida uuden teknologian hyödyllisyys ja mahdollisuudet. Työvälineet ja aineistot julkaistaan Living Laboratory -tyyppisinä avoimina verkkopalveluina ja hankkeen tulokset avoimesti hyödynnettävissä olevassa muodossa (open source, open data) mahdollisuuksien mukaan.

Tavoitetutkimus fokusoituu LD-alueen keskeisimpiin haasteisiin: metadatan automaattiseen tuotantoon, automaattisesti tuotetun tiedon laatuongelmiin, ja heterogeenisten metadatajoukkojen siltaamismenetelmiin. Kehitettävän ydinteknologian mahdollisuuksia ja haasteita arvioidaan liiketoimintalähtöisten case-demonstraattoreiden avulla.

TAUSTA



Kuva 1. Britannian data.gov.uk -palvelu julkaisee valtakunnan tietoaineistoja LD-periaatteella.

Kansainvälinen Linked Data -liike

Yhdistetyn ja avoimen tiedon määrä webissä on viimeisen parin vuoden aikana kasvanut räjähdysmäisesti erityisesti kansainvälisen Linked Data -liikkeen (LD) seurauksena[1] (Bizer et al., 2009; Heath, Bizer,

[1] http://linkeddata.org/

2011), jonka keskeisenä keulahahmona toimii webin ”isä” ja W3C:n johtaja Tim Berners-Lee. LD:n idea perustuu joukkoon semanttisia teknologioita ja käytäntöjä (best practices), joiden avulla voidaan julkaista laajoja tietosisältöjä siten, että ne yhdistyvät toisiinsa käsitteellisellä tasolla. Näin muodostettu datapilvi (Data Cloud, Web of Data), kuten avoin Linked Open Data -pilvi (LOD), on kustannustehokkaasti hyödynnettävissä sekä hyvin määritellyn standardoidun RDF-perustaisen (Resource Description Framework) rakenteensa että rajapintojensa (SPARQL, REST ym.) kautta[2] . Uusi lähestymistapa on otettu käyttöön mm. Britannian koko julkisen sektorin tietoaineistojen julkaisemisessa data.gov.uk -palvelun kautta (kuva 1), joka sisältää sekä avoimesti julkaistuja tietoaineistoja että niiden päälle kehitettyjä lukuisia sovelluksia. *Linked Data -liikkeen ympärille on syntynyt uudentyyppinen, sovelluksia kehittävien yritysten ja avointa tietoa tarjoavan julkisen sektorin ekosysteemi.*

Kansainvälinen LOD-pilvi on syntynyt erittäin nopeasti. Se koostuu tätä hakemusta jätettäessä 295 tietojoukosta. Vuonna 2010 pilveen kuului 203 tietoaineistosta kuvan 2 mukaisesti eri aihealueisiin jaoteltuna. Vuonna 2007 tietojoukkoja oli 28, vuonna 2008 45 ja vuonna 2009 95, joten kasvu on ollut eksponentiaalisen nopeaa (n. 100% vuodessa). Lisäksi kaavion LOD-pilvi edustaa vain keskeisimpiä datajoukkoja laajemmasta kansainvälisestä CKAN-rekisteristä[3], johon oli rekisteröity yli 1600 tietojoukkoa (v. 2010).

LOD-pilvi koostuu solmuista, jotka ovat datajoukkoja (dataset), kuten miljoonia paikkoja sisältävä Geonames, erikielisistä Wikipedioista automaattisesti louhittu RDF-muunnos DBPedia[4], Ruotsin kansalliskirjaston viitetietokanta Libris tai EU:n virallisten tilastojoukko Eurostat. Kaaret siltaavat eri tietojoukkojen käsitteitä toisilleen kertoen esimerkiksi sen, miten englanninkielisen Wikipedian artikkeli ”Helsinki” liittyy Geonames-paikka-aineiston metatietoihin Helsingistä paikkana. Datajoukot muodostavat jättiläismäisen, miljardeista soluista ja kaarista muodostuvat semanttisen RDF-verkon, josta on muodostunut maailmanlaajuisen semanttisen webin ydin.

Linked Data -aineistojen julkaisemisen ideana on, että niiden varaan voidaan kustannustehokkaasti kehittää käytännön sovelluksia, kun tiedon saatavuuteen, yhteentoimivuuteen ja yhdistämiseen liittyvät ongelmat on jo osin tiedon julkaisijayhteisön voimin ratkaistu, ks. esim. W3C:n use case -lista[5]. Tunnettu alan sovellus on esimerkiksi BBC:n kotisivut, jotka perustuvat DBPedian käyttöön organisaation tuottamien tietosisältöjen yhdistämisessä ja sisällön rikastamisessa.

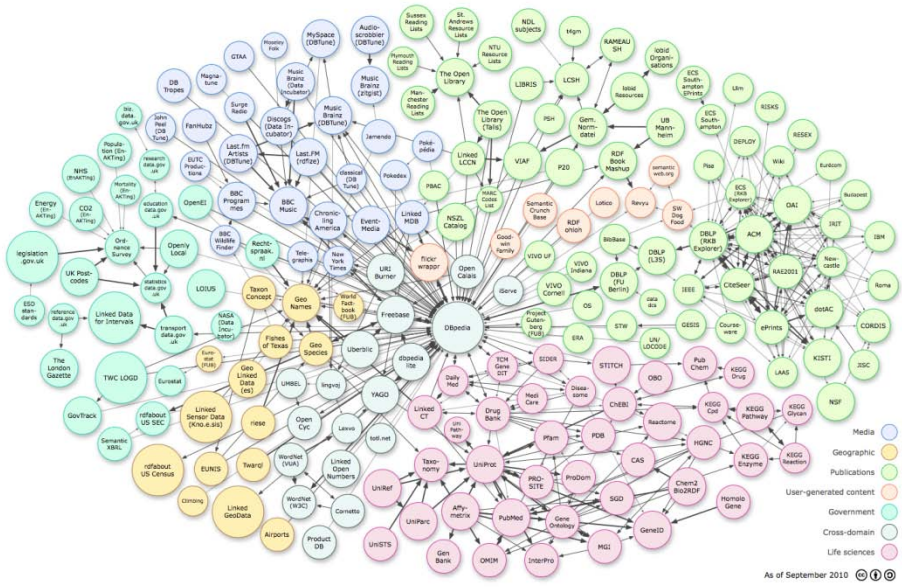
---

22 http://www.w3.org/2001/sw/

3 http://ckan.net/

4 http://dbpedia.org/

5 http://www.w3.org/2001/sw/sweo/public/UseCases/



Kuva 2. Linked Open Data -pilven ydin v. 2010 koostui 203 laajasta, toisiinsa sillatusta datajoukosta. Kokonaisuuden ytimenä on erikielisistä Wikipedioista louhittu semanttinen verkko DBPedia.

Yleisten ja avoimien WWW:n laajuisten datajoukkojen ohella LD-teknologia soveltuu myös ei-avointen ja intranet-kohtaisten järjestelmien kehittämiseen, kuten esimerkiksi BBC:n esimerkki osoittaa (Kobilarov et al., 2009). Keskeisinä tuotannollisina ja liiketoiminnallisina etuina ovat 1) mahdollisuus heterogeenisten aineistojen yhteentoimivuuden luomiseen, 2) semanttisten aineistojen sisällöllinen uudelleenkäyttö uusissa sovelluksissa sekä 3) älykkäiden verkkopalveluiden kehittäminen datajoukkojen sisältöjen semantiikkaa ja päättelyä hyödyntäen.

Suomessa Linked Data -filosofiaa ja teknologiaa on kehitetty mm. Kulttuurisampo.fi-palvelussa (Hyvönen et al., 2009). Ensimmäisistä avoimista datajoukoista kuten pääkaupunkiseudun kirjastojen Helmet-aineistosta on tuotettu Linked Data -muotoista dataa. Nyt pitäisi panostaa aineistojen systemaattiseen ja koordinoituun, yhteentoimivaan tuotantoon, kuten Britanniassa sekä USA:n ja Australian hieman vastaavissa avoimen datan hankkeissa, sekä tulosten hyödyntämiseen verkkopalveluina. Pelkkä tiedon avaaminen yhdistettynä tietona ei riitä vaan tarvitaan näyttöjä teknologian tarjoamista uusista mahdollisuuksista.

LDF–tutkimushankkeen tavoitteena on käynnistää ja edistää edistää yhdistetyn tiedon – niin avoimen kuin ei-avoimenkin – yhteentoimivaa tuotantoa ja hyödyntämistä laajassa kansallisessa projektissa, johon kuuluu sekä julkisen että yksityisen sektorin toimijoita. Hanke on hyvin verkottunut sekä kansainvälisen LD-tutkimus- ja kehitysyhteisön että kotimaisen suunnitteilla olevan Tivit Oy:n Data to Intelligence -ohjelman (Paajanen, Kuosmanen, 2010) toimijoiden kanssa.

Kansallisia tavoitteita Suomessa

Yhdistettyyn ja avoimeen tietoon liittyvät kysymykset ovat viimeisen vuoden aikana olleet voimakkaasti esillä kansallisella tasolla:

- Valtiovarainministeriön ja ValtIT:n Valtiotason tietoarkkitehtuurit -hankkeen loppuraportti ja osaraportit valmistuivat[6]. Tietoon liittyvät semanttisen yhteentoimivuuden kysymykset ja ontologiatekniikat ovat niissä voimakkaasti esillä. Tuloksia ollaan kokoamassa toimenpidesuosituksiksi. Valtiovarainministeriö on tulossa mukaan LDF-hankkeeseen.
- Valtiovarainministeriön valmistelema laki julkisen hallinnon tietohallinnon ohjauksesta (tietohallintolaki) astui voimaan 1.9.2011. Sen tavoitteena on tietojärjestelmien yhteentoimivuuden parantaminen valtion asetusvaltaa lisäämällä. Yhä tärkeämpään rooliin tulee VM:n alaisuudessa tapahtuva tietoarkkitehtuurityö ja yhteentoimivuuden koordinointi.
- Liikenne- ja viestintäministeriön julkisen tiedon saatavuus työryhmä[7] on laatinut avoimen tiedon selvityksiä ja oppaita ja valmistumassa on tähän liittyvä valtioneuvoston periaatepäätös[8] ”julkisen sektorin tietoaineistojen saatavuuden parantamisesta ja uudelleenkäytön edistämisestä”. Liikenne- ja viestintäministriö on tulossa mukaan LDF-hankkeeseen.
- OKM:n rahoittamassa laajassa Kansallinen digitaalinen kirjasto -hankkeessa tieto ja metatietointegraatio ovat ydinkysymyksiä, samoin siihen liittyvässä Euroopan laajuisessa Europeana-hankkeessa. KDK-hankkeen tiedon julkaisu- ja asiakasliittymäosaa vetävä Kansalliskirjasto on tulossa mukaan LDF-hankkeeseen.
- OKM:n tutkimuksen tietoaineistot –selvityshankkeen opas ”Tutkimuksen tietoaineistot - olennaisen käsikirja päättäjille” (CSC, 2011) ilmestyi ja loppuraportti on juuri valmistunut[9]. Hankkeen käytännön toteutusta vetävä opetusministeriön CSC Tieteellinen laskenta Oy on tulossa mukaan LDF-hankkeeseen.
- Maamme kansallisen IT-klusterin Tivit:n keskeisin kehityskohde jatkossa on suunnitelmien mukaan tieto ja sen hyödyntäminen Data to Intelligence -ohjelmassa (D2I) (Paajanen, Kuosmanen, 2010). LDF-hanke on neuvottelemassa toim. joht. Reijo Paajasen kanssa yhteistyöstä sekä D2I-hankkeen että jo käynnissä olevan Next media -hankkeen kanssa.
- Laajassa FinnONTO-hankkeessa on kehitetty ja pilotoitu kansallista semanttisen webin ontologiainfrastuktuuria, jonka viemisestä Living Laboratory -ympäristöstä tuotantokäyttöön neuvotellaan. LDF-hanke jatkaa FinnONTO:n päättyessä siinä tehtyä työtä uusiin suuntiin, joita alla tarkemmin esitellään.

LDF-hankkeen tavoitteena on pilotoida maahamme uutta teknologista osaamista ja järjestelmiä, joilla edellä lueteltuja tietolähtöisiä pyrkimyksiä voidaan käytännön tietojärjestelmätasolla edistää ja implementoida. Kansainvälisen LOD-työn tavoin tietojoukkojen pääpaino on avoimilla, erityisesti julkisen sek-

[6] http://www.vm.fi/yhteentoimivuus

[7] http://www.lvm.fi/web/fi/tyoryhmat/tyoryhma/view/1130660

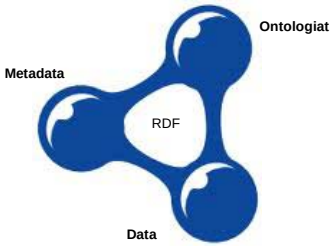
[8] http://www.lvm.fi/c/document_library/get_file?folderId=1591058&name=DLFE-11755.pdf

[9] http://www.csc.fi/csc/julkaisut/oppaat/2010/tutkimuksen-tietoaineistot

torin tai yhteisöllisten portaalien tuottamissa aineistoissa. Liiketoiminnallinen hyödyntäminen syntyy näiden aineistojen päälle kehitettävistä sovelluksista, verkkopalveluista ja uusista innovaatioista. Tieto-yhteiskuntaan on syntymässä uudentyyppinen julkisen sektorin ja yritysten ekosysteemi.

Teknologinen perusta

Yhdistetty tieto koostuu kolmesta pääkomponentista kuvan 3 mukaan. Kaiken perustana on tieto (data), vaikkapa Helsingin Sanomien artikkeli talouskriisistä v. 2009, Flickr-palvelun valokuva Egyptin mellakoista tai YLE:n Elävän arkiston video. Metadata on tietoa datasta, kuten että videon on tuottanut Suomi-Filmi Oy ja se on käytettävissä tietyllä Creative Commons -lisenssillä. Webin metadata voi myös olla tietoa reaalimaailmasta, kuten että Helsinki on Suomen pääkaupunki, jossa asuu 589 000 asukasta. Jotta metatieto olisi koneellisesti tulkittavissa ja yhdistettävissä muihin metatietoihin, pitää sen kuvailujen perustua yhteiselle semanttiselle perustalle ja sanastoille eli ontologioille.



Kuva 3. Yhdistetyn tiedon kolme komponenttia.

LDF-hanke rakentuu vuosina 2003–2011 Suomalaiset semanttisen webin ontologiat (FinnONTO) -hankejatkumon[10] rakentamalle kansalliselle ontologiainfrastruktuurille, johon kuuluu kymmeniä kansallisia ja kansainvälisiä ontologioita julkaistuina ONKI-ontologiapalveluna (Viljanen et al., 2009, Tuominen et al., 2009). ONKI-palvelua käytetään ihmisten toimesta n. 14 000 uniikista domain-osoitteesta kuukaudessa, ja koneellista käyttöä varten palveluun on rekisteröitynyt yli 400 yritystä, julkista organisaatiota sekä yksityistä tutkijaa. Ontologioita vähemmälle huomiolle hankkeessa on jäänyt yhdistetyn tiedon kaksi muuta komponenttia, ts. metatietoon ja tietoon liittyvät kysymykset. Nämä muodostavat LDF-hankkeen tutkimus- ja kehitystyön ytimen. Yhdistettynä FinnONTO:n ontologiatyön tuloksiin voidaan LDF-hankkeessa kattaa koko yhdistetyn tiedon tuotannon, julkaisemisen ja hyödyntämisen kenttä. Teknologisena perustana tässä toimivat W3C:n ja LD-yhteisön nykyiset ja kehitteillä suositukset ja parhaat käytännöt (RDF(S), OWL, RIF, GRDDL, R2RML, POWDER ym.) sillattuna semanttisen laskennan (semantic computing) uusimpiin tutkimustuloksiin (Sheu el al., 2010).

Tutkimusryhmä

Hanketta vetää Aalto-yliopiston mediatekniikan laitoksen ja Helsingin yliopiston tietojenkäsittelytieteen laitoksen Semanttisen laskennan tutkimusryhmä SeCo. SeCo on johtamiensa FinnONTO-hankkeiden ja muiden projektien kautta saavuttanut arvostetun aseman niin kansainvälisesti kuin kotimaassa. Esimerkiksi vuonna 2010 ryhmän EU-hankkeessa kehittämä SmartMuseum-sovellus voitti YK:n UNIDA:n World

[10] http://www.seco.tkk.fi/projects/finnonto/

Summit Award Mobile 2010 -palkinnon yli 400:n kansainvälisen kilpailutyön joukossa, ja kotimaassa yhdistetyn avoimen tiedon julkaisuun pilotoitu avoimen tiedon tietoaineistojen DataSuomi-sovellus voitti yrityssarjan Apps4Finland – Doing Good with Open Data –kilpailussa. Aiemmista palkinnoista mainittakoon kansainvälisen tutkijayhteisön myöntämä Semantic Web Challence Award -palkinto kahdesti (2004, 2008), sekä pääministerin kunniamaininta (2004). Ryhmässä valmistui v. 2010 neljä väitöskirjaa ja parikymmentä kansainvälistä julkaisua[11] (suomalaisten artikkelien lisäksi).

TUTKIMUSALUEET JA -KYSYMYKSET

LD:n hyödyntämisen keskeiseksi haaste- ja tutkimusalueiksi ovat muodostuneet:

1. Metadatan kustannustehokas tuotanto. Tässä tarvitaan yhä automaattisempia menetelmiä.
2. Metadatan laadun ja luotettavuuden parantaminen. Erityisesti automaattisesti tuotetussa metadatassa on tyypillisesti paljon laatuongelmia.
3. Metadatajoukkojen yhdistäminen. LD pilvi koostuu olennaisesti joukosta eri metadataskeemoilla kuvattuja tietoja ja näiden välisiä siltauksia. Haasteena on toisaalta se, miten siltauksia eri tilanteissa voidaan kuvata, toisaalta se, miten spesifikaation mukaisia siltauksia voidaan mahdollisimman laadukkaasti ja automaattisesti tuottaa.
4. Demonstraatiot hyötykäytöstä. Yhdistettyä avointa tietoa ei tuoteta vain poliittisista syistä. Alalle tehtyjä investointeja vastaan tarvitaan vakuuttavia näyttöjä LD-lähestymistavan uusista mahdollisuuksista ja hyödyistä käytännön sovelluksissa.

Metadatan kustannustehokas tuotanto

Tutkimusalueella 1 keskitytään seuraaviin tutkimuskysymyksiin:

Miten monikielisistä teksteistä voidaan tuottaa automaattisesti LD-aineistoja?

Kun käytettävissä on runsaasti dataa, esimerkiksi lehtiartikkeleita tai uutisia, ei aineistojen täsmällisen sisällönkuvailun ja muun metadata tuottaminen käsin ole kustannustehokasta tai edes käytännössä mahdollista. Tarvitaan automaattisen tiedoneristyksen ja annotoinnin menetelmiä. LDF-hankkeessa näitä kehitetään erityisesti 1) ei-rakenteisille tekstiaineistoille kieliteknologian ja tiedon louhinnan menetelmin sekä 2) rakenteiselle tiedoille tilastollisen päättelyn ja oppimisen kautta.

Suurin osa verkkoaineistoista on rakenteetonta tekstiä, jossa on mukana lähinnä tiedon esittämiseen liittyvää HTML:n mukaista rakennetta. Tiedon eristys -tutkimuksessa (information extraction) kehitetään menetelmiä, joilla voidaan tunnistaa tekstistä entiteettejä kuten henkilöitä ja paikkoja (named entity recognition, NER), näiden välisiä suhteita (relation extraction) ja tapahtumia (topic detection).

SeCo-ryhmässä on testattu ja kehitetty semanttisen tiedon kieliteknologiaan perustuvia tiedoneristimiä ja sovellettu niitä mm. sanomalehtiaineistoihin (Frosterus, Hyvönen, 2009; Ahonen, Hyvönen, 2009) ja

[11] http://www.seco.tkk.fi/publications/

THL:n Sosiaaliportti-portaaliin (Sinkkilä et al., 2010). Näiden pohjalta LDF-hankkeessa kehitetään suomalaisille aineistoilla ja ontologioille ARPA-verkkopalvelu, joka on kytkettävässä ulkoisiin sisällöntuotannon järjestelmiin ONKI-palvelun tapaan.

Miten voidaan automatisoida manuaalisesti tuotettavan ja tarkistettavan metatiedon tuotantoa?

Vaikka käytettävissä olisi rakenteistakin tietoa, joudutaan sitä monasti korjaamaan käsin, jotta saadaan riittävän laadukas lopputulos. Esimerkiksi Googlen käyttämässä semanttisessa Freebase-kannassa arvioitiin automaattisilla menetelmillä päästävän 60-70% oikeellisuuteen ja vasta ihmistyön avulla siedettävämpään yli 90% tasoon. Freebase hyödyntää mm. koneellisesti annotoitua DBPediaa, mutta sen laatua joudutaan parantamaan.

Miten automaattinen annotointi voidaan toteuttaa verkkopalveluna?

Kielteknologiaan ja semantiikkaan perustuvat annotointipalvelut voidaan monilta osin toteuttaa verkkopalveluina, jolloin niitä hyödyntävien yritysten ei tarvitse sellaisia itse kehittää ja ne voivat keskittyä ydinliiketoimintaansa. Hankkeessa kehitetään uusia palvelukonsepteja ja rajapintoja tällaisille palveluilla.

Miten voidaan tunnistaa ja kuvata yhdistetyn tiedon laatua ja luetettavuutta?

Tiedon ylläpidon ja luetettavuuden arviointi perustuu korkeamman asteiseen provenienssitietoon siitä, mistä tieo on peräisin, miten se on tuotettu, kuinka varmaa se on yms. Tämä tiedon dimensio on vielä niukasti tutkittu LD:n osa-alue, mutta havaittu keskeiseksi käytännön työssä mm. muuttuvaa tietoa ylläpidettäessä ja tiedon luetettavuutta arvioitaessa. Hankkeessa kehitetään menetelmiä tiedon laadun ja luetettavuuden tunnistamiseksi ja kuvaamiseksi.

Metadatan laatu ja luotettavuus

Kun linkitettyjen aineistojen tuotannossa joudutaan tekemisiin yhä laajempien ja hajanaisempien tieto aineistojen kanssa, ja käyttöön on otettu yhä automaattisempia annotoinnin menetelmiä, keskeiseksi ongelmaksi sovelluskäytön kannalta on muodostunut aineiston semanttinen laatu. LDF-hankkeessa laatuongelmaan paneudutaan kehittämällä menetelmiä ja työkaluja aineistojen semanttiseen validointiin. Ideana on tunnistaa automaattisesti metadatasta semanttisesti virheelliset tai puutteelliset kohdat skeemamäärittelyjen ja niihin liittyvien ontologioiden avulla, jolloin kohtien automaattinen tai manuaalinen korjaaminen tulee mahdolliseksi ja kustannustehokkaaksi. Mekanismi yleistää XML:stä tutun dokumenttien syntaktisen validoinnin semanttiselle tasolla.

Yhdistetyn tiedon laadun kvaliteetteja ovat: 1) semanttiset virheet (schema validation), 2) epätäsmällisyydet (uncertainty, ambibuity, precision) ja 3) puutteellisuudet (incompleteness) ja 4) luotettavuus. Semanttisena virheenä voidaan pitää esimerkiksi sitä, että teoksen kirjoittajana on organisaatio. Epätäsmällisyyttä on se, että ”Pyhäjärvi” nimellä voidaan viitata jopa kymmeniin eri järviin tilanteesta riippuen. Esimerkki puutteellisesta tiedosta on ns. orpo tieto (orphan), joka ei linkity riittävästi muuhun tietoon. Tiedon luetettavuutta voidaan arvioida mm. samaa tietoa eri lähteiden kautta verraten ja arvioimalla tiedon alkuperän kautta. Esimerkiksi Wiki/DBPediassa Helsingin asukasluku eri kieliversioissa saattaa poiketa huomattavasti toisistaan. Jos erot ovat suuria, ei luotettavaa arviota voida antaa, mutta luultavasti

suomenkielisen version arvio on paikallistuntemuksen kautta luotettavampi kuin monen muun. Vähintään loppukäyttäjä olisi syytä tehdä tietoiseksi siitä, että väkiluvusta on olemassa erilaisia käsityksiä.

Hankkeessa kehitetään uusia menetelmiä, joilla jo olemassa olevan LD-aineiston ja koneellisen päättelyn avulla avulla järjestelmä kykenee ehdottamaan ihmiskäyttäjälle annotointeja, jolloin sisällönkuvailu helpottuu huomattavasti ja metadatan laatu eri annotoijilla saadaan tasalaatuisemmaksi ja kattavammaksi.

Metadatajoukkojen yhdistäminen

Tutkimusalueella 3 keskitytään seuraaviin haasteisiin:

Miten metadataa ja sanastoja kannattaa sillata (align)?

Linked data -aineistojen yhteentoimivuus edellyttää eri tietojoukkojen siltaamista (mapping, alignment) toisiinsa. Tässä ongelmakentässä voidaan erottaa kaksi pääosaa: sanastojen/ontologioiden siltaaminen (ontology mapping/alignment) ja metadata-skeemojen siltaaminen eli skeemaintegraatio. LDF-hankkeessa kehitetään menetelmiä molempiin.

Siltaamisongelma on keskeinen esimerkiksi LOD-pilven aineistoissa – kuvan 2 tietojoukkoja (pallot) yhdistävät kaaret kuvaavat näiden välisiä siltauksia. LDF-hankkeessa tutkitaan kahta lähestymistapaa. Yhtenä lähtökohtana on Vocabulary Mapping Framework (VMF) -järjestelmän[12] kaltainen RDF(S)-perustainen siltaustapa käsitteille ja niiden ominaisuuksille hierarkioiden kautta, toisena aineistojen muunnos alla olevalle harmonisoivalle tietomallille, kuten esimerkiksi ISO:n standardoimassa CIDOC CRM[13] -mallissa ja siihen liitettävässä funktionaalisen luetteloinnin FRBR-järjestelmässä.

Demonstraatiot hyötykäytöstä ja kriittinen arvio haasteista

Hankkeessa kehitettyjä menetelmiä, työkaluja ja palvelualustoja sovelletaan todellisissa käyttötapauksissa ja todellisilla tietoaineistoilla. Tavoitteena ovat toisaalta vakuuttavat demonstraatiot uuden teknologian mahdollisuuksista toisaalta kriittiset arviot teknologian haasteista.

PILOTTIALUEET JA CASE-HANKKEET

Pilotointi jakautuu seuraavassa kuvattavilla sovellusalueille, jotka muodostavat omat temaattiset työpajansa. Kukin pajan osalta esitellään tutkimushankkeen tavoitteena oleva uusi hyöty yleisemmin, ja sitten hieman tarkemmin pilottikohteita, joiden case-demonstraattorien kautta uutta teknologiaa päästään tutkimaan, kehittämään ja arvioimaan konkreettisesti.

Pilottien kautta hankkeessa mukana olevat organisaatiot pääsevät perehtymään mahdollisten asiakkaittensa ongelmiin ja saavat projektissa karttuvan osaamisen ja verkottumisen kautta etulyöntiaseman onnistuneiksi osoittautuvia pilotteja jatkossa tuotteistettaessa.

[12] http://www.jisc.ac.uk/whatwedo/projects/vocab-framework.aspx

[13] http://www.cidoc-crm.org/

Kunkin pilotointialueen lopuksi luetellaan siihen liittyviä tärkeimpiä tutkimusongelman omistajia. Projektin resursseja suunnataan eri suuntiin hankkeeseen saatavan yritysrahoituksen suhteessa.

Ydinteknologiat ja työkalut

Piloteissa kehitetään ja hyödynnetään ristiin synergeettisesti samoja LD-teknologioita ja työkaluja joista keskeisimpiä on lueteltu alla. Monien välineiden osalta saadaan lentävä lähtö FinnONTO:ssa jo valmistuneiden välineiden avulla (esimerkiksi ONKI) ja kyse on enempi käyttöönotosta kuin jatkokehittämisestä.

Automaattisen annotoinnin työkakalut. Käytetään olemassa oleva Connexor Oy:n ja avoimia NER kieliteknologia ohjelmistoja (esim. Stanford Tagger), jotka kytketään joustavasti verkkopalveluiden kautta korkeammanasteisiksi palveluiksi. Tieteellisenä tavoitteena on tunnistetaan entiteettien ohella näiden välisiä semanttisia suhteita metadatamallien perusteella monikielisistä aineistoista. Työn lähtökohtana on FinnONTO:n ARPA-järjestelmä, jota kehitetään näihin uusiin suuntiin.

Metadatan siltausvälineet. Erityyppisten tietojen yhdistämisessä keskeinen ongelma on skeemaintegraatio. Case-tutkimuksena kehitetään prototyyppi skeemojen julkaisualustaksi ja keskinäistä siltaamista varten. Lähtökohtana työssä on Vocabulary Mapping Framework (VMF) -ontologia, joka mahdollistaa metadataformaattien elementtien välisten suhteiden määrittelyn RDFS-standardiin perustuen. Järjestelmään on mahdollista lisätä asteittain uusien metadataformaattien elementtejä siltä osin kuin ne sieltä puuttuvat. Järjestelmän sekä elementtien siltausten oikeellisuuden arviointiin käytetään case-tutkimusten aineistoja.

Semanttiset haku- ja suosittelupalvelut. FinnONTO:ssa on valmistunut maailmankin mittaluokassa kilpailukykyinen älykäs triple store –ratkaisu EMO (Mäkelä, Hyvönen, 2011), joka tarjoaa semanttisen haun peruspalveluita useille muille järjestelmille, mm. verkossa koekäytössä olevalle yleisten kirjastojen Kirjasampo.fi palvelulle, Wärtsilä Oyj:n vomalaitosdokumenttien hallinna demolle, Datasuomi-pilotille, Puolustusvoimien normitietokantapalvelulle, ja museoiden ja kirjastojen Kulttuurisammolle. Uudessa hankkeessa kehitetään uusi, korkeammanasteinen haku- ja suosittelupalvelukerros ns. suhteikkohaun tarpeisiin (relational search). Innovaationa on hakukohteiden (dokumenttien) sijasta kysellä ja hakea 1) niissä esiintyvien entiteettien välisiä suhteita (esim. miten henkilö liittyy johonkin yritykseen) ja 2) visualisoinnin sekä datajournalismin pohjana tarvittavia moniulotteista suhdekuvauksia (esim. yritykseen liittyvien erilaisten uutistyyppien aikasarja matriisina).

Datan validointi ja kollaboratiivinen korjaus. Tässä osiossa lähtökohtana on SAHA-HAKO-VERA-järjestelmä (Laitio, 2011), joka on yhdistetty EMO:oon. Sen avulla potentiaaliset epätäsmällisyydet voidaan tuoda esille loppukäyttäjälle tarkastettavaksi. Erityisenä uutena tutkimuskysymyksenä on useamman tietojoukon yhdistämisessä syntyvät epätäsmällisyydet, ts. miten niitä voidaan automaattisesti tunnistaa ja merkata. Nykyinen järjestelmä käsittelee vain yhtä tietojoukkoa ja skeemaa kerrallaan erillisenä. Esimerkiksi RDF(S)-muotoisessa tiedossa transitiivisten periytymisketjujen (rdfs:subClassOf) katkeaminen tietojoukkojen välillä on haasteellista sanastoja yhdistettäessä, metatietomallisen domain- ja range-rajoitukset saattavat muodostua ristiriitaisiksi jne. Tähän liittyy myös kiinnostavia hajautetun kollaboratiivisen sisältötyön haasteita, koska tietojoukkoja kehitetään lokaalisti hajautetusti, mutta tietyt riippuvuudet edellyttävät globaalin tason tarkastelua.

Case ”Laki”: Suomalainen lakitieto semanttisena palveluna

Suomessa on käytössä lukuisia lakiin liittyviä verkkopalveluita: Edita Publishing Oy tarjoaa Suomen lakiteksteistä tuotettuun, XML-muotoiseen Finlex-aineistoon liittyviä maksullisia hakupalveluita. Toinen sähköisiä lakipalveluita tarjoava suomalainen yritys on Talentum Suomen laki -palvelullaan[14] .WSOY Pro tarjoaa verkossa lakipalveluita. Laki24.fi mainostaa olevansa ”Suomen suosituin lakialan palvelu”[15]. Nettilaki.com[16] tarjoaa ilmaisia lakipalveluita, tavoitteena näistä kumpuava maksullinenkin toiminta. Keskeinen tietosisältö näissä on julkinen kuudesta eri osiosta koostuva Finlex – Valtion säädöstietopankki, joka on saatavilla XML-muodossa.

Finlexin painopiste on lakiteksteissä, sisällönkuvailuun on kiinnitetty vähemmän huomiota, ja esimerkiksi aineistoa käyttävä Suomi.fi-palvelu joutuu asiasanoittamaan tarvitsemiaan sisältöjä itse. Yksi tähän liittyvä lisähaaste on oikeusalan aineistoihin liittyvien sanastojen moninaisuus ja hajanaisuus. Siihen oikeusministeriön ”Oikeushallinnon asiasanaston kehittämishanke”[17] on parhaillaan kehittämässä uusia ratkaisuja sekä arvioimassa ontologisoinnin hyötyjä liittyen mm. FinnONTO:ssa kehitettyyn julkishallinnon JUHO-ontologiaan ja eduskunnan sanastoihin.

Laki on kenties vanhin hallinnollisen avoimen tiedon muoto. Oikeusvaltion periaatteisiin kuuluu, että kansalaisten, yritysten ja hallinnon toimijoiden oletetaan tuntevan lain. Haasteena on kuitenkin lakien ja säädösten muuttuminen yhä moninaisemmiksi ja monimuotoisimmiksi mm. EU:n lainsäädännön vaikutuksesta. Verkkopalveluista on muodostunut keskeinen juridisen tiedon julkaisukanava, kansainvälisenä esimerkkinä mm. Britannian UK Legistlation[18] ja Hollannin Wetten.nl[19] ja kotimaassa Edita Publishing Oy, Talentum Oy:n ja WSOY Pron palvelut. Myös ensimmäiset seuraavan polven Linked Data –teknologiaa hyödyntävät säädöspalvelut ovat jo verkossa, kuten MetaLex Document Server (Hoekstra, 2011) Hollannissa.

Case-tutkimuksessa selvitetään LD-teknologian mahdollisuusia tehostaa lakeihin liittyviä haku ja suosittelupalveluita automaattisen lakiteksteihin perustuvan annotoinnin avulla. Visiona on demonstraatio Suomen lakitieto semanttisena verkkopalveluna. Tällaisen palvelun käyttötapauksia ja loppukäyttäjiä olisivat mm. (Hoekstra, 2011):

1. *Yritysten* kannalta keskeinen käyttötapaus on niiden omaan toimintaan liittyvien säädösten tuntemuksen hallinta ja säädösten muuttumisen tarkkailu kuten ts. palvelu, jonka avulla yrityksiä voitaisiin automaattisesti tiedottaa niiden toimintaan liittyvien säädösten ja lakien muutoksista.

[14] http://www.suomenlaki.com/

[15] http://www.laki24.com

[16] http://www.nettilaki.com/

[17] http://www.hare.vn.fi/mHankePerusSelaus.asp?h_iId=16997

[18] http://www.legislation.gov.uk/

[19] http://wetten.overheid.nl/zoeken/

2. *Lakimiesten* kannalta kriittinen käyttötapaus on mm. säädösten ja lakien aiempien versioiden hallinta, joita tarvitaan taannehtivasti mm. oikeusjutuissa.
3. *Hallinnon ja lainsäädännön* kannalta merkittävä käyttötapaus on tarjota tukea uusien säädösten vaikutuksista ja mahdollisista ristiriidoista toisten säädösten ja lakien suhteen. Kriittisten lakipalveluiden tarjoaminen mahdollistaa julkisen datan varaan rakentuvaa liiketoimintaa.
4. Lakia huonosti tuntevat *henkilöt* tarvitsivat palvelua, jolla voisivat hakea ja selailla lainsäädäntöä nykyistä tehokkaammin.
5. *Toiset verkkopalvelut* voisivat hyödyntää lakiaineistoja palveluina ulkoisissa tietojärjestelmissä, esimerkiksi Suomi.fi-portaali tai Puolustusvoimien normitietokanta PAHVI Finlexiä LD-rajapintojen kautta.

Pilotin ydinaineistoina on Finlex, juridiikan sanastot sekä tarpeen mukaan yritysten palveluiden aineistoja, kuten kommentaareja, artikkeleita jne.

Hankkeessa kehitetään malli Suomen lakien ja säädösten (erityisesti FinLex) ja siihen liittyvien ontologioiden julkaisemisesta Linked Data –muodossa. Tavoitteena olevan mallin ideana on tukea älykästä juridista tiedonhakua edellä kuvatuissa käyttötapauksissa, eli kehittää nykyisiä palveluita entistä ”älykkäämmiksi”, sekä ideoida kokonaan uusia palvelumuotoja uuden teknologian avulla. Demonstraatiojärjestelmänä kehitetään 1) semanttinen LD-perustainen rajapintapalvelu ja 2) ihmiskäyttäjille suunnattu järjestelmä, joka em. rajapintapalvelun avulla luo lisäarvoa edellä kuvatuissa käyttötapauksissa. Näitä tarkennetaan yhteistyössä nykyisten lakipalveluiden tarjoajien kanssa. Ajatuksena on luoda avoin kansallinen LD-versio nykyisestä Finlex-palvelusta, jonka varaan kaupalliset palvelutarjoajat ja julkishallinto voivat kustannustehokkaasti kehittää uusia aiempaa älykkäämpiä lisäpalveluita.

Case ”Media”: Mediayrityksen sisällönhallinta ja datajournalismi

Case-tutkimuksen kohteena on LD-aineistojen ja teknologian hyödyntäminen mediayritysten sisällöntuotannon, -hallinnan ja -julkaisun kannalta. Alan mahdollisuuksista on olemassa jo ensimmäisiä lupaavia kansainvälisiä referenssejä, kuten The New York Timesin LOD-aineisto[20] ja BBC:n LD-perustainen WWW-julkaisujärjestelmä (Kobilarov, 2009). Siinä yhtiön eri osastojen aineistojen metatietoja harmonisoidaan ja yhdistetään toisiinsa DBPedian avulla ja rikastetaan semanttisesti paitsi toisillaan myös ulkoisilla LOD-pilven tietojoukoilla. Tutkimuksen motiivina on, että laadukkaan yhdistetyn metatiedon varaan voidaan kehittää aiempaa älykkäämpiä haku- ja suosittelupalveluita sekä tarjota asiakkaille semanttisesti rikastettuja tietosisältöjä. Lisäksi datajournalismin keinoin yhdistetystä tiedosta voidaan luoda uutta tietoa, jota ei erillisistä aineistoista erikseen ole nähtävissä, ja visualisoida tietoa uusilla yhdistetyn tiedon visualisoinnin menetelmillä (Dadzie, Rowe, 2011).

Mediayrityksen sisällöntuotannon keskeinen elementti tiedon haun ja yhdistämisen kannalta on metatieto. Kun yrityksen sisällöntuotannossa pyritään koko yrityksen toiminnan kattavaan yhdistettyyn tietoon sirpaleisten tietosiilojen sijasta, haasteeksi muodostuu toisaalta eri aloilla käytettyjen käsitteistöjen yh-

[20] http://data.nytimes.com/

teentoimimattomuus, toisaalta heterogeenisten aineistojen metatietomallien siltaaminen (alignment, mapping) toisiinsa. Media-alalla yksi erityisongelma on nopeasti muuttuvien sisältöjen ja sanastojen hallinta. Esimerkiksi uusia uutistapahtumia syntyy päivittäin ja uusia henkilöitä nousee nopeasti julkisuuden valokeilaan. Jotta eri tahoilla samaan tapahtumaan tai henkilöön liittyvät tiedot voidaan yhdistää, vaikkapa kaksi uutista samaan tapahtumaan liittyen, on sisällöntuotannossa voitava reaaliaikaisesti luoda ja jakaa näitä kuvaavia käsitteitä. Lisäksi sisällönkuvailu esimerkiksi jatkuvan uutisvirran osalta pitäisi kustannussyistä saada käytännössä lähes automaattiseksi, mikä johtaa haasteisiin metatietojen laadun ja validoinnin osalta.

Case-tutkimuksessa kehitetään malli ja välineistö yhdistetyn metatiedon sisällöntuotantoprosessista mediayrityksen erityistarpeiden lähtökohdista. Mallin keskeisiä komponentteja ovat sanastopalvelu, metatietojen siltaamismalli ja -palvelu, automaattinen annotointipalvelu sekä metadatan validointi- ja korjauspalvelu. Ratkaisumallissa pyritään hyödyntämään jo olemassa olevia komponentteja ja open source -työkaluja, monessa tapauksessa komponentteja kuitenkin joudutaan kehittämään edelleen. Dynaamisten sanastojen reaaliaikaiseen hallintaan kehitetään ratkaisu nykyisen staattisen ONKI-palvelun laajennuksena, metatietojen siltaamispalvelun osalta lähtökohtana on aiemmin mainittu VMF-ratkaisumalli, jota kuitenkin pitää soveltaa media-alalle. Annotointipalvelun osalta lähtökohtana ovat olemassa olevat Named Entity Recognition (NER) -ohjelmistot ja Connexor Oy:n Machinese, joita hyödynnetään relaatioiden tunnistamiseen ja monikielisten aineistojen (suomi, ruotsi, englanti) annotointiin. Metadatan validoinnin ja korjailun osalta testataan ja kehitetään edelleen FinnONTO-hankkeessa kehitettyyn SAHA-HAKO-ympäristöön vast'ikään luotua uutta metadatan merkkaus- ja validointijärjestelmää (Laitio, 2011).

Hankkeessa tuotettavat avoimesti julkaistavissa olevat tietosisällöt julkaistaan avoimena yhdistettynä tietona LD-periaatteiden (Heath, Bizer, 2011) mukaisesti (source tiedostot, URI-ohjaukset, SPARQL-rajapinta), käyttäen hyväksi LDF-hankkeen ydinosassa kehitettäviä välineitä.

Työpajassa kehitetään kaksi demonstraattoria.

1) Kansallinen media LOD-palvelu "Melod". Demonstraattorin tietosisältönä on YLE:n, HS:n ja kansainvälisten toimijoiden julkaisemaa avointa tietoa. Kehitettävän mallin ja komponenttien case-ympäristönä on Yleisradion kehitteillä oleva semanttinen Drupal-perustainen julkaisujärjestelmä, jolle tarjotaan tarvittavat palvelurajapinnat.

2) Semanttinen Kansallisbiografia. Toisena case-ympäristönä on SKS:n julkaisema ja myymä Kansallibiografia (n. 6000 toimitettua elämänkertaa) ja SLS:n vastaava ruotsinkielinen julkaisu, sekä avoimena tietona muualta saatava elämänkertatieto (mm. Freebase, DBPedia ja kirjastojen ns. auktoriteettitiedostot). Pilottijärjestelmänä kehitettä biografia-aineistoille semanttinen palvelu, jonka avulla elämänkertatietoon voi tutustua verkon kautta semanttisesti rikastetusti ja visualisointien kautta. Semanttista Kansallisbiografiaa voivat käyttää sekä ihmiset että toiset palvelut, kuten Melod, rajapintojen kautta.

Semanttisesti annotoitujen aineistojen hyödyntämistä testataan kahdella tavalla: 1) Aineistojen semanttista hakua ja suosittelua demonstroidaan ja testataan avoimen datan hakukoneilla ja suositterurajapintojen kautta. 2) Tiedonlouhintaa ja aineistojen analyysiä testataan datajournalismin, suhteikkohaun (relation search, knowledge disvovery) ja yhdistetyn tiedon visualisoinnin keinoin. Datajournalismin tavoittee-

na on löytää uutta tietoa yhteiskunnan ilmiöistä yhdistämällä eri tietoaineistoja Linked Data -periaatteella, analysoimalla niitä algoritmisesti ja visualisoimalla tuloksia.

Yhteistyössä Helsingin Sanomien kanssa järjestetään tähän liittyviä HS Open -hakkerointipäiviä, joissa avoimen tiedon hyödyntäjille tarjotaan mahdollisuus kehittää sovelluksia ja analyyseja projektin tuottamista avoimista yhdistetyistä tietoaineistoista. Tilaisuudet ovat jatkoa Semanttisen laskennan tutkimusryhmän ja Helsingin Sanomien HS Open -tapahtumille[21], joiden tulokset ovat osoittautuneet lupaaviksi, ja joista on raportoitu HS:n sivuilla.

Case ”Yritys”: Yritystietojen semanttinen rikastaminen uutisaineistoilla

LD-ajattelun keskeinen tavoite on sisältöjen rikastaminen tietoa yhdistämällä, mikä voi tapahtua pienemmässä mittakaavassa esimerkiksi yrityksen sisällä tai äärimmilleen vietynä yli koko webin. Yhdistämällä tietoa voidaan hajautetusti tuotetut aineistot saada kustannustehokkaasti laajempaan käyttöön, yhdistetystä tiedosta voidaan löytää uutta tietoa jota erillisistä aineistoista ei ole mahdollista nähdä, ja yhdistämisessä käytettävä uusi web-teknologia tarjoaa uudenlaisia rakenteisia mahdollisuuksia tiedon louhintaan ja visualisointiin. Case-tutkimuksen aiheena on tutkia julkisen tiedon hyödyntämistä osana yksittäisen yrityksen tarjoamaa palvelua esimerkkinä yritys, jossa julkisen tiedon yhdistämistä on aiemmin tehty perinteisin menetelmin. Näin tarjoutuu mahdollisuus kehittää ja arvioida uuden teknologian lisäarvoa, mahdollisuuksia ja haasteita todellisessa liiketoimintatapauksessa.

Suomen Asiakastieto Oy yhdistää tietokantoihinsa dataa kymmenistä eri lähteistä, kuten kapparekisteristä ja luottotietorekisteristä. Perinteisin tavoin yhdistetyn tiedon avulla yritys tarjoaa yrityksille, pankeille, vakuutuslaitoksille ja muille luottotietoja tarvitseville tietoa ja analyysejä yrityksen ja henkilöiden liiketoiminnasta, kytkennöistä, maksuhäiriöistä jne. Perustana ovat julkiset viranomaisrekisterit, joissa olevaa perustietoa yhdistetään ja rikastetaan lisäksi käsityönä mm. yritysten web-sivuja analysoiden.

Rakenteisen ”kovan” tiedon lisäksi olisi hyödyllistä tarjota yritystietoa käyttäville asiakkaille myös ns. ”pehmeää”, rakenteetonta tietoa yrityksistä ja henkilöistä, jota on saatavilla rekisterien ulkopuolta eri lähteistä. Esimerkiksi yrityksiin liittyvien talousuutisten ja webin verkkosivujen kautta voidaan yrityksestä saada tärkeää lisätietoa luottopäätöksen tueksi, ennustettaessa yrityksen liiketoiminnan jatkonäkymiä tai markkina- ja kilpailija-analyyseja tehtäessä. Tutkimuksen ongelmana on selvittää, miten yritystieto myyvän yrityksen asiakkaille voitaisiin tuottaa lisäarvoa verkon uutisaineistoja ja hakutuloksia koneellisesti rakenteistamalla ja siltaamalla aineistoja RDF-muodossa olevaan rekistereiden yritystietoon ja muihin verkon LD-aineistoihin. Käytännön tasolla tämä merkitsisi sitä, että asiakkaalle verkkopalvelussa generoidun yritysanalyysin yhteyteen voitaisiin koneellisesti liittää esimerkiksi yritykseen ja sen toimiaan liittyviä tuoreita uutisia, tietoa yrityksen taustalle olevista ydinhenkilöistä, yrityksen yhteistyöstä toisten yritysten kanssa yms. Tällainen tieto antaisi lisävalaistusta heikkoina signaaleina yrityksen tulevaisuudesta, mikä täydentäisi rekistereistä saatavaa enemmän historiaan liittyviä aineistoja, kuten tilinpäätös- tai maksuhäiriötietoja. Nyt tämän tyyppistä työtä tehdään ihmisvoimin esim. hakukonetta eri hakusanoilla kokeilemalla ja tuloksia yhdistelemällä sekä yritysten verkkosivuja analysoimalla. Tutkimushypoteesi-

[21] http://blogit.hs.fi/hsnext/tama-on-kutsu-hs-open-14-3-sanomatalossa

na on, että kone voisi toimia ihmisen apuna tiedon mielekkäässä harvestoinnissa, jäsentämisessä ja liittämisessä yritysten muihin tietoihin.

LOD-verkko ja metadatasiltaukset muodostetaan ydinteknologiaosuudessa kehitetyillä välineillä semanttiseksi tietorepositorioksi (triple store) ja kehitetään semanttiselle suhteikkohaulle (relation search, knowledge discovery) tarvittavat rajapinnat. Ihmiskäyttäjää varten kehitetään palvelurajapintaa hyödyntävä logiikkaperustainen raporttigeneraattori ja visualisaattori, joka semanttisen haun, yhteyshaun, ja suosittelun keinoin tuo lisäarvoa Asiakastiedon asiakkaille ja datajournalisteille.

Case ”Palvelu”: Semanttinen palvelukartta

Yhdistettyyn tietoon liittyy monesti paikka- ja aikatietoa, mikä mahdollistaa tiedon karttapohjaisen haun, paikka- ja aikakontekstiin perustuvan käytön sekä tietojen visualisoinnin ajan suhteen ja kartalla. Yksi kiinnostava tietolaji tässä mielessä ovat palvelut, joita tarjotaan tietyssä paikassa tiettyyn aikaan, ja joita myös usein tarvitaan tietyssä paikka/aikakontekstissa. Alalle on Suomessakin syntynyt liiketoimintaa ja palvelukonsepteja, kuten Tässä.fi ja Citynomadi.fi. Klassinen esimerkki yhdistyn tiedon tarjoamisesta karttanäkymän kautta on Wikipedia- ym. aineistoja kartoilla ja mobiililaitteella tarjoileva Mobile DBPedia (Becker, Bizer, 2008).

Case-tutkimuksen kohteena ovat kunnalliset palvelut ja niiden hyödyntäminen verkko- ja mobiilisovelluksissa aika- ja paikkakontekstissa. SeCo-tutkimusryhmä on ollut mukana kehittämässä uutta, vuoden 2011-2012 vaihteessa valmistuvaa JHS 145 –suositusta palveluiden metatietomallista. Kyseessä on ensimmäinen semanttisen webin teknologiaa hyödyntävä JHS-suositus maassamme. Case-tutkimuksessa mallia sovelletaan ja testataan Helsingin ja pääkaupunkiseudun palvelukartta-aineistoon, joka on tällä hetkellä monipuolisin saatavilla oleva tietoaineisto kuntien palveluista. Ideana on muodostaa palvelukartan tietokannasta avoimen datan aineisto, joka on yhdistetty kansalliseen FinnONTO-ontologiainfrastruktuuriin. Näin kunnallinen palvelutieto voidaan yhdistää muihin niihin liittyviin aineistoihin, esimerkiksi terveyspalvelut TerveSuomi.fi:n terveystietoihin ja kulttuuripalvelut museoiden ja kirjastojen kokoelmatietoihin. Ideana on julkaista palveluaineistot LD-periaatteiden mukaisina rajapintapalveluina, joiden varaan voidaan kehittää sovellukset sekä verkko- että mobiilikäyttäjille. Jälkimmäisessä tapauksessa hyödynnetään vuoden 2012 keväällä päättyvän Tekesin SUBI-hankkeen tuloksia ja open source –ratkaisuja, joissa on kehitetty semanttista, personoitua, mobiilia Matkailusampo-palvelua kulttuurikohteille (Mäkelä et al., 2011b).

Tutkimuksen käyttötapauksena on Helsinki World Design Capital -juhlavuoden aikana Helsinkiin saapuva ulkomainen tai kotimainen matkailija. Hän tarvitsee toisaalta tietoa sekä lähikaupungeissa (Espoo, Vantaa, Lahti) eri aikoina ja eri paikoissa järjestettävistä tapahtumista (esimerkiksi näyttelyt, konsertit yms.), toisaalta tietoa saatavilla olevista muista palveluista, kuten terveyspalveluista mahdollista sairastapausta varten ja poliisin palveluista taskuvarkauden johdosta, kulttuuripalveluista tai vaikkapa tietoa kaupungin ulkoilureiteistä virkistäytymistä varten.

Lisäarvona markkinoilla jo oleviin navigaattoreihin ja Tässä.fi-tyyppisiin järjestelmiin on mahdollisuus päästä käsiksi paitsi kohteisiin kartoilla, myös niihin eri tavoilla liittyviin muihin tietoihin ja palveluihin. Haun ja suosittelun osalta lisääntynyt semanttinen kuvailu mahdollistaa mm. aukioloaikojen (esimerkiksi museon kiinniolo maanantaisin) tai palvelun varaustilanteen (esimerkiksi loppuunmyyty tilaisuus) hyö-

dyntämisen käyttötilanteessa. Tärkeää on myös mahdollisuus järjestelmän personointiin käyttäjän preferenssien mukaan, mihin semanttiset merkkaukset antavat uusia mahdollisuuksia.

Tuloksena syntyy LD-perustainen rajapintapalvelu kunnallisten palveluiden löytämiseksi paikka- ja aikakontekstissa. Palvelun käyttökelpoisuutta testataan liittämällä se 1) ulkoiseen portaaliiin ja 2) mobiililaitteeseen käytettäväksi paikan päällä. Uutuusarvona on palveluiden ja tapahtumien semanttinen kuvailu, jonka avulla voidaan loppukäyttäjälle tarjota automaattisesti lisätietoa yhdistetyn tiedon kautta. Työn on jatkoa tutkimusryhmän kansainvälisesti palkitulle SmartMuseum- järjestelmälle ja Matkailusampolle.

HYÖDYNTÄMISSUUNNITELMA

Yleiset tavoitteet

Hanke kehittää avointa, W3C:n suosituksia pääsääntöisesti noudattavaa LD-ydinteknologiaa. Tulokset julkaistaan pääsääntöisesti MIT-lisenssin alla, mikä edistää ratkaisujen kansallista käyttöönottoa ja mahdollistaa maksimaalisen kaupallisen hyödyntämisen:

MIT License

Copyright (c) <year> <copyright owner> Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software. THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Hanke toteutetaan yhteistyössä Tivit Oy:n Data to Intelligence ja NextMedia –-ohjelmien kanssa, joiden tarpeita ja mahdollisesti saman suuntaista kehitystyötä pyritään huomioimaan ydinteknologian ja työkalujen kehittämisessä. LDF-hankkeen käyttämä MIT-lisenssi on ongelmaton tulosten siirron kannalta. LDF-hankkeen tutkimustyö ja kehitettävä teknologia ei ole yksittäisen yrityksen tuotekehitystä, vaan pyrkii löytämään case-ongelmiin yleisempiä ratkaisuja, joita voidaan soveltaa paitsi kyseisessä case-ongelmassa myös laajemmin muissa vastaavissa. Ydinteknologiaa sovelletaan yritysmaailman haasteista syntyneissä case-tutkimuksissa uuden teknologian käyttökelpoisuuden testaamiseksi.

Yhteistyön muodoista sovitaan tarkemmin D2I- ja NextMedia-ohjelmien kanssa, kun niiden tulevien vuosien tutkimustavoitteet ja toiminta selviävät syksyn 2011 kuluessa.

	Johtoryhmän kokoonpano (24.11.2011)		
	Yritys / organisaatio	**Jäsen**	**Varajäsen**
1	Citynomadi Oy	Merja Taipaleenmäki	
2	Connexor Oy	Pasi Tapanainen	Mirkka Tapanainen
3	CSC - Tieteen tietotekniikan keskus Oy	Pirjo-Leena Forsström	
4	Edita Publishing Oy	Jari Linhala	Päivi Helander
5	Lean Development Oy	Mikko Mäkelä	
6	Sanastokeskus TSK	Katri Seppänen	
7	Sanoma News Oy	Esa Mäkinen	
8	Suomalaisen Kirjallisuuden Seura SKS	Lauri Harvilahti	
9	Suomen Asiakastieto Oy	Pertti Vahermaa	
10	Svenska Litteratursällskapet SLS	Karola Söderman	
11	Talentum Oy	Stina Wikberg	
12	Tieto Oyj	Lasse Akselin	
13	Yleisradio Oy	Sami Kallinen	
1	Oikeusministeriö	Silja Korvenmaa	
2	Liikenne- ja viestintäministeriö	Taru Rastas	
3	Valtiovarainministeriö	Jukka Uusitalo	Anne Kauhanen-Simanainen
4	Helsingin kaupunginkirjasto	Erkki Lounasvuori	Matti Sarmela
5	Helsingin kaupunki	Mirjam Heikkinen	Leila Oravisto
6	Suomenlinnan hoitokunta	Maire Mattinen	
	Asiantuntijajäsenet		
7	W3C Suomen toimisto	Ossi Nykänen	
	Tutkimustahot		
8	Aalto-yliopisto	Eero Hyvönen	
9	Helsingin yliopiston tietojenkäsittelytieteen laitos	Jouni Tuominen	
	Päärahoittaja		
10	Tekes	TBA	

Kuva 3. LDF-hankkeen konsortio ja johtoryhmä.

Tulosten siirto

Projekti järjestää säännöllisesti projektin sisäisten tilaisuuksien ohella myös avoimia tilaisuuksia, joissa esitellään projektin tuloksia ulkopuolisille tahoille ja keskustellaan toiminnan suuntaamisesta. Aiemmassa FinnONTO-hankeessa tällaisiin tilaisuuksiin on saapunut poikkeuksetta satoja osanottajia.

Tuloksista raportoidaan paitsi kansainvälisillä tieteellisillä foorumeilla, myös kotimaisessa lehdistössä ja tilaisuuksissa.

Tutkimustulosten siirto yritysten ja tietoyhteiskunnan käyttöön on yksi projektin ydinlähtökohdista: hankkeen keskeinen tavoitehan on yhteentoimivien tietojoukkojen ja metatietomallien hyödyntäminen edistäminen maassamme mahdollisimman tehokkaasti LD-teknologian ja uudenlaisten, nykyistä ONKI-palvelua laajentavien verkkopalvelukonseptien ja sovellusten kautta.

PROJEKTIN ORGANISOITUMINEN

Hankkeen johtoryhmä projektin alussa sekä ja hankkeeseen osallistuvat organisaatiot on lueteltu kuvassa 3. Vastuullisena johtajana toimii prof. Eero Hyvönen.

AIKATAULU

Hanke jakautuu kahteen vaiheeseen, josta ensimmäiselle on 1.1.2012-30.6.2013 on myönnetty Tekesin tukea. Hankeen volyymi kasvaa kevään 2012 aikana täyteen mittaansa SeCo-tutkimusryhmässä käynnissä olevien SUBI- ja FinnONTO-hankkeiden päättyessä ja näiden tukijoiden siirtyessä uusiin hankkeisiin.

YHTEENVETO

Kehitettävä teknologia, innovaato ja osaaminen

Hankkeessa tutkitaan ja saadaan käytännön kokemusta uusimman linked data -teknologian mahdollisuuksista synnyttää tietovarantolähtöisiä innovaatioita maahamme, Näistä on muodostumassa keskeinen kehitysalue monissa maamme tietoyhteiskuntahankkeissa ja liiketoiminnan piirissä. Tutkimuksellisina uusina kehityssuuntina ovat yhdistetyn tiedon laadun varmistus, muutoksen hallinta sekä kontekstuaalinen hyödyntäminen. Työssä jatketaan FinnONTO:ssa luotua akateemista tutkimusperinnettä ja aktiivista tulosten julkaisemista alan huippukonferesseissa ja journaaleissa. Hanke on laaja-alainen ja muodostaa alansa moniin kansainvälisiin huippuyliopistoihin linkittyneen kansallisen keihäänkärkihankkeen. Tuloksena on uusimpien web-teknologioiden osaamisen kehittyminen Suomessa niin yliopistoissa kuin laajan konsortion sisällä, konkreettisia uusia tietoaineistoja yhteiskunnan ja liike-elämän käyttöön, sekä avoimia ohjelmistoja ja uusia pilotoituja verkkopalvelukonsepteja.

Yhteistyössä koko arvoketju

Hanke on laaja-alainen ja perustuu läheiseen yhteistyöhön koko LD-arvoketjun edustajien lävitse: Mukana on tutkimustahojen lisäksi tietoaineistojen tuottajia, palvelun tarjoajia, sisältötyön kehittäjä, ohjelmistotalo ja kansainvälinen standardointiorganisaatio. Hankkeeseen palkataan tarpeen mukaan erityisalojen osaajia konsortion jäsenorganisaatioista, mikä edesauttaa sekä projektin työtä että tulosten siirtoa käytäntöön. Tästä on hyviä kokemuksia aiemmasta FinnONTO-hankkeesta.

Hyödyntäminen

Hankkeessa tutkitaan ja kehitetään uutta teknologiaa ongelmalähtöisesti todellisten käyttötapausten ja aineistojen pohjalta. Tulokset ovat kuitenkin yhtä sovellusta yleisempiä ja ne julkaistaan Living Laboratory -tyyppisinä verkkopalveluina sekä konsortion jäsenten käytettäväksi että laajemmankin yleisön.

Resurssit

LDF-hankkeen tutkijoiksi saadaan kansainvälisen tason semanttisen webin tutkijatiimi päättyvistä FinnONTO- ja Semanttiset jokapaikan palvelut -hankkeista. Projektiin liittyy useita kansainvälisiä yhteistyö-

hankkeita. Aalto-yliopisto ja Helsingin yliopisto maamme kahtena suurimpana yliopistona tarjoavat hankkeelle hyvän tutkimuksen infrastruktuurin.

Hyvinvointitekijät yhteiskunnalle / ympäristölle

Avoimen tietoon liittyvät kysymykset, kuten oikeus tiedon demokraattiseen käyttöön, tietovarantoperustainen innovaatiotoiminta ja valtiotason yhteentoimivuuskysymykset, ovat nopeasti nousseet Iso-Britannian ja USA:n vanavedessä tietoyhteiskunnan ydinkysymyksiksi monissa eri maissa Suomi mukaan lukien. Hankkeessa tutkitaan ja kehitetään teknologian lähtökohdista näihin kysymyksiin liittyviä toiminta- ja teknisiä ratkaisuratkaisumalleja, joita myös pilotoidaan käytössä.

KIRJALLISUUTTA

1. Eeva Ahonen and Eero Hyvönen: Publishing Historical Texts on the Semantic Web - A Case Study. Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC2009) (forthcoming), Berkeley, CA, USA, September, 2009.
2. Sören Auer, Thomas Riechert, and Sebastian Dietzold: OntoWiki - A Tool for Social, Semantic Collaboration. Proceedings of ISWC 2006, Athens, USA. Springer-Verlag, 2006.
3. Christian Bizer, Tom Heath, Tim Berners-Lee (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, Vol. 5(3), Pages 1-22.
4. C. Becker, C. Bizer: DBpedia Mobile: A Location Enabled Linked Data Browser. Proc. of LDOW2008, April 22, 2008, Beijing, China.
5. CSC – Tieteen tietoteknikan keskus: Tutkimuksen tietoaineistot - olennaisen käsikirja päättäjille. CSC, 2001, 146 pp. http://www.csc.fi/csc/julkaisut/oppaat/2010/tutkimuksen-tietoaineistot
6. Aba-Sah Dadzie, Matthew Rowe: Approaches to Visualising Linked Data: A Survey. Journal of Semantic Web, 2011, IOS Press, forth-coming. http://www.semantic-web-journal.net/sites/default/files/swj118_1.pdf
7. Marc Ehrig: Ontology Alignment. Bridging the Semantic Gap. Springer-Verlag, 2007.
8. Jerome Euzenat, Pavel Shvaiko: Ontology Matching. Springer-Verlag, 2007.
9. Dieter Fensel, James A. Hendler, Henry Lieberman, and Wolfgang Wahlster (eds.): Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, MIT Press, 2005.
10. Matias Frosterus and Eero Hyvönen: Bridging the Search Gap between the Web of Pages and Web of Data by Combining Ontological Document Expansion with Text Search. Proceedings of the International Conf*erences* on Digital Libraries and the Semantic Web 2009 (ICSD2009), Trento, Italy, September, 2009.
11. Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 2011.

12. Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. Gumo – the general user model ontology. In User Modeling 2005: 10th International Conference, 2005.

13. Rinke Hoekstra: The MetaLex document server. Legal documents as versioned Linked Data. In: Proc. of ISWC 2001. Springer-Verlag, forth-coming.

14. Eero Hyvönen, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila and Suvi Kettula: MuseumFinland - Finnish Museums on the Semantic Web. Journal of Web Semantics, vol. 3, no. 2, pp. 25, 2005.

15. Eero Hyvönen, Kim Viljanen, Eetu Mäkelä, Tomi Kauppinen, Tuukka Ruotsalo, Onni Valkeapää, Katri Seppälä, Osma Suominen, Olli Alm, Robin Lindroos, Teppo Känsälä, Riikka Henriksson, Matias Frosterus, Jouni Tuominen, Reetta Sinkkilä and Jussi Kurki: Elements of a National Semantic Web Infrastructure - Case Study Finland on the Semantic Web (Invited paper). Proceedings of the First International Semantic Computing Conference (IEEE ICSC 2007), Irvine, California, September, 2007a. IEEE Press.

16. Eero Hyvönen, Kim Viljanen and Osma Suominen: HealthFinland-Finnish Health Information on the Semantic Web. Proceedings of ISWC/ASWC 2007, Pusan, Korea. Springer-Verlag, 2007.

17. Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, Katariina Nyberg: CultureSampo - Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user. Proceedings, Museums and the Web 2009, Indianapolis, USA, April 15-18, 2009.

18. Eero Hyvönen, Jouni Tuominen, Tomi Kauppinen, Jari Väätäinen: Representing and Utilizing Changing Historical Places as an Ontology Time Series. Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications (Naveen Ashish and Amit Sheth (eds.)), Springer-Verlag, 2011, forthcoming. Book chapter.

19. Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov and Damyan Ognyanoff: Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web.Volume 2, Issue 1, 1 December 2004, Pages 49-79.

20. Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Robert Lee: Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. European Semantic Web Conference, Proceedings, Springer-Verlag, 2009.

21. Jussi Kurki and Eero Hyvönen: Collaborative Metadata Editor Integrated with Ontology Services and Faceted Portals. Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece, CEUR Workshop Proceedings, http://ceur-ws.org/, June, 2010.

22. Joonas Laitio: Semantic Web Data Quality Control. MSC Thesis, Aalto University, Department of Media Technology, 2011.

23. Robin Lindroos: Paikkatiedon ontologiapalvelu. MSc Thesis, Helsinki University of Technology (TKK), May, 2008

24. Eetu Mäkelä, Kim Viljanen, Olli Alm, Jouni Tuominen, Onni Valkeapää, Tomi Kauppinen, Jussi Kurki, Reetta Sinkkilä, Teppo Känsälä, Robin Lindroos, Osma Suominen, Tuukka Ruotsalo and Eero Hyvönen: Enabling the Semantic Web with Ready-to-Use Mash-Up Components. Proceedings of the First Industrial Results of Semantic Technologies Workshop, ISWC2007, Busan, Korea, 2007.

25. Eetu Mäkelä, Aleksi Lindblad, Jari Väätäinen, Rami Alatalo, Osma Suominen and Eero Hyvönen: Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources -- The TravelSampo System. 2011. Accepted for publication, Terra Cognita Workshop, ISWC2011

26. Eetu Mäkelä, Eero Hyvönen and Tuukka Ruotsalo: How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability, 2011b. Accepted for publication.

27. Reijo Paajanen and Pauli Kuosmanen: Tivit White Paper: Finland and Data Reserves, Tivit Oy, Espoo, 2010.

28. Jouni Tuominen, Nina Laurenne and Eero Hyvönen: Biological Names and Taxonomies on the Semantic Web - Managing the Change in Scientific Conception. Submitted for review, 2011.

29. Phillip Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (eds.): Semantifc Computing, IEEE Wiley - IEEE Press, May, 2010.

30. Reetta Sinkkilä, Osma Suominen and Eero Hyvönen: Automatic Semantic Subject Indexing of Web Documents in Highly Inflected Languages. December, 2010. Submitted for review

31. Osma Suominen, Eero Hyvönen, Kim Viljanen and Eija Hukka: HealthFinland-a National Semantic Publishing Network and Portal for Health Information. Journal of Web Semantics, vol. 7, no. 4, pp. 271-376, Dec, 2009.

32. Osma Suominen and Eero Hyvönen: Expressing and Aggregating Rich Event Descriptions. Proceedings of the 6th Workshop on Scripting and Development on the Semantic Web, Heraklion, Greece, May, 2010.

33. Jouni Tuominen, Matias Frosterus, Kim Viljanen and Eero Hyvönen: ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services. Proceedings of ESWC-2009, Springer-Verlag, 2009.

34. Onni Valkeapää, Olli Alm and Eero Hyvönen: Efficient Content Creation on the Semantic Web Using Metadata Schemas with Domain Ontology Services. Proceedings of the European Semantic Web Conference ESWC 2007, Innsbruck, Austria, Springer, June 4-5, 2007.

35. YiwenWang, Natalia Stash, Lora Aroyo, Peter Gorgels, Lloyd Rutledge, and Guus Schreiber: Recommendations based on semantically enriched museum collections. Web Semantics: Science, Services and Agents on the World Wide Web, 6(4):283 – 290, 2008.

36. Antti Vehviläinen, Eero Hyvönen and Olli Alm: A Semi-Automatic Semantic Annotation and Authoring Tool for a Library Help Desk Service. Proceedings of the first Semantic Authoring and Annotation Workshop, ISWC-2007, Atlanta, USA, November, 2006.
37. Kim Viljanen, Jouni Tuominen, Teppo Känsälä and Eero Hyvönen: Distributed Semantic Content Creation and Publication for Cultural Heritage Legacy Systems. Proceedings of the 2008 IEEE International Conference on Distributed Human-Machine Systems, IEEE Press, Athens, Greece, March 9-12, 2008.