HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Semantic Computing Research Group

UNIVERSITY OF HELSINKI

27.1.2004

Suomalaisen semanttisen webin ontologiat

Hanke-esittely 1.9.2003-30.8.2005

Eero Hyvönen

Hankkeen kotisivu:
http://www.cs.helsinki.fi/group/seco/ontologies/

L 26 (Teollisuuskatu 23)
0014 HELSINGIN YLIOPISTO
uhelin: (09) 191 51
elefax: 1914 4441

PB 26 (Industrigatan 23)
00014 HELSINGFORS UNIVERSITET
Tfn: (09) 191 51
Telefax: (09) 1914 4441

P.O. BOX 26 (Teollisuuskatu 23)
FIN-00014 UNIVERSITY OF HELSINKI
Tel: +358 9 191 51
Telefax: +358 9 1914 4441

Tiivistelmä

Artikkelissa esitellään kansallisen Suomalaisen semanttisen webin ontologiat *–hankkeen visio ja tavoitteet. Hankkeessa kehitetään pilottijärjestelmä kansallisesti merkittävien semanttisen webin ontologioiden hajautettua kehitystyötä varten. Tavoitteena on siirtyä tiedon indeksoinnissa ja haussa nykyisestä asiasanatekniikasta semanttisesti rikkaampaan ontologiateknologiaan, mikä mahdollistaa aiempaa käyttäjäystävällisemmän täsmätiedonhaun, tietojärjestelmien semanttisen yhteentoimivuuden ja älykkäät palvelut webissä. Ontologioiden käyttö luo uusia liiketoimintamahdollisuuksia ja mahdollistaa arvokkaiden tietosisältöjen aiempaa tehokkaamman hyötykäytön tietoverkoissa. Ontologiakehityksessä keskeisiä kohteita ovat erityisesti Yleinen suomalainen asiasanasto YSA, museoalan asiasanasto MASA, taidealan ICONCLASS sekä paikkatietoon ja toimijoihin liittyvät ontologiat. Teknologiaa sovelletaan konkreettisissa case-sovelluksissa, joiden käyttökelpoisuus arvioidaan kokeellisesti. Tärkenä testialustana toimii mm.* MuseoSuomi — Suomen museot semanttisessa webissä *-portaali semanttisine hakukoneineen ja suosittelujärjestelmineen*[1].

Tutkimuskonsortio ja johtoryhmä

Hankkeessa vuonna 2003 aloittanut tutkimuskonsortio edustajineen on esitetty alla olevassa taulukossa. Jatkossa konsortiota voidaan laajentaa uusilla organisaatioilla.

KONSORTION JÄSEN	EDUSTAJA	VARAEDUSTAJA
AAC Global Oy	Janne Nevasuo	
AlmaMedia Oyj	Hannele Vihermaa	Marko Turpeinen
TietoEnator Oyj	Timo Pehkoranta	Jukka Junttila
Connexor Oy	Pasi Tapanainen	Sirkku Paajanen
Leiki Oy	Petrus Pennanen	Sami Linnavuo
M-Cult ry	Minna Tarkka	
Museovirasto	Sirkka Valanto	Vesa Hongisto
Kiasma	Perttu Rastas	Jyrki Simovaara
Suomen valokuvatait. museo	Asko Mäkelä	Anne Isomursu
Tekniikan Sanastokeskus ry	Lena Jolkkonen	Katri Seppälä
Valtion taidemuseo / Kehys	Riikka Haapalainen	Marjatta Levanto
HY/Tietojenkäs. tieteen laitos	Eero Hyvönen	Tomi Kauppinen
HY/Yleisen kielitieteen laitos	Lauri Carlson	Kimmo Koskenniemi
Kansalliskirjasto	Juha Hakala	Jani Stenvall
Tampereen yliopisto	Kalervo Järvelin	Jaana Kekäläinen
Tekes	Marko Heikkinen	Matti Sihto

Projektin vastuullinen johtaja on prof. Eero Hyvönen ja sen koordinoinnista vastaa Helsingin yliopiston tietojenkäsittelytieteen laitoksella ja HIIT:ssä toimiva Semanttisen laskennan tutkimusryhmä. Helsingin yliopiston kielitieteen laitoksen osuuden vastuullinen johtaja on prof. Lauri Carlson ja Tampereen yliopiston osuudesta vastaa prof. Kalervo Järvelin.

[1] http://www.cs.helsinki.fi/group/seco/museosuomi/

1. Lähtökohta: Miksi hanke on tärkeä?

Loppukäyttäjä kommunikoi tietojärjestelmän kanssa mieluiten hänelle itselleen tuttujen käsitteiden ja sanojen avulla – ei koneen ehdoilla. Jotta käyttäjäkeskeinen vuorovaikutus olisi mahdollista, on koneen tunnettava ihmiskäyttäjänsä käsitteistö ja sanasto sekä yhdistettävä omat sisäiset esitysmuotonsa siihen.

Esimerkiksi paikkatietosovelluksessa koneläheiset x-y–koordinaatit eivät aina ole ihmiselle luonteva tapa etsiä tietoa. Vaalitulospalvelussa käyttäjä jäsentää maailmaa luontevimmin ”vaalipiireinä”, ”kaupunkeina” jne. Ontologian avulla voidaan kuvata tarvittavat käsitteet ja näiden ilmentymien ominaisuudet, kuten tietyn kaupungin erikieliset nimet (Espoo, Esbo), keskinäiset sisältyvyyssuhteet (Espoo on osa Uudenmaan vaalipiiriä ja Etelä-Suomen lääniä), hallinnolliset suhteet (Kauniainen on itsenäinen kaupunki, vaikka sijaitsee Espoon ”sisällä”) jne.

Osa käyttäjän tarvitsemista ontologioista, esimerkiksi edellä mainittu paikkaontologia, on luonteeltaan yleistä kansallista tietämystä määritteleviä *yleisontologioita*, osa taas yritys- ja sovelluskohtaista erityistietämystä esittäviä *erityisontologioita*. Sovelluksissa joudutaan yleensä samanaikaisesti hyödyntämään sekä yleisiä että eritysalakohtaisia käsitteitä ja termejä.

Suomalaisia koneluettavia ontologioita ei tällä hetkellä ole juurikaan käytettävissä. Sen sijaan näiden ”esiasteita”, *asiasanatesauruksia* (Foskett, 1980), käytetään hyvin yleisesti niin yritysmaailmassa kuin julkishallinnon piirissä. Tietokantoihin talletettu tieto on perinteisesti indeksoitu luokituksilla ja asiasanotuksen avulla, mikä mahdollistaa tietojen myöhemmän haun sisällön perusteella. Käytetyin suomalainen asiasanatesaurus on Yleinen Suomalainen Asiasanasto (YSA), jota ylläpitää maassamme Kansalliskirjasto. Suuri joukko itsenäisiä toimijoita kehittää itsenäisesti omia sanastojaan yleensä YSA:a hyödyntäen, esimerkiksi merkkaamalla sanastossa YSA:ssa olevat vastaavat käsitteet. Asiasanastoja kehittävät ja ylläpitävät mm. mediayritykset, Museovirasto ja museot, kirjastot, eduskunta, puolustusvoimat, terveydenhuoltoala ja Stakes, erityisalojen tutkimuslaitokset jne.

Menestyksekkäästä soveltamisesta huolimatta nykyisiin asiasanoitusmenetelmiin liittyy vakavia ongelmia.

1. *Yhteentoimivuuden ongelma*. Eri organisaatioiden ja henkilöiden käyttämät terminologiat eivät ole yhteensopivia. Esimerkiksi eri tieteenalat ja sovellusalat käyttävät toisistaan poikkeavia sanoituksia samastakin asiasta, yksittäisillä museoilla on omia sisäisiä sanastoja, eri kielen termien välillä on semanttisia eroja jne. Semanttinen yhteensopimattomuus vie pohjan täsmälliseltä tiedonhaulta ja estää järjestelmien yhteiskäytön. Suomalaisissa semantic web -sovelluksissa yritykset, kirjastot, museot, tutkimuslaitokset ja muut toimijat tarvitsevat omien sovelluskohtaisten käsitteistöjen

ohella yleisiä suomenkielen sanastoja ja käsitteistöjä (ontologioita) koneluettavassa muodossa.

2. *Semanttinen köyhyys*. Nykyiset tesaurukset sisältävät vain hyvin primitiivisiä semanttisia suhteita, kuten ”laajempi termi”, ”suppeampi termi” ja ”rinnakkaistermi”. Tämä estää käsitteistöjen käytön syvällisempää analyysia edellyttävissä, merkitysperustaisissa sovelluksissa. Esimerkiksi edes peruserottelua yläluokka-alaluokka ja kokonaisuus-osa-suhteisiin ei perinteisissä tesauruksissa tehdä.

3. *Laajuuden hallinta*. Mikään taho ei yksin pysty eikä ole halukaskaan yksin ylläpitämään kaikkien eri alojen sanastoja tai ontologioita. Esimerkiksi geologian tai taiteen käsitteistön hallinta ei voi kuulua yleiskäsitteistöä tällä hetkellä ylläpitävälle Kansalliskirjastolle, eikä Geologian tutkimuskeskuksella tai Valtion taidemuseolla ole mahdollisuuksia laajojen yleiskielen sanastojen ylläpitoon. Työ on kyettävä hajauttamaan eri intressiryhmille ja yhdistämään tulokset WWW:n kautta sovelluksissa, joissa usein tarvitaan samanaikaisesti eri ontologioita.

4. *Muutosten hallinta*. Käsitteistö ja sen kuvaamisessa käytetyt ontologia muuttuvat ajan kuluessa. Esimerkiksi Tsekkoslovakiaa ei enää ole olemassa valtio-ontologiassa, mutta paljon aineistoa on indeksoitu tämän valtion nimellä. Tarvitaan menetelmiä muutosten hallintaan, jotta eri aikoina erilaisilla ontologioilla kuvattuja aineistoja voitaisiin hakea hyvin määritellyllä tavalla.

5. *Kuvausten tuottamisen ongelmat*. Asiasanoitusten tuottaminen sisällöille eli *annotointi* on vaikeaa, aiheuttaa kustannuksia ja sen laatu vaihtelee suuresti indeksoijasta riippuvasta. Kun sisältöjä tuottaa ja niitä indeksoi jatkossa yhä laajempi ja siksi ei-ammattimaisempi webin käyttäjäkunta, tarvitaan yhä parempia välineitä annotoinnin tukemiseksi ja automatisoimiseksi. Sisältöjen annotointien vaivaton tuottaminen on Semanttisen Webin menestyksen avainkysymyksiä.

6. *Monikielisyys*. Suomalaiset WWW-sisällöt ovat paljolti suomenkielisiä, mutta WWW on leimallisesti monikielinen ja kansainvälinen julkaisu- ja palveluntarjonnan mekanismi. Käsitteet ovat lähtökohtaisesti kieliriippumattomia abstraktioita. Ne mahdollistavat erikielisten termien väliset kuvaukset ja tätä kautta monikieliset WWW-palvelut, laajan saavutettavuuden ja käyttömukavuuden. Täsmällisesti määritellyt monikieliset ontologiat ovat perusedellytys monikielisten sisältöjen, esimerkiksi eri maissa toimivien yritysten tuoteluetteloiden, uutisten, kirjastojen tutkimusaineistojen tai museoiden kokoelmatietojen julkaisemiselle webissä siten, että käyttäjät kykenevät hakemaan niistä helposti tietoa.

7. *Käytettävyys*. Semanttiset kuvaukset antavat uusia sovellusmahdollisuuksia WWW:n käyttöliittymien kehittämiseen, esimerkiksi sisältöjen visualisointiin ja multimodaaliliittymiin, jossa yhdistyy puheen/äänen ja näyttöruudun sekä näppäinten/hiiren käyttö.

8. *Sovellettavuus*. Nykyiset sanastot eivät ole yleensä vapaasti saatavilla ja hyödynnettävissä sovelluksissa. Yleiset käsitteistöt ja tesaurukset, vaikkapa ”Suomen kunnat ja

paikat" -ontologia tai taiteilijoiden "Auktoriteettiontologia" pitäisi olla esitetty avointen standardien avulla ja olla vapaasti yritysten ja muiden toimijoiden saatavilla. Koneen ymmärrettävässä avoimessa muodossa saatavaa aineistoa voitaisiin helposti hyödyntää erityyppisissä sovelluksissa.

Nämä ongelmat ovat yhteisiä lukuisille suomalaisille yrityksille ja organisaatioille, mutta liian laajoja ratkottavaksi ilman laajemman konsortion tukea. Erityisen polttavaksi ja ajankohtaiseksi ongelmatiikan on nostanut WWW:n kehitys kohti semanttista webiä, jossa ydinkysymyksenä on nimenomaan sisältöjen semantiikan saattaminen tietojärjestelmien "ymmärtämään" muotoon ja sovellusten hyötykäyttöön. Semanttisessa tiedon täsmähaussa yksinkertaiset asiasanatesaurukset ja luokittelut eivät riitä, vaan tarvitaan näistä jalostettuja täsmällisesti määriteltyjä ontologioita. Näissä käsitteet on määritelty ei vain ihmistä vaan erityisesti tietoteknisiä sovelluksia varten koneen ymmärtämässä muodossa.

Esimerkiksi englanninkieltä varten on vapaasti saatavilla RDF(S)-muotoinen WordNet ontologia, joka perustuu alun perin jo 70-luvulla käynnistyneeseen WordNet hankkeeseen. Järjestelmä on avoin, ilmainen ja laajassa käytössä. WordNet on sittemmin siirretty monille eurooppalaisille kielille (EuroWordNet), mukaan lukien mm. eesti. On suorastaan kulttuuriskandaali, että suomenkielelle ei ole olemassa WordNetin kaltaisia laajempia ja syvällisempiä koneluettavia tesauruksia.

2. Tavoite ja tulokset

Projektin tavoitteena on kunnianhimoisesti suomalaisen semanttisen webin tarvitseman, kansallisen ontologiajärjestelmän kehittäminen. Työ konkretisoituu pilottijärjestelmän toteutushankkeeksi, jonka puitteissa teknologian tutkimusta ja kehitystyötä tehdään.

Pilotti koostuu kolmesta osasta, jotka muodostavat projektin päätulokset:

1. *Kansallinen ontologiakirjastojärjestelmä* (ontology library system) *ONKI*. ONKI on ohjelmistokokonaisuus, jonka avulla ontologiatyö voidaan hajauttaa koordinoidusti eri toimijoille WWW:n välityksellä, kehittää ontologioita ja hallita niiden versioitiin (ontology versioning) ja yhdistämiseen (ontology merging) liittyviä ongelmia (Davies et al., 2002). ONKI-järjestelmässä pyritään hyödyntämää ulkomailla ja Suomessa jo kehitettyjä avoimia työkaluja, kuten ontologiatoimittimia (esim. Protege-2000-perheen välineet), RDF-tietokantoja ja moottoreita (Jena, Sesame jne.) ja selaimia. Tarvittavaa teknologiaa on tutkittu ja kehitetty hakijaorganisaation toimesta tätä projektia edeltäneessä Tekesin Semantic Web –hankkeessa.

2. *Yleinen Suomalainen Ontologia YSO*. YSO perustuu Kansalliskirjaston ylläpitämän Yleisen Suomalaisen Asiasanaston (YSA) käsitteisiin semanttisesti analysoituna ja rikastettuna. Ontologia toteutetaan avoimilla, tietokoneen ymmärtämillä semantic web –teknologioilla.

3. *Joukko ONKI-kirjastoon sisällytettyjä kansallisia erityisontologioita.* Koska sovelluksissa joudutaan yleensä yhdistelemään yleis- ja erityisontologioita, hankkeessa kehitetään myös joukko YSO:oon liittyviä, sitä tarkentavia vertikaalisia erityisontologioita hankkeeseen osallistuvien yritysten ja organisaatioiden keskeisimpien sovellusalueiden tarpeisiin.

4. *Hyödyllisyyden demonstointi.* Ontologiateknologian hyödyllisyyttä demonstroidaan ja tuloksia arvioidaan osallistujaorganisaatioiden sovelluksissa, tuotteissa ja aineistoilla.

Projektin tuloksena ei ole vain kertaluonteinen ONKI-järjestelmän oteutus ja joukko ontologiaprotoja, vaan

> *tavoitteena on maamme eri asiantuntijaorganisaatioille hajautetun, jatkuvan WWW-perustaisen prosessin kehittäminen*.

Tällaisen mekanismin synnyttäminen on välttämätöntä, jotta suomalaisen semanttisen webin keskeisten ontologioiden kehitystyötä voidaan projektin päätyttyäkin jatkaa eteenpäin. Ajatuksena on, että hankkeeseen osallistuvat tahot, kuten Kansalliskirjasto, Museovirasto ja muut toimijat voisivat projektin kehittämän välineistön avulla ryhtyä ylläpitämään asiasanastojensa sijasta paremmin hyödynnettävissä olevia ontologioita.

3. Osatehtävät, tulokset ja aikataulu

Projekti on 2+2-vuotinen. Tämä suunnitelma kattaa ensimmäiset 2 vuotta alkaen 1.9.2003 ja päättyen 31.8.2005. Tulosten ja kokemusten perusteella laaditaan jatkosuunnitelma jälkimmäiselle 2-vuotiskaudelle.

Projektin osatehtävät (work package, WP) on esitetty alla. Jokaisen osatehtävän kohdalla on lueteltu odotettavissa olevat tulokset ja arvioitu näiden valmistumisaika projektin alusta lukien.

Ensimmäisenä tehtävänä on perehtyminen kansainvälisiin yleisontologioihin, näissä käytettyihin semantiikkoihin ja esitysmuotoihin sekä YSO-ontologiassa tarvittavan rakenteen suunnittelu tamän pohjalta (n. v. 2003 loppuun mennessä). Tämän perusteella voidaan ryhtyä porrastetusti muiden tästä selvitystyöstä riippuvien osatehtävien suunnitteluun toteuttamiseen. Osahankkeiden porrastukseen vaikuttaa myös hankkeeseen käytettävissä olevien tutkijoiden vapautuminen toisista tutkimushankkeista (osa hankkeen keskeisistä tutkijoista siirtyvät projektiin Tekesin Semantic Web –hankkeen päättyessä).

Oletusarvoisena etenemismallina eri osahankkeissa on:

Suunnitelma ja teknologiaselvitys. Osakokonaisuudesta laaditaan ensin (n. 3 kk) selvitys ja toteutussuunnitelma johtoryhmälle.

Demonstraatio. 6-12kk:n kuluessa tästä valmistuu ensimmäinen demonstraatio kunkin osatehtävän osalta erikseen. Tämän perusteella voidaan suunnitella tarvittavia lisätoimenpiteitä ja resursseja.

Tulosten arviointi. Kunkin osakokonaisuuden tulokset testataan lopuksi loppukäyttäjillä ja tulokset arvioidaan.

WP1: YSO-ontologiajärjestelmä

Semanttisen rakenteen suunnittelu YSO-ontologialle. Työssä luodaan synteesi aiemmista vastaavantyyppisistä hankkeista sekä käytettävissä olevasta valmiista termiaineistosta. Tällaisia ovat mm. YSA-tesaurus, WordNet-malli ja siitä saadut kokemukset (Fellbaum, 1998), IEEE:n Standard Upper Ontology standardointihanke (www.suo.ieee.org), muut ”upper ontogy” -hankkeet kuten CYC (www.opencyc.org) ja DMOZ (dmoz.org) ja sovellusontologioiden vaatimukset. Tavoitteena on pyrkiä modulaariseen rakenteeseen, jossa ontologian erilaiset luokkahierarkiat ja semanttiset suhteet (esimerkiksi hyponymiset ja meronymiset suhteet) muodostavat erillisiä, toisiinsa RDF(S) muodossa yhdistettäviä komponentteja. Tämä helpottaa määrittelyjen ylläpitoa ja laajan YSO-ontologian sovittamista laskennallisilta ja semanttisilta vaatimuksiltaan erilaisiin sovellustarpeisiin. Yksittäisessä sovelluksessa ei välttämättä tarvita koko YSO:aa kaikkine ominaisuuksineen, mutta voidaan toisaalta joutua laajentamaan sitä erillisontologioilla.

Tulokset: YSO-spesifikaatio (3kk). Päivitykset hankkeen edetessä tarpeen mukaan.

WP2: YSA-YSO-muunnos

YSA-tesauruksen muunnos MARC-formaatista RDF(S) –muotoon. Ontologisen rakenteen modularisointi ja semanttisten suhteiden rikastaminen valituille YSA:n osille. YSO laaditaan monikielisenä nykyisen YSA:n pohjalta suomen, ruotsin ja englannin kielellä, mikä mahdollistaa kieliriippumattoman tiedon haun. Yhteydet ja yhteensopivuus ulkomaisten semanttisten yleisontologioiden, erityisesti WordNetin ja EuroWordNetin kanssa selvitetään.

Tulokset: RDF-raakaversio ontologisointia varten (3kk). Raportointi tämän jälkeen 3kk välein.

WP3: ONKI-arkkitehtuuri ja sen toteutus

Hajautetun ontologiakirjaston (Ding, Fensel, 2002) kehittämisarkkitehtuurin (ONKI) suunnittelu ja toteutus. ONKI mahdollistaa toisaalta edellä kuvatun YSO:n modulaarisen jakamisen sekä osien yhdistelyn, toisaalta uusien vertikaalisten sovellusontologioiden liittämisen kokonaisuuteen. ONKI:n olennainen osa on ohjelmisto, jonka avulla ontologioita voidaan kehittää ja hallita hajautetusti webissä.

Tulokset: Ensimmäinen spesifikaatio (6kk), demo (12kk). Raportointi tämän jälkeen 3kk välein.

WP4: Versionhallintamekanismi

Hankkeessa kehitetään malli ja ohjelmisto ontologioiden muutosten hallintaa varten. Keskeisenä ongelmana on, miten muutokset ontologioissa, esimerkiksi käsitteen lisääminen, poistaminen tai siirtäminen ontologiahaarasta toiseen, vaikuttavat aiemman ontologian perusteella annotoidun tiedon hakuun. Aihepiiriä on tutkittu aiemmin erityisesti USA:ssa Marylandin yliopistossa (Heflin, 2001)

Tulokset: Ensimmäinen spesifikaatio (6kk), demo (12kk). Raportointi tämän jälkeen 3kk välein.

WP5: Epätäsmällisyyden hallintamekanismi

Ontologiat ovat yleensä loogisia diskreettejä malleja, mutta niiden avulla kuvattu maailma ja tietomme siitä monin tavoin epätäsmällistä. Hankkeen yhtenä teknologisena haasteena on laajentaa perinteistä ontologiatekniikkaa sumeiden ja muilla tavoin epätäsmällisten käsitteiden (Russell, Norvig, 2002) ja tiedon määrittelyä varten. Esimerkiksi sekoitteiset kangaslaadut kuuluvat usein samanaikaisesti sekä ”luonnonkuituihin” että ”keinokuituihin”, ”järven” ja ”lammen” ero on veteen piirretty viiva jne. Epätäsmällisyyden ja sumeuden esittämistä ja hallintaa ei vielä ole kovinkaan paljoa tutkittu semanttisen webin ontologioiden yhteydessä, muissa yhteyksissä runsaastikin.

Tulokset: Ensimmäinen spesifikaatio (6kk), demo (12kk). Raportointi tämän jälkeen 3kk välein.

WP6: Erityisontologiat

Edellä esitelty tutkimus- ja kehitystyö tehdään seuraavien case-ontologioiden pilotointiin liittyen:

- *Media-alan ontologia*. AlmaMedia Oy:n tietokannan ja aineiston pohjalta.

- *Taiteen ja mediakulttuurin ontologia*. Ontologioita kehitetään Valtion taidemuseon käyttämän ICONCLASS järjestelmän, Gettyn säätiön AAT:n, YSA:n sekä m-cultin, Valokuvataiteen museon ja Kiasman omien, kehitteillä olevien sanastojen pohjalta ja niitä tuottaen yhteistyössä kunkin yhteistyökumppanin kanssa. Kohteena ovat erityisesti valokuvan, mediataiteen sekä mediakulttuuri- ja sisältötuotannon uudet lajityypit, joiden kuvaus edellyttää taiteen, tieteen ja teknologian ontologioiden monialaista yhdistämistä. Kehitettävillä ontologioilla on kiinnostavia yhtymäkohtia edellisen kohdan media-alan ontologiatyöhön ja liiketoiminnalliseen sovellukseen.

- *Museoalan ontologia*. Ontologia kehitetään Museoviraston ylläpitämän MASA:n, Gettyn säätiön AAT:n ja Ouline of Cultural Material -järjestelmän pohjalta. Myös tämä ontologia liittyy osiltaan edellisiin ja tarjoaa ONKI-teknologian kehittämisen kannalta hyvän case-haasteen.

Tulokset: Asteittain laajentuvat ontologiat. YSO-arkkitehtuurin suunnitelman jälkeen. Suunnitelmat ja ensimmäiset kokeilut (6kk). Raportointi tämän jälkeen 3kk välein.

WP7: Suomenkielen ontologinen tiedonhaku

Kehitettyjen ontologioiden hyödyllisyyttä tiedonhaussa testataan sekä tietokantakyselyissä että WWW-perustaisessa haussa.

Käsiteperustainen tiedonhaku tietokannasta

Tampereen yliopisto on aiemmissa tutkimus- ja kehityshankkeissa toteuttanut tesaurusperustaisen kyselymekanismin ja testiympäristön tiedonhakuun terminlavennustekniikan avulla. Tämä työhön liittyvää tukimusta on palkittu sekä Suomessa (Tietojenkäsittelytieteen seuran väitöskirjapalkinto) että ulkomailla (mm. kutsuesitelmä arvostetussa IJCAI-01 konferenssissa) (Kekäläinen 1999; Järvelin et al., 2000). Soveltavaa kehitystyötä on tehty mm. TietoEnator Oyj:n TRIP-järjestelmään ja AlmaMedian mediatietokantaan liittyen. Johdonmukainen jatkokehityksen kohde on siirtyminen tesaurusperusteisuudesta (esim. YSA) hyödyntämään semanttisesti rikkaampaa ontologista mallia (esim. YSO). Toinen kehityskohde on sovittaa järjestelmä yhteensopivaksi semantic web –teknologioden kanssa (RDF, RDS(S), OWL, jne,), mikä mahdollistasi menetelmien hyödyntämistä WWW-ympäristössä.

Tutkimushypoteesina on, että ontologian ja käyttäjän antaman suomenkielisen kyselyn avulla voidaan konstruoida ulkoiselle hakukoneelle (esim. Google) tarkkuuden ja saannin suhteen ”parempaan” vastaukseen johtava kysely. Paremmuus perustuu siihen, että ontologian ja kieliteknologian avulla voidaan toisaalta huomioida suomen taivutusmuotoihin liittyviä ongelmia, toisaalta ottaa haussa huomioon hakutermeihin semanttisesti liittyvät toiset termit, kuten synonyymit, antonyymit, hyponyymit, meronyymit jne. paremmin kuin nykyisessä tesaurusperustaisessa menetelmässä.

Semanttinen haku annotoidusta RDF(S)-materiaalista

Hankkeessa kehitetään semanttinen, suomenkielen morfologian ja ontologisia suhteita huomioon ottava indeksointimekanismi ja hakukone OSUMA. Tavoitteena ei ole Googlen kaltainen yleishakukone, vaan sellaisen mekanismin kehittäminen, jolla tietyn ontologian mukaisesti annotoitu, rajoitettu RDF(S)-perustainen www-aineisto voidaan indeksoida ja kohdistaa siihen ontologiaperustaisia hakuja. Tällaisia järjestelmiä tarvitaan esimerkiksi yritysten intraneteissä olevissa knowledge management –sovelluksissa ja luotaessa sisäisellä hakumekanismilla varustettuja semanttisia portaaleja internetiin. Googlen kaltaiset yleishakukoneet eivät tehtävään sovellu suomenkielen morfologiasta ja ontologioiden käytöstä johtuen.

OSUMA:a kehitetään ja testataan Tampereen yliopiston aineistojen ohella mm. Helsingin yliopiston ja HIIT:n semantic web –sovellusaineistoilla, kuten ”Finnish Museums on the Semantic Web” -portaaliprotolla, joka sisältää eri museoiden kokoelmatietoja ontologisesti annotoituna. OSUMA:n kehitys on jatkoa Helsingin yliopistossa oppilastyönä tehtyyn pieneen esiprotoon, jossa on testattu Connexor Oy:n morfologista ja syntaktista

suomenkielen analysaattoria ja indeksoijaa RDF(S)-muotoisen data-aineiston indeksointiin.

Tulokset: Tutustuminen case-järjestelmään ja aineistoihin. Suunnitelmat (6 kk) toteutustyöstä ja tarkemmasta aikatauluista. Ohjelmistodemo (18kk), testaus ja tulosten arviointi (24kk).

WP8: Monikielisyystuki

YSO-ontologian esitysmuodolle kehitetään RDF(S)-perustainen, ulkoinen kuvaustapa, jolla ontologian kieliriippumattomiin käsitteisiin (luokat ja ominaisuudet) voidaan liittää erikielisiä termejä. Tämä mahdollistaa monikielisyystuen ja modulaarisen kehitystyön eri kielten osalta.

Monikielisyystukea demonstroidaan kehittämällä monikielinen terminologinen käännösportaali WWW:iin. Se yleistää nykyisen, YSA:aan perustuvan VESA-asiasanaportaalin.

Tulokset: RDFS-spesifikaatiot erikielisten sanastojen liittämiseksi ontologiaan (3 kk). YSO:n ja erillisontologioiden RDF-termikuvaukset em. spesifikaatioiden mukaisesti (12kk ja 24 kk). Selvitys eri kielten välisten kuvausten ontologisista ongelmista ja ratkaisumallin kehittäminen (12 kk ja 24 kk).

WP9: Teknologian ja standardointityön seuranta

Tutkimus- ja standardointityö semantic web –alueella etenee nopeasti ja edellyttää valpasta kehityksen seurantaa sekä osallistumista kansainvälisiin verkostoihin. Projektia vetävä Semantic Computing -tutkimusryhmä on jäsen mm. Eurooppalaisen OntoWeb-verkoston sisällön esittämisen standardeja selvittävässä työryhmässä ja on osana Helsingin yliopiston tietojenkäsittelytieteen laitosta mukana W3C:n työssä. Kansalliskirjaston kautta ollaan mukana kirjastoalan kansainvälisessä yhteistyössä (mm. Dublin Core –standardi). Yhteistyötä www-palveluihin liittyvästä tutkimuksesta on viereillä Carnegie-Mellonin yliopiston ja Innsbruckin yliopiston kanssa ja laitoksella ollaan mukana valmistelemassa kahta aihepiiriin liittyvää EU-projektia.

Tulokset: Raportti ontologiastandardeista (3kk), päivitysraportit (12kk) ja (24kk)

WP10: Projektin johtaminen, teknologian siirto ja hallinto

Hankkeeseen kuuluu useita, eri tieteenaloja ja sovellusalueita edustavia tahoja niin yritysmaailmasta kuin julkishallinnostakin. Tulosten siirtämiseksi eri tahojen hyötykäyttöön ja ylläpidettäviksi projektissa tarvitaan erityisiä teknologian siirtoon liittyviä toimenpiteitä, kuten tiedotus- ja koulutustilaisuuksien järjestämistä. Konsortioita ja rahoituspohjaa on tarkoitus laajentaa hankkeen kuluessa, mihin tarvitaan myös resursseja. Alustavia keskusteluja on tarkoitus jatkaa mm. Opetusministeriön ja Sitran kanssa.

Tulokset: Johtoryhmälle raportointi n. neljännesvuosittain, jatkohakemukset ja Tekes-raportointi (12kk ja 24kk). Sisällölliset tilaisuudet konsortio-organisaatioiden henkilöstölle.

4. Miten tuloksiin päästään: sovellettava teknologia

Hanke rakentuu ONKI-pilottijärjestelmän kehitystyölle, jonka yhteydessä tutkitaan ja kehitetään ratkaisuja ja ohjelmistoja. Kehitettävät ohjelmistot, kuten sumea ontologiatekniikka ja versionhallinta ovat kuitenkin yleisemminkin sovellettavissa. Teknologiapohjan muodostaa oletusarvoisesti semantic web -standardit, suositukset ja työkalut kuten RDF(S), OWL, Jena, RQL, RDQL, Sesame, jne. Ontologioiden osalta pyritään hyödyntämään mahdollisimman paljon jo olemassa olevia sanastoja ja ontologioita (esimerkiksi WordNet).

Järjestelmä testataan toteuttamalla sillä yleisontologia YSO, liittämällä tähän caseina olevia erityisontologioita ja soveltamalla näitä TRIP-järjestelmään ja AlmaMedian tietokantaan sekä muihin konsortiopartnereiden järjestelmiin. Suuri osa jälkimmäisestä työstä tehdään osallistujien toimesta yritysten/organisaatioiden omissa hankkeissa. Ydinhankkeessa kehitettyjen ontologioiden hyödyntämistä tiedonhaussa demonstroidaan myös HIIT:n ja HY:n Semantic Web –projektissa valmistuvilla tiedonhakujärjestelmillä (Java-perustainen kuvatietokantaselain, WWW-perustainen Finnish Museums on the Semantic Web -järjestelmä).

5. Hyödyntämissuunnitelma

Tesauruksiin perustuvan asiasanoituksen käyttö on keskeisimpiä tapoja hallita tiedonhaun ja sisällöllisen yhteentoimivuuden ongelmia kolmelta kannalta:

1. *Tiedon loppukäyttäjälle* se tarjoaa sanaston ja yksinkertaiseen semantiikan, jolla tietoa voidaan tietokannoista ja Webistä hakea.

2. *Tiedon tuottajalle* se tarjoaa kehyksen, jonka avulla sisältöjä voidaan kuvata eli *annotoida* rutiininomaisesti ja tasalaatuisesti.

3. *Järjestelmien kehittäjille* avoimet semantic web -esitysmuodot sekä standardit ontologiat tarjoavat teknologisen perustan tiedonhakujärjestelmien kehittämiseen webiin. Lisäksi ne tarjoavat perustan hajautettujen järjestelmien yhteiskäytölle verkossa, esimerkiksi museo-, terveydenhuollon- tai kirjastojärjestelmien yhdistämiselle.

Konsortion jäsenten hyödyt

Kaikki konsortion jäsenet käyttävät omassa toiminnassaan asiasanoitukseen perustuvaa indeksointia ja tiedonhakua tai kehittävät sitä varten ohjelmistotuotteita. Jokaisella osallistuvalla sisällöntuottajataholla on laadittu omia asiasanastoja ja törmätty käytännön työssä hakemuksessa esitettyihin tesaurusten käytön ongelmiin. ONKI-järjestelmän avulla osallistujat toivovat voivansa ryhtyä kehittämään käsitteistöjään ja sanastojaan tavalla, joka olisi suoraan hyödynnettävissä tietoteknisissä sovelluksissa ja toisi lisäarvoa myytäviin ohjelmistotuotteisiin.

Media-aineistojen sähköinen hyödyntäminen on ydinliiketoimintaa AlmaMedia Oyj:n kaltaiselle mediayritykselle yhä voimakkaammin verkottumassa Internet-maailmassa. Semanttisen webin mukana tulevat indeksointimenetelmät ja kielet ja niihin perustuvat web-perustaiset hakumenetelmät ovat jatkossa muodostumassa alan standardiksi. Niiden tukeminen ja hyödyntäminen alan ohjelmistotuotteissa, kuten TietoEnatorin TRIP-tietokantajärjestelmässä on välttämätöntä kilpailuedun säilyttäminen kannalta.

Taideala on yksi keskeinen sisällöntuottaja tulevassa semanttisessa webissä. Alueella on sekä kulttuuri- teknologia- että elinkeinopoliittinen merkitys. Hankkeessa kehitettävät uusien digitaalisten sisältötuotteiden ontologiat ja ohjelmistotyökalut hyödyntävät sekä kulttuurialan instituutioita, jotka palvelevat laajoja käyttäjäryhmiä (yleisöt, yhteisöt, yritykset) että kotimaisia uusmedia- ja sisältöteollisuuden toimialoja, joiden yhä vakiintumaton sanasto ja luokitukset ovat parhaillaan eurooppalaisten harmonisointihankkeiden kohteina. Semanttisten standardien luominen alalle edistää myös alan kansainvälistymistä.

Museot kuuluvat arkistolaitoksen ja kirjastojen ohella ns. muistiorganisaatioihin, joiden keskeisenä tehtävä on kansallisen tiedon tallettaminen. Aineistojen sähköinen julkaiseminen webissä muodostaa uuden kanavan yleisön, tutkijoiden ja yritysasiakkaiden palvelemiselle, mutta on toisaalta merkittävä kustannuserä museoiden budjeteissa. Esimerkiksi Kansallismuseossa on kokoelmien digitointiin arvioitu tarvittavan n. 500 henkilötyövuotta ja joka vuosi kokoelmat karttuvat entisestään. Tähän jättiurakkaan sisältyvä asiasanoitus olisi syytä tehdä kerralla mahdollisimman ”oikein”, jotta tiedonhaku – keskeinen syy koko urakalle – ja aineiston julkaiseminen webissä onnistuisivat mahdollisimman hyvin. Hankkeeseen sisältyvä ontologiakehitys on johdonmukaista jatkoa nykyisille asiasanatesauruksille ja on erittäin keskeisessä roolissa museosisältöjen esittämistapojen harmonisoinnissa ja tietoteknisessä hyödyntämisessä. Esimerkiksi museoalan kansainvälisen CIDOC-kattojärjestön CRM-ontologiasta luettelointia varten on valmistumassa ISO standardi.

Tekniikan Sanastokeskus ry (www.tsk.fi) on maamme johtava terminologian asiantuntija. Sanastoja on perinteisesti kehitetty normatiivisesti ihmisen käyttöön, mutta semantic webin ontologiatekniikat ovat mullistamassa tilannetta. Ontologioiden avulla määrittelyt voidaan tehdä riittävän formaalilla tavalla, siirtää koneen ymmärtämään muotoon ja ottaa käyttöön yritysten tietojärjestelmissä ja julkishallinnon sovelluksissa. Systemaattisten käsitteistön ja sanastojen käyttö on välttämätöntä esimerkiksi yritysten välissä sähköisessä kaupankäynnissä. Esimerkiksi elektroniikka- ja puolijohdeteollisuuden RosettaNet standardin yksi keskeinen osa on ontologiset sanastot (RosettaNet Dictionaries). Kuvaavaa on, että seuraavassa pohjoismaisessa terminologiakonferenssissa Nordterm 2003 semanttinen web on valittu tilaisuuden yhdeksi teemaksi.

Monikielisyyteen liittyvät ongelmat ja toisaalta niihin liittyvät liiketoiminnalliset mahdollisuudet ovat maapalloistumisen myötä arkea yhä useammassa suomalaisyrityksessä. Aihepiiriin liittyviä sovellusalueita ovat mm. yritysten monikielisen viestinnän terminologiapalvelut, kääntäminen sekä tuotteiden ja palveluiden lokalisointi.

Kansallinen hyöty

On kansantaloudellista resurssien tuhlausta, jos WWW:n sisällöntuottajayritykset ja julkishallinnon organisaatiot ryhtyvät laatimaan kansallisista yleisontologioista omia kilpailevia versioitaan. Väistämättömästi Suomessakin edessä olevan ontologiaurakan yhteinen osa kannattaa tehdä Suomessa vain kerran eri tahojen välisenä yhteistyönä ja säästää resursseja yleisosaa tarkentavien, sovelluskohtaisten erityisontologioiden kehittämiseen. Moninkertainen ontologisointi johtaa myös tilanteeseen, jossa eri systeemejä käyttävät järjestelmät eivät ole keskenään semanttisesti yhteentoimivia. Vaarana on suomalaisen semanttisen webin fragmentoituminen eri standardeja kättäviin saarekkeisiin.

Siksi on tarkoituksenmukaista laatia julkisella rahoituksella kaikille toimijoille avoimia suomenkielen yleisontologioita kansantaloudellisesti tärkeimmillä alueilla, kehittää ja toteuttaa tekninen mekanismi, jolla eri aloilla tehtävät sovelluskohtaiset ontologiat voidaan siihen yhdistää. Näin luodaan teknologinen perusta sille, että eri aihepiirien sisällöntuottajat voisivat jatkossa kehittää oman erityisalueensa ontologiaa ja sanastoja koordinoidusti muiden tahojen kanssa. Täydellinen yhteensopivuus ja koordinointi on aihepiirin laajuudesta johtuen tietysti mahdotonta, mutta tavoitteeseen pitää pyrkiä aktiivisesti, sillä osittainenkin onnistuminen joillain keskeisellä alueella voi olla merkittävä edistysaskel. Yksittäiselle toimijalle näin laaja ontologiahanke olisi taloudellisesti ja teknisesti liian raskas toteuttaa.

Jotta työn taloudellinen hyöty maksimoituisi, yleisontologioiden tulisi olla ”kansallisomaisuutta” ja kaikkien vapaasti käytettävissä. Näin taloudellinen hyöty jakaantuu suurelle joukolle suomalaisia WWW:n sisällöntuottajia, yrityksiä, muistiorganisaatiota, yliopistoja ja muita toimijoita.

Lisensiointimalli

Hankkeen ohjelmistot ja ontologiat toteutetaan pääosin seuraavan Massachusetts Institute of Technologyn IT-lisenssimallin puitteissa, joka antaa hankkeeseen osallistuville ja muille hyödyntäjätahoille mahdollisuuden hyödyntää tuloksia vapaasti omissa tuotteissaan:

> **MIT Licence**
> 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.

Yrityskriittisissä osissa voidaan kuitenkin soveltaa tapauskohtaisesti suljetun koodin malleja. MIT-lisenssimalli ei pakota tulosten käyttäjiä julkaisemaan omia uusia koodejaan avoimena, mikä rajoittaisi kaupallista hyödynnettävyyttä. Lähteen ilmoittaminen riittää. Koodi ei näin ”saastu”, kuten esimerkiksi GPL-lisenssissä. Koodit ja ontologiat voidaan kuitenkin julkaista myös GPL-tyyppisesti MIT-lisenssin ohella, mikäli tämä on joltain osin tarkoituksenmukaista. Koodista voi syntyä sekä yksityisesti ja julkisesti ylläpidettävät haarat.

Viitteet

Davies J., Fensel D., Harmelen F. (eds), Towards the Semantic Web. Wiley, 2003.

Ding Y., Fensel D., Ontology library systems: The key to successful ontology re-use. Free University of Amsterdam, 2002. URL: citeseer.nj.nec.com/455227.html

Fellbaum C. (ed.), WordNet: An Electronic Lexical Database. MIT Press, 1998.

Fensel D., Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, 2001.

Foskett D., Thesaurus. In: Encyclopaedia of Library and Information Science, Vol. 30, Marcel Dekker, 1980.

Heflin J., Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment. PhD Thesis, University of Maryland, 2001.

Hyvönen E. (ed.), Semantic Web Kick-Off in Finland. HIIT Publications 2002-01, 2002.

Hyvönen E., Klemettinen M. (eds), Towards the Semantic Web and Web Services. Proceedings of XML Finland 2002. HIIT Publications 2002-03, 2002.

Järvelin K., Kekäläinen J., IR evaluation methods for retrieveing highly relevant documents. Proceedings of ACM SIGIR’00, ACM Press, New York, 2000 (Receiver of “Best Paper Award”).

Kekäläinen J., The Effects of Query Complexity, Expansion, and Structure on Retrieval Performance in Probabilistic Text Retrieval. PhD Thesis. Acta Universitatis Tamperensis 678, Tampereen yliopisto, 1999. (Tietojenkäsittelytieteen seuran vuoden väitöskirjapalkinto).

Russel S., Norvig P., Artificial Intelligence. A Modern Approach. Prentice-Hall, 2002.