TEKNILLINEN KORKEAKOULU
Viestintätekniikka

8.10.2005

Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita?

Eero Hyvönen
TKK Viestintätekniikka ja Helsingin yliopisto, tietojenkäsittelytieteen laitos
PL 5500, 02015 TKK
eero.hyvonen@tkk.fi
http://www.tkk.fi/~eahyvone

Asiasanastoja käytetään yleisesti tiedon kuvailussa kirjastoissa, museoissa, arkistoissa, mediayhtiöissä jne. Kontrolloidun sanaston käytöllä voidaan parantaa ja yhdenmukaistaa tiedon indeksointia, jolloin tiedon haussa päästään myöhemmin parempaa tarkkuuteen (precision) ja saantiin (recall). Asiasanastoja on organisoitu perinteisesti merkitysten perusteella tesauruksiksi, jolloin termeihin merkityksen perusteella liittyvät toiset termit voidaan löytää helpommin kuin esimerkiksi aakkosellisesta hakemistosta. Viime aikoina monia asiasanastoja on alettu kehittää ns. ontologioiksi (Fensel, 2003; Staab, Studer, 2003). Tässä artikkelissa kerrotaan miksi ja milloin ontologian kehittämistä kannattaisi harkita tesauruksen sijaan ja esitetään, miten perinteinen tesaurus voidaan muuttaa ontologiaksi. Esimerkkinä käytetään Yleistä suomalaista asiasanastoa (YSA), jota ollaan parhaillaan ontologisoimassa kansallisessa Suomalaiset semanttisen webin ontologiat -hankkeessa (FinnONTO)[1] (Hyvönen et al., 2005a).

1 Koneet asiasanastojen käyttäjiksi

Tesaurusta voidaan käyttää indeksoinnin (asiasanoituksen, luokituksen) apuvälineenä, tiedon haun apuvälineenä tai molemmissa tehtävissä. Tämä näkökulma on keskeinen informaatiotutkimuksen alueella ja kirjastotieteissä. Tesauruksen avulla voidaan myös kehittää termien määrittelyä ja ohjata sekä harmonisoida kielen käyttöä – tämä on erityisesti terminologiatutkimuksen näkökulma sanastotyöhön (Suonuuti, 2001). Tieto- ja tietämystekniikan (knowledge engineering) näkökulmasta tarkasteltuna tesaurukset ovat eräänlaisia ontologisia kuvauksia maailmasta ja niitä voidaan potentiaalisesti käyttää hyväksi tiedon haussa ja päättelyssä monin eri tavoin.

Aiemmin indeksoinnissa käytettiin painettuja sanastoja, mutta nykyisin yleensä näiden sähköisiä versioita. Esimerkiksi YSA julkaistaan nykyisin ajantasaisena vain verkkoversiona[2]. Myös tiedon haussa sanastojen ja luokitusten tietokoneperustainen hyödyntäminen on yleistynyt. Esimerkiksi Googlesta on käytössä Google Suggest[3] -versio, joka osaa täydentää käyttäjän kirjoituksen mielekkäiksi hakusanoiksi, mikä mm. estää kirjoitusvirheitä. Googlen hakemisto-

[1] http://www.cs.helsinki.fi/group/seco/ontologies/
[2] http://vesa.lib.helsinki.fi
[3] http://www.google.com/webhp?complete=1&hl=en

osa taas hyödyntää Open Directory Project -hankkeen[4] (ODP) 590 000 kategorian luokittelua. ODP-hankkeessa ollaan indeksoimassa n. 70 000 vapaaehtoisen voimin koko webin sisältöä. Tällä hetkellä indeksoituna on yli 5 mijoonaa sivua. Järjestelmä perustuu semanttisen webin avoimeen RDF(S)-formaattiin, joka on yksinkertainen kieli ontologioiden ja niihin liittyvän metadatan esittämiseksi. Tyypillinen tesauruksen hyödyntämistapa on termilaajennus, jonka avulla voidaan automaattisesti parantaa haun saantia. Esimerkiksi hakusana Pohjoismaat voidaan tesauruksen avulla laventaa hakusanoiksi Suomi, Ruotsi jne., jolloin myös nämä sanat sisältävät dokumentit löytyvät. Termilaajennus on usein tarpeen, vaikka tällöin haun tarkkuus usein vastaavasti yleensä heikkeneekin. Edellisessä esimerkissä yksittäistä maata käsittelevät dokumentit eivät esimerkiksi välttämättä vastaa tietotarpeeseeni koskien Pohjoismaita kokonaisuutena.

Jotta sanastojen käytöstä saataisiin maksimaalinen hyöty, pitää tietokoneiden tavalla tai toisella ”ymmärtää” tesauruksessa käytettyjä semanttisia suhteita ja merkityksiä, esimerkiksi että Suomi kuuluu Pohjoismaihin. Ontologiatekniikoiden tavoitteena on nimenomaan tehdä sanastoista tietokoneelle ymmärrettävämpiä.

Konesemantiikka

Monien nykyisten asiasanastojen ongelmana on, että niitä on kehitetty lähinnä ihmiskäyttäjän, erityisesti indeksoijan ehdoilla. Sanastojen avulla voidaan helpottaa indeksoijan työtä löytää kuvailtavaa kohdetta parhaiten luonnehtivat asiasanat. Sanastojen käyttäjäkunta indeksoinnissa ja erityisesti tiedon haussa on kuitenkin nopeasti laajentunut indeksoinnin asiantuntijoista maallikoihin ja tietokoneisiin, joilla ei ole samaa tietämystä maailmasta kuin meillä. Termit kuten ”Suomi” ovat koneille pelkkiä kasvottomia, merkityksettömiä merkkijonoja. Suomi on merkkijonona koneelle samanarvoinen kuin vaikkapa ”Zxc5b”. Koneen kannalta merkitys syntyy termien välisten suhteiden kautta, esimerkiksi siitä, että Suomi on hierarkkisessa osa-kokonaisuussuhteessa Pohjoismaihin. Suhteiden täsmällinen kuvaaminen on siksi ensiarvoisen tärkeää koneiden kannalta.

Asiasanaston ontologisoinnissa on kysymys siitä, että ihmiselle tarkoitetun sanaston käsitteistö ja suhdeverkosto täsmennetään sellaiseksi, että tietokone pystyy sitä hyödyntämään aiempaa paremmin. Tästä ei ole haittaa ihmislukijalle, jolle sanaston täsmällisempi rakenne ja merkitys on myös selkeä etu. Haittana on lisätyö ja kustannus, joka sanaston täsmällisemmästä määrittelystä seuraa. Sanaston kehittäjän arvioitavaksi jää, milloin saavutettavat edut sovelluksissa ovat suurempia kuin tarkemman määrittelyn aiheuttama lisäkustannus. Ontologisointi voidaan tehdä tarpeen mukaan joko kevyesti vain välttämättömimpiä muutoksia tehden (light-weight ontologies) tai syvällisemmin sanaston käyttötarkoituksesta riippuen. Tässä artikkelissa tarkastellaan vain hyvin kevyttä ontologisointia ja tekniseltä näkökulmalta. Ontologia-termillä tarkoitetaan tällöin teknistä formaalia, koneen tulkittavissa olevaa kuvausta niistä käsitteistä ja näiden välisistä suhteista, joita älykkäällä toimijalla tai toimijajoukolla on käytettävissään. Ontologiatutkimus syvällisemmässä mielessä muodostaa oman laajan tutkimusalueensa ja traditionsa filosofiassa ja tekoälytutkimuksessa (Sowa, 2000).

[4] http://dmoz.org

Ontologiat ja semanttinen web

Usko ontologioiden avulla saavutettavista sovelluseduista on viime vuosina kasvanut voimakkaasti. Yksi merkittävä syy tähän on ollut webin isänä tunnetun Tim Berners-Leen visio semanttisesta webistä, jonka kehittämiseksi webin infrastruktuuria kansainvälisesti kehittävä W3C-järjestestö käynnisti vuonna 2001 erityisen Semantic Web Activity -ohjelman[5]. Semanttisen webin ehkä keskeisimpänä teknologisena elementtinä ovat ontologiat. Sovellusten kesken jaettujen ontologioiden avulla voidaan parantaa tietojärjestelmien yhteentoimivuutta (interoprability), mikä on ollut jatkuvasti yksi tietotekniikan suurimpia käytännön haasteita. Semanttisen webin ontologiat ovat sanastoja, joiden avulla tietoverkkojen sisällöt (metatiedot) voidaan ilmaista koneen ymmärtämällä tavalla. Koneymmärrettävät sisällöt mahdollistavat yhteentoimivuuden ohella entistä älykkäämpien järjestelmien kehittämisen.

Esimerkki semanttisen webin ontologioiden mahdollisuuksista Suomessa on kansainvälisesti ja kotimaassa palkittu[6] MuseoSuomi-järjestelmä[7] (Hyvönen et al. 2005b). Siinä eri museoiden tietosisällöt voitiin yhdistää sisällöllisesti seitsemän yhteisen ontologian avulla yhdeksi virtuaaliseksi kokoelmaksi: käyttäjä löytää ”yhden luukun takaa” eri museoiden aineistot saumattomasti toisiinsa liitettyinä. Pilottijärjestelmässä ovat mukana Kansallismuseo, Espoon kaupunginmuseo sekä Lahden kaupunginmuseo, joiden kokoelmat käyttävät erilaisia museotietojärjestelmiä, luettelointikäytäntöjä ja sijaitsevat eri kaupungeissa. Täsmällisemmän semantiikan avulla MuseoSuomi pystyy tarjoamaan lisäksi käyttäjilleen älykkään ”semanttisen hakukoneen” sekä mahdollisuuden ”semanttiseen samoiluun”, jossa kone päättelysääntöjensä avulla pystyy suosittelemaan käyttäjälle linkkejä eri tavoin toisiinsa liittyviin tietosisältöihin. MuseoSuomessa hyödynnettiin n. 6000 termin Museoalan asiasanastoa MASA:a (Leskinen, 1997), mutta ilman sen ontologisointia Museoalan ontologiaksi MAO ei näitä järjestelmiä olisi voitu toteuttaa. Yleisradio Oy:lle tehdyssä semanttisessa Orava-portaalissa[8] MuseoSuomen tietosisällöt voitiin linkittää vielä yli 2000 videoklipin opetusaineistoon.

Tesaurusten ontologisointikysymys on tärkeä asiasanastojen laajan käytön ja webin käytön nopean yleistymisen takia. Aihe on otettu käsiteltäväksi mm. W3C:n Semantic Web Best Practices –työryhmässä[9], joka on valmistellut tähän liittyvää ohjeistusta ja SKOS-suositusta. Tässä yhteydessä on syytä korostaa, ettei pelkkä SKOS-suosituksen mukainen syntaktinen muunnos tesaurusformaatista XML- tai RDF-muotoon riitä, vaikka parantaakin tesauruksen käytön edellytyksiä web-ympäristössä. Jotta semanttista lisäarvoa syntyisi, pitää samalla yleensä täsmentää, laajentaa ja korjata tesauruksissa olevia semanttisia suhteita. Muuten kone tekee samat virhepäätelmät kuin tesaurusta käytettäessä.

Seuraavassa esitetään, millaisia muutoksia tesaurukseen käytännössä joudutaan tekemään suoritettaessa kevyttä muunnosta ontologiaksi. Esimerkkinä käytetään n. 23 000 termin YSA:a,

---

[5] http://www.w3.org/2001/SW

[6] MuseoSuomi sai v. 2004 semanttisen webin kansivälisen tutkijayhteisön Semantic Web Challenge Award –sovelluspalkinnon ja pääministerin kunniamaininna innovatiivisimmasta sovelluksesta Tietoyhteiskuntaohjelman Laaatua verkkoon –kilpailussa.

[7] MuseoSuomi on vapaasti käytettävissä osoitteessa http://www.museosuomi.fi. Pilottijärjestelmää ylläpitää Teknillisen korkeakoulun viestintätekniikan laboratorio ja Semanttisen laskennan tutkimusryhmä (SeCo).

[8] Orava-portaali on kokeiltavissa osoitteessa http://www.museosuomi.fi/orava.

[9] http://www.w3.org/2001/sw/BestPractices/

jota TKK:ssa ja Helsingin yliopistolla toimiva Semanttisen laskennan tutkimusryhmä[10] on parhaillaan ontologisoimassa Yleiseksi Suomalaiseksi Ontologiaksi YSO. Yhtä hyvin esimerkit olisi voitu valita MuseoSuomen yhteydessä käytetystä Museoalan asiasanastosta MASA tai muista samantyyppisistä suomalaisista tesauruksista, joita FinnONTO-projektin puitteissa on tutkittu ja ontologisoitu.

2 Asiasanaston rakenne

Asiasanastot esitetään yleensä ISO 2788, BB 5723 (Iso-Britannia) ja ANSI/NISO Z39.19 (USA) standardien mukaisessa muodossa. Sanasto koostuu termeistä, joiden merkitys on kuvattu joukolla suhteita toisiin termeihin. Tärkeimmät suhteet, joita käytetään mm. suomalaisissa tesauruksissa, on esitetty taulukossa 2.1.

Suhde englanniksi	**Suhde suomeksi**	**Merkitys**
BT Broader term	LT	Laajempi termi, hierarkkinen suhde, käänteissuhde ST
NT Narrower term	ST	Suppeampi termi, hierarkkinen suhde, käänteissuhde LT
RT Related term	RT	Rinnakkaistermi, assosiatiivinen suhde
USE Use	KÄYTÄ	Ilmaisee suositeltavan synonyymin, käänteisuhde KORVAA
UF Used for	KORVAA	Ilmaisee ei-suositeltavan synonyymin, käänteissuhde KÄYTÄ
SN Scope note	HUOMAUTUS	Luonnehtii termin merkitystä ja käyttöä

Taulukko 2.1. Tärkeimpiä semanttisia suhteita tesauruksissa.

Sanaston semanttiset suhteet jakaantuvat kolmeen ryhmään (Aitchison et al., 2000):

Ekvivalessisuhteet

Ekvivalenssisuhteilla ilmaistaan termien synonyymit ja näille käytettävä suositeltava termi (preferred term). Nämä tiedot kuvataan KÄYTÄ/KORVAA-suhteilla. Esimerkki YSA:sta:

agenttikirjallisuus KÄYTÄ vakoilukirjallisuus

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

Synonyymien osalta voidaan tehdä ero aidosti eri termeille ja leksikaalisille varianteilla (lexical variant), jotka ovat saman termin erilaisia ilmiasuja. Kvasisynonyymilla (quasi-synonym) tarkoitetaan lähes samaa merkitsevää termiä.

Hierarkkiset suhteet

Hierarkkisia suhteita kuvataan LT ja ST suhteilla. Tärkeimmät suhdetyypit ovat osa-kokonaisuus (part of), kuten

komeetat LT aurinkokunnat

ja ala-yläkäsite (subclass of), kuten

sairaalat LT terveydenhuoltolaitokset

Assosiatiiviset suhteet

Muita kuin ekvivalenssi ja hierarkkisia suhteita kuvataan RT-suhteella. Esimerkiksi:

sairaus RT parantuminen

3 Asiasanaston muuttaminen ontologiaksi

Tesaurus voidaan muuttaa yksinkertaiseksi ontologiaksi seuraavissa vaiheissa:

1. Syntaktinen muunnos ontologiaformaattiin
2. Terminologian erottaminen ontologiasta
3. Termien päämerkitysten erottelu
4. Luokkien ja yksilöiden erottelu
5. Hierarkkisten suhteiden erottelu
6. Hierarkkisten suhteiden systematisointi hierarkiaksi
7. Assosiatiivisten suhteiden esittäminen

Vaihe 1 on luonteeltaan syntaktinen. Vaiheet 2-3 liittyvät semanttisiin ekvivalenssisuhteisiin, vaiheet 4-6 hierarkkisten suhteiden täydentämiseen, korjaamiseen ja täsmentämiseen ja vaihe 7 assosiatiivisten suhteiden tarkempaan määrittelyyn. Seuraavassa esitellään lyhyesti nämä vaiheet ja saavutettava lopputulos.

3.1 Syntaktinen muunnos ontologiaformaattiin

Ensin asiasanasto muutetaan syntaktisesti tavoitteena olevalle ontologiakielelle, joka voi olla esimerkiksi RDF(S)[11] tai OWL[12]. Näitä kieliä varten on saatavilla erityisiä ontologiaeditoreita,

[11] http://www.w3.org/RDF/
[12] http://www.w3.org/2004/OWL/

joiden avulla muunnoksen seuraavat vaiheet on helpompi suorittaa. Käytetyin ontologiaeditori on Stanfordin yliopistossa kehitetty Pretege-2000[13] lisäohjelmineen (plug-in).

YSA:n tapauksessa sanasto oli saatavilla MARC-formaatissa[14], josta se muutettiin ensiksi XML-muotoon ja sitten RDF(S)- ja OWL-kielelle. Näin sanasto voitiin ottaa edelleen kehitettäväksi Protege-2000-editorissa.

3.2 Terminologian erottaminen ontologiasta

Tesaurukset ovat normatiivisia, termien käyttöä ohjaavia terminologioita. Ontologioissa lähtökohtana eivät niinkään ole ihmisen käyttämät termit kuin niiden alla olevat koneen ymmärtämät merkitykset. Harkittavaksi tulee, miten termit ja merkitykset esitetään ontologiassa. Tässä voidaan käyttää kahta periaatteellista tapaa.

Termit käsitteiden yhteydessä

Suoraviivainen tapa on esittää kunkin käsitteen termit aina käsitteen yhteydessä jonkin ominaisuuden arvona. Esimerkiksi RDF-kehyksessä voidaan käyttää rdfs:label-ominaisuutta ja tämän tarkentamiseen kielen ilmaisevaa xml:lang-attribuuttia. Lähestymistavan ongelmana on, että käsitteeseen voi liittyä hyvin paljon termejä ja näiden leksikaalisia variantteja. Esimerkiksi Art and Architecture Thesaurus (AAT) -tesauruksessa[15] käsitteellä voi olla termit englannin eri murremuodoille. Henkilötesauruksessa Union List of Artist Names[16] (ULAN) löytyy Ivan Aivazovskille = Ivan Aivazovsky (Valtion taidemuseon käyttämät nimet) syntaktisina variantteina seuraavat translitteroidut muodot

Ayvazovsky, Ivan (Russian painter, 1817-1900) [500021161]
Aivasovsky, Ivan Konstantinovitsch
Aivazovskii, Ivan Konstantinovich
Aivazovski, Ivan Constantinovich
Aivazovsky, Ivan Konstantinovich
Aivazowsky, Ivan Konstantinovich
Aiwasoffski, Ivan Konstantinovich
Ajvazovskij, Ivan Konstantinovic
Ajvazovskij, Ivan Konstantinovic
Alwasoffski, Ivan Konstantinovich
Ayvazovsky, Ivan Konstantinovich
Ivan Ayvazovsky
Ivan Konstantinovic Ajvazovskij

Taiteilijoilla voi lisäksi olla eri pseudonimiä eri aikoina, nimi voi muuttua avioliiton seurauksena, jokin tietty nimi voi olla henkilön itsensä käyttämä jne.

Toinen ongelma on, että samankin kielen sisällä eri käyttäjäryhmillä ja organisaatiolla saattaa olla toisistaan poikkeavia terminologisia käytäntöjä eri aikoina. Esimerkiksi vuosituhannen

[13] http://protege.stamford.edu
[14] http://www.loc.gov/marc/
[15] http://www.getty.edu/research/conducting_research/vocabularies/aat/
[16] http://www.getty.edu/research/conducting_research/vocabularies/ulan/

alussa käytetty suomenkieli poikkeaa nykyisestä ja Suomessa toiset museot käyttävät yksikkömuotoisia asiasanoja, toiset taas monikkomuotoa.

Termit erillisinä termiontologioina

Terminologiset määrittelyt voidaan erottaa itse ontologiasta mutkikkaiden terminologisten määrittelyiden toteuttamiseksi ja käyttäjäryhmäkohtaisten terminologioiden luomiseksi. Lähestymistapa on luonteva esimerkiksi monikielisten ontologioiden toteuttamiseksi. Tätä lähestymistapa kehitettiin ja kokeiltiin MuseoSuomi-hankkeen yhteydessä. Siinä jokainen semanttiseen portaaliin aineistoa tuottanut museo saattoi käyttää omaa terminologiaansa ilmoittamalla kullekin termille sitä vastaavan ontologisen käsitteen yhteisissä ontologioissa.

Menetelmän käytännöllisenä hankaluutena on, että nykyisten ontologiaeditorien ja ontologian toimittajan kannalta on usein luontevampaa ylläpitää asiasanastoja käsitteiden yhteen kerättyinä ominaisuuksina eikä erillisinä termiontologioina. Ongelmaa voidaan lähestyä kehittämällä erityisesti asiasanastotyötä tukevia ontologiaeditoreita.

3.3 Termien päämerkitysten erottelu

Sanat ovat monimerkityksisiä. Tesauruksia ja ontologioita luotaessa keskeinen tehtävä on eri merkitysten erottaminen termeiksi ja käsitteiksi. Ilmeinen ongelma ovat homonyymiset termit, joissa sama sana tarkoittaa kahta aivan eri käsitettä. Esimerkiksi:

Nokia (paikkakunta)
Nokia (yritys)

Polyseemisillä termeillä taas on useita toisiinsa liittyviä merkityksiä. Esimerkiksi sanalla ”johtaminen” on mm. seuraavia merkityksiä, joista kaksi ensimmäistä liittyvät toisiinsa melko läheisesti:

johtaminen (musiikki)
johtaminen (liiketalous)
johtaminen (sähkötekniikka)
johtaminen (matematiikka)

Johtaminen-hakusanan käyttö tiedonhaussa johtaa (tässä jälleen yksi johtaminen-sanan merkitys) siksi huonoon tarkkuuteen erityisesti silloin, kun haku kohdistuu samanaikaisesti eri alojen tietosisältöihin. Kun tesaurus ontologisoidaan, tulee termien keskeisimmät eri merkitykset eri alueilla erottaa omiksi termeiksi (käsitteiksi) ja erikseen näiden suhteet toisiin termeihin erikseen.

YSA:ssa termillä johtaminen on 20 alakäsitettä, joilla tarkoitetaan lähinnä erilaisia johtamismenetelmiä, kuten arvojohtaminen, henkilöstöjohtaminen jne. Merkitys johtaminen (musiikki) taas vastaa seuraavia termejä:

orkesterit – johtaminen,
yhtyeet – johtaminen
kuorot – johtaminen.

Asiasanan ”musiikin johtaminen” käyttö on erityisesti kielletty. Merkitystä johtaminen (sähkötekniikka) ja johtaminen (matematiikka) ei YSA:ssa ole terminä. Näillä termeillä ei ole eksplisiittisiä ST/LT/RT-suhteita, vaan semanttinen yhteys johtamiseen on ilmaistu implisiittisesti termin nimessä päätteellä ” – johtaminen”. Indeksointia suorittavalle ihmiselle merkitys selviää nopeasti termin nimen perusteella, mutta kone joutuu tässä isomman haasteen eteen yrittäessään ymmärtää termin merkitystä esimerkiksi tiedonhaun yhteydessä. Koneen kannalta merkitys tulisi luonnollisen kielen sijasta esittää eksplisiittisesti semanttisten suhteiden avulla. Kone ei esimerkiksi voi älytä, missä johtaminen-sanan merkityksessä liite ”—johtaminen” terminimessä ”orkesterit – johtaminen” esiintyy, vaikka ihminen tietämyksensä avulla tulkinnan pystyykin tekemään.

Ontologisessa mielessä orkesterin, yhtyeen ja liikkeen johtamisen voisi asettaa johtamiskäsitteen alakäsitteiksi, koska niissä on kysymys jonkinlaisen organisaation johtamisesta. Sähkön ja kaavojen johtaminen taas on merkitykseltään erilaista. Eri merkityksille pitää lähtökohtaisesti luoda omat käsitteet ja määritellä omat semanttiset suhteet.

Asiasanaston ontologisoinnin yhteydessä jokaiselle käsitteelle annetaan yksilöivä tunniste. Web-maailmassa tunnisteina käytetään yleisesti URI-osoitteita[17] (Universal Resource Identifier). Muitakin tunnistejärjestelmiä on kehitetty, kuten ITU-järjestön OID[18] (Object Identifier). Tunnisteen ideana on erottaa käsite tai muu olio toisista ja antaa tapa viitata siihen. Tunnisteen tulee olla ajallisesti pysyvä (persistent), sillä tunnisteen muuttamisen jälkeen pitäisi muuttaa myös kaikki tunnisteen ilmentymät aiemmin indeksoiduissa aineistoissa. Esimerkiksi sanaston laajentaminen uusilla termeillä ei saa vaikuttaa aiemmin nimettyjen käsitteiden tunnisteisiin.

Yksilöivän tunnisteen avulla kahdella eri käsitteellä voi olla sama ilmiasu. Esimerkiksi asiasana ”kilvet” voi viitata samassa sanastossa vaikkapa kilpikonnan tai muun eläimen kilpeen, jonka URI-tunnus on kilvet1, tai taistelussa käytettyyn suojaan, jonka URI-tunnus on kilvet2. Monissa asiasanastoissa merkitykset on yksilöity niiden ilmiasun eikä tunnisteen perusteella, jolloin eri käsitteillä ei voi olla samaa ilmiasua. Tämä voi johtaa kömpelöihin ja pitkiin käsitteiden nimiin, kuten:

kilvet (eläimen osa)

Ilmiasuun perustuva yksilöinti on ongelmallista myös silloin, kun merkitykseen pitää liittää useita ilmiasuja. Esimerkiksi monikielisissä asiasanastoissa yhteen käsitteeseen pitäisi voida viitata eri kielistä symmetrisesti, jolloin viittaus kieliriippumattomaan tunnisteeseen eikä jonkun tietyn kielen ilmaukseen on luontevaa.

Monissa asiasanastoissa ja luokitteluissa käytetään sanastokohtaisia pysyviä tunnisteita. Ongelmana tällöin voi olla se, että kahdessa tesauruksessa esiintyy sama tunniste eri käsitteille. URI-mekanismiin perustuvan yksilöivän tunnisteen avulla voidaan estää nimikonfliktit myös eri sanastojen välillä, ts. ettei kukaan muu ole sattumalta ottanut käyttöön samaa merkintätapaa jollekin toiselle käsitteelle. Tämä johtaisi sekaannuksiin sanastoilla kuivailtuja tietoja yhdistettäessä.

[17] http://www.w3.org/Addressing/
[18] http://www.itu.int/ITU-T/asn1/uuid.html

Globaalin käsitteiden erottelukyvyn aikaansaamiseksi URI-tunniste on muodoltaan web-osoitteen kaltainen, esimerkiksi:

http://www.sahko.fi/sanasto#johtaminen

Ideana on, että osoitteen aluenimi (domain), tässä www.sahko.fi, rekisteröidään webissä normaaliin tapaan sanastoa hallinnoivan organisaation toimesta, jolloin kenelläkään muulla ei voi olla samannimistä aluenimeä. Joku toinen organisaatio voi ottaa vapaasti käyttöön nimen johtaminen toisella alueella esimerkiksi seuraavan URI:n avulla:

http://www.yritys.fi/sanasto#johtaminen

Edellisessä tapauksessa sama paikallinen tunnus johtaminen voi tarkoittaa sähkön johtamista ja jälkimmäisessä yrityksen johtamista. Merkitykset erottuvat toisistaan ilman pelkoa nimien sekaantumisesta käyttämällä koko URI:a. XML:n nimiavaruuksien (name space) avulla voidaan eri sanastojen käsitteisiin viitata ekonomisesti lyhenteillä. Esimerkiksi http://www.sahko.fi/sanasto/ voidaan määritellä sahkö-nimiavaruudeksi, jolloin merkintä sahkö:johtaminen viittaa sähkönjohtamiseen.

URI-tunnisteen tärkeä ero muihin tunnistekoodauksiin on, että sillä voidaan paitsi tunnistaa identiteetti myös kertoa, missä päin webiä käsite on määritelty. Esimerkiksi liikkeenjohtamisen käsite tulisi määritellä tiedostossa http://www.yritys.fi/sanasto, josta sovellukset voivat käydä lukemassa sen. Tiedostossa oleva RDF- tai OWL-kielinen kuvaus määrittelee liikkeenjohtamisen merkityksen siihen liittyvien semanttisten suhteiden verkoston avulla..

3.4 Luokkien ja yksilöiden erottelu

Tesaurukset eivät yleensä erottele yksilöitä luokista. Esimerkiksi YSA:ssa on määrittely:

Halleyn komeetta LT komeetat

Halleyn komeetta on kuitenkin semanttiselta olemukseltaan yksilö eikä yksilöiden luokka, kuten komeetat. Kone ei tätä erottelua yleensä pysty tesauruksen perusteella tekemään. Iso alkukirjain tai se, ettei yksilöillä olisi alakäsitteitä, ei riitä perusteeksi: Esimerkiksi Pohjoismaat-termillä on suppeammat termit Suomi, Ruotsi jne. Yksilöiden ja näitä määrittävien luokkien erottelu on keskeinen idea tietotekniikassa; se on perustana mm. koko modernille oliokeskeiselle ohjelmoinnin paradigmalle. Ajatuksena on, että luokka määrittelee yksilöiden yhteiset ominaisuudet[19]. Esimerkiksi kaikilla komeetoilla on nimi ja jokin rata. Luokasta voidaan luoda yhä uusia yksilöitä, jolloin luokasta perityvien ominaisuuksien arvot kiinnittyvät: esimerkiksi Hale-Bobin komeetalla on nimi Hale-Bob ja tietty, Halleyn komeetasta poikkeava rata. Tesaurus standardit antavat pariaatteessa mahdollisuuden ilmaista yksilön ja luokan suhde hierarkkista BT/NT-suhdetta tarkentavalla BTI/NTI-suhteella, mutta tätä mahdollisuutta ei esimerkiksi YSA:ssa käytetä, joten se jää ontologisointiprosessin tehtäväksi.

3.5 Hierarkkisten suhteiden erottelu

Hierarkkiset suhteet esitetään tesauruksissa ST/LT-suhteella. Tämä suhdetyyppi jakautuu käytännössä kolmeen suhdetyyppiin: edellä mainittuun yksilö-luokka suhteeseen, ala-yläluokka-

[19] Lisäksi luokalla voi olla koko luokkaan liittyviä ns. luokkaominaisuuksia. Esimerkiksi luokan yksilöiden lukumäärä voisi olla koko luokkaa kuvaava luokkaominaisuus.

suhteeseen (hyponymia) ja osa-kokonaisuus-suhteeseen (meronymia). Esimerkiksi YSA:n määrittelyssä

sairaalat LT terveydenhuoltolaitokset

LT-suhde kuvaa luokkasuhdetta, mutta määrittelyssä

komeetat LT aurinkokunnat

kyse on osa-kokonaisuussuhteesta. Ongelmana on, että vaikka ihmislukija usein tajuaa eron automaattisesti, kone ei sitä voi yksinkertaisissakaan tapauksissa ymmärtää vaan tekee virhepäätelmiä. Edellisessä tapauksessa sairaalat perivät terveydenhuoltolaitosten ominaisuudet, mutta jälkimmäisessä tapauksessa komeetoilla ei ole aurinkokuntien ominaisuuksia, kuten omaa planeettajärjestelmää.

Jotta kone ei sekoittaisi hyponymiaa meronymiaan ja tekisi virheellisiä päätelmiä, tesauruksen hierarkkiset suhteen pitää erottaa toisistaan. Hyponymia voidaan esimerkiksi korvata rdfs:subClassOf ominaisuudella ja meronymia Dublin Core standardin dcterms:isPartOf ominaisuudella.

Tarkemmassa tarkastelussa osa-kokonaisuussuhde jakaantuu vielä moneen semanttisesti erilaiseen tapaukseen, joita on lueteltu taulukossa 3.1 (Fellman, 1998). Laajasti käytetty englanninkielinen WordNet-tesaurus[20] (Fellman, 1998) käyttää kolmea meronymiasuhdetta: component part of, member of, made form stuff.

Meronyminen suhde	**Esimerkki**
part / whole	branch / tree
member / collection	tree / forest
piece / whole	piece-of-cake / cake
material / object	aluminium / airplane
feature / activity	paying / shopping
phase /process	childhood / growing-up
place / region	Helsinki / Finland

Taulukko 3.1. Osa-kokonaisuussuhdetyyppejä.

WordNetissä tunnistetaan hierarkkisten luokka- ja osa-kokonaisuussuhteiden lisäksi verbien välinen troponymiasuhde, joka on eräänlainen toimintojen ja tapahtumien ala-yläluokkasuhde. Esimerkiksi ”kulkea” käsitteen troponyminen alakäsite voisi olla ”kävellä”, jolla taas voisi olla alakäsite ”ontua”.

[20] http://wordnet.princeton.edu/

YSA-YSO-muunnoshankkeessa ei erilaisia osa-kokonaisuussuhteita ainakaan hankkeen alkuvaiheessa erotella toisistaan.

3.6 Hierarkkisen suhteiden systematisointi hierarkiaksi

Asiasanaston hierarkkiset suhteet on usein kehitetty vain osittain, jolloin kaikkia käsitteitä ei ole kytketty toisiinsa eikä käsitteistä muodostu systemaattisista hierarkkista rakennetta. Esimerkiksi MASA:n 6000 käsitteestä n. 2300:lla ei ole yläkäsitettä ja sanaston rakenne on eräitä osia lukuun ottamatta varsin litteä. Tilanne on vastaava YSA:ssa.

Sanaston hyödyntämisen kannalta käsitteet olisi hyvä järjestää kattavasti hierarkioihin. Nämä ovat tarpeen mm. termejä lavennettaessa, sanaston esittelemisessä loppukäyttäjälle sovelluksissa sekä hakutulosten ryhmittelyssä. Systematiikka helpottaa myös sanaston ylläpitoa. Monissa kehittyneissä sanastoissa kuten AAT ja ICONCLASS[21] hierarkkiset rakenteet on kehitetty systemaattisesti huipulla olevan kaikkeuden käsitteestä alkaen. Kullakin alatasolla voi hyvän käytettävyyden kannalta olla vain rajoitettu joukko alakäsitteitä. MASAn ontologisoinnin yhteydessä hierarkioiden täydentäminen johti lukuisten uusien käsitteiden luomiseen.

Samalla kun hierarkkiset luokka- ja osa-kokonaisuussuhteet erotellaan ja täydennetään systemaattisiksi hierarkioiksi, pitää vielä tarkastaa, että näiden transitiivisuusominaisuudet ovat kunnossa. Tämä tarkoittaa sitä, että luokkahierarkiassa kaikki yläluokan ominaisuudet ovat myös kaikkien – ei ainoastaan seuraavan – alaluokkien ominaisuuksia ja käänteisesti että yksilöt ovat kaikkien yläluokkiensa yksilöitä. Tesauruksissa näin ei aina ole, mikä johtaa mm. virheellisiin hakutuloksiin termejä lavennettaessa. Vastaavasti on tarkistettava osa-kokonaisuussuhteiden toimivuus läpi koko hierarkian.

Tarkastellaan esimerkkiä YSA:sta:

mäntyöljy LT öljy
polttoöljy LT öljy
öljy LT kaivannaiset

Termiä laventavassa haussa haku suoritetaan myös hierarkiassa alempana olevilla termeillä. Ongelma on, että tässä esimerkissä ”kaivannaiset” hakusana lavenee aivan oikein ”öljyksi” ja ”polttoöljyksi”. Öljy lavenee mänty- ja polttoöljyksi, mikä sekin tuntuu perustellulta. Mutta kun molemmat laajennukset tehdään peräkkäin, lavenee kaivannaiset virheellisesti ”mäntyöljyksi”, joka ei ole peräisin maaperästä. Ongelma on ratkaistavissa jakamalla öljyn käsite kahtia ja luomalla kaksi eri hierarkiaa:

mäntyöljy LT öljy (biovalmiste)
polttoöljy LT öljy (kaivannainen)
öljy (kaivannainen) LT kaivannaiset

Meronymioiden osalta samantyyppiset suhteet ovat yleensä transitiivisia. Esimerkiksi

Helsinki isPartOf Suomi
Suomi isPartOf Eurooppa

[21] http://www.iconclass.nl/

joten voidaan päätellä:

Helsinki isPartOf Eurooppa

Erityyppisten osa-kokonaisuussuhteiden ketjuttaminen voi johtaa virhepäätelmiin. Esimerkiksi oksa on osa puuta (part/whole) ja puut muodostavat metsän (member/collection), mutta tästä ei seuraa, että oksat muodostavat metsän (member/collection).

3.7 Assosiatiivisten suhteiden esittäminen

Hierarkkisia suhteita suuremman semanttisen mallittamisen haasteen asettavat erilaiset assosiatiiviset suhteet. Näitä esitetään tesauruksissa kollektiivisesti RT-suhteella. Tesaurusstandardeissa annetaan esimerkkejä useista erilaisista assosiatiivisista suhteista, joita on esitelty taulukossa 3.2.

Assosiatiivinen suhde	**Esimerkki**
osa / kokonaisuus	paineastiat RT ydinreaktorit
ala / tutkimuskohde	seismologia RT maanjäristykset
prosessi / instrumentti	kampaaminen RT kammat
ammattityö / henkilö	kirjanpito RT kirjanpitäjät
toiminta / tuote	julkaiseminen RT kirjat
toiminta /kohde	opettaminen RT oppilaat
käsite / ominaisuus	naiset RT naisellisuus
käsite / alkuperä	vesi RT kaivot
syy / seuraus	kitka RT kuluminen
objekti / vastavoima	tuholaiset RT tuholaismyrkyt
raaka-aine / tuote	nahka RT nahkavyöt
toiminta / ominaisuus	ajanmittaus RT tarkkuus
käsite / vastakohta	suvaitsevaisuus RT suvaitsemattomuus

Taulukko 3.2. Assosiatiivisia suhdetyyppejä (Aitchison, 2000). Suomenkieliset esimerkit ovat kuvitteellisia eivätkä perustu YSA:an.

Sovelluksista ja kuvaustarkkuudesta riippuen assosiatiivisten suhteiden kirjo muuttuu helposti hankalan suureksi. Esimerkiksi lääketieteellisen UMLS-tesauruksen[22] (Unified Medical

[22] http://www.nlm.nih.gov/research/umls/

Language System) semanttisessa verkossa käytetään 54 eri semanttista suhdetta, joista suurin osa on luonteeltaan assosiatiivisia.

YSA-YSO-muunnostyössä assosiatiivisten suhteiden tarkempi mallittaminen jätetään ensivaiheessa eri sovellusten kontolle. Tällainen on esimerkiksi MuseoSuomesta kehitettävä kulttuurisisältöjä yhdistävä semanttinen KulttuuriSampo-portaali[23]. Siinä ajatuksena on soveltaa kehysperustaista (frame) lähestymistapaa, jossa olennaista ovat tapahtumat sekä näihin eri semanttisissa rooleissa (semantic role) liittyvät tekijät.

Esimerkiksi myyminen-tapahtuma voitaisiin kuvata kehyksellä, johon liittyy roolit toimija (myyjä), osallistuja (myytävä asia) ja kohde (ostaja). Ostaminen voitaisiin kuvata samalla semanttisella kehyksellä roolien täyttäjiä vaihtamalla. Tunnettu kehyksiin perustuva järjestelmä on esimerkiksi amerikkalainen FrameNet[24]. Tietämyksen esittäminen kehystyyppisinä rakenteina on tekoälytutkimuksessa ja tietämyksen esittämisessä laajalti käytetty menetelmä (Sowa, 2000).

Taulukossa 3.3. on hahmoteltu yksi mahdollinen sovelluksessa käytettävä roolijärjestelmä:

Rooli	**Selite**	**Esimerkki**
toimija (agent)	Toiminnan aktiivinen alullepanija tai suorittaja.	laulaminen – lintu laulaminen - Kirka
väline (instrument)	Toiminnassa käytettävä väline tai toiminnan edistäjä	naulaaminen - vasara
tulos(goal)	Toiminnan tulos	rakentaminen – talo opettaminen – oppiminen
kohde (patient)	Toiminnan kohteena oleva objekti.	opettaminen – oppilaat
paikka (location)	Toiminnan paikka, jolle voidaan antaa koordinaatit.	maraton - Helsinki
ympäristö	Toiminnan ympäristötyyppi.	puunhakkuu - metsä
aika (time)	Toiminnan ajankohta, joka voidaan sijoittaa aikajanalle.	tuhoutuminen – 2002-09-11
ajankohta	Toiminnan syklinen ajankohta.	iltarukous - ilta

Taulukko 3.3. Semanttisia rooleja.

[23] http://www.cs.helsinki.fi/group/seco/ontologies/kulttuurisampo/
[24] http://framenet.icsi.berkeley.edu/

Tarkastellaan esimerkkinä YSA:ssa olevaa seuraavaa assosiatiivista suhdetta, joka ilmaisee syytä ja seurausta:

aurinkotuuli RT revontulet

Tämä voitaisiin kehysmallissa esittää aiheuttaminen-tapahtuman yhtenä alityyppinä seuraavasti:

aiheuttaminen1	toimija	aurinkotuuli
	tulos	revontulet

Vaikka kehysmallinen esitysmuoto on näennäisesti mutkikkaampi kuin alkuperäinen perinteinen assosiatiivinen binäärirelaatio, saavutetaan kehysten avulla mm. seuraavia etuja:

Assosiatiivisen roolin semantiikka voidaan kuvata täsmällisesti toisen käsitteen avulla. Edellisessä esimerkissä syy-seuraus-suhde perustuu aiheuttamiseen. Tätä voidaan hyödyntää koneellisessa päättelyssä.

Määrittelyjen semantiikka kehysjärjestelmissä on täsmällisempi kuin perinteisissä assosiatiivisissa suhteissa. Mekanismi erottaa selkeästi toiminnat ja tapahtumat ja ilmaisee roolit näiden suhteen.

Roolien valinta riippuu sovelluksesta ja siitä, miten tarkkoja semanttisia roolierotteluja tarvitaan halutun toiminnallisuuden toteuttamiseksi. YSO:ssa tavoitteena on mahdollisimman monelle sovellusalueelle soveltuva yleisontologia, joten muunnostyön yhteydessä on mielekästä ottaa mukaan vain hyvin yleiskäyttöisiä rooleja. Harkittavaksi jää, missä vaiheissa ja missä laajuudessa assosiatiivisten suhteiden esittäminen on syytä tehdä YSO:n tulevissa versiossa.

3.8 Terminologialogiikat ja ontologisen kuvailun rikastaminen

Edellä on hahmoteltu asiasanastojen kevyttä ontologisointia. Semanttisen webin ontologiatekniikat ja kielet (Staab, Studer, 2003) tarjoavat mahdollisuuksia huomattavasti rikkaampaankin semanttiseen kuvailuun ja valmiita päättelykoneita, joilla ontologisista määritelmistä voidaan johtaa uutta tietoa. Tarkastellaan esimerkkinä OWL-kieliperhettä, joka on standardoitu W3C:n toimesta korkeimpaan mahdolliseen ”Recommendation” –tasoon.

OWL on tarkoitettu sanastojen formaaliin kuvaamiseen ns. kuvauslogiikan avulla (description logic) ja perustuu rakenteeltaan XML- ja RDFS-kieliin. Kuvauslogiikat ovat ensimmäisen kertaluvun predikaattilogiikan osajoukkoja, joille on onnistuttu kehittämään ratkeavia ja tehokkaita algoritmeja. Käytännössä OWL tarjoaa ontologian kehittäjälle mm. seuraavia mahdollisuuksia alemman tason RDF(S) kieleen verrattuna:

- Uusia käsitteitä voidaan muodostaa aiemmista mm. loogisten konnektiivien ”ja”. ”tai” avulla, ja luokkien samuus tai erillisyys voidaan ilmaista. Esimerkiksi ihmiset ovat miehiä tai naisia, mutta jokainen yksilö kuuluu vain jompaankumpaan sukupuoleen.

- Käsitteiden ominaisuudet voidaan määritellä luokkakohtaisesti. Voidaan esimerkiksi kertoa, että koirien pennut ovat koiria ja kissojen pennut kissoja. RDFS-kielen perussemantiikassa tämä ei ole mahdollista (ilman kielen laajentamista omien määrittelyjen kautta), sillä luokkien ominaisuudet (edellä ”pentu”) ovat globaaleja.

- Ominaisuuksilla voi olla kardinaliteetti. Voidaan esimerkiksi ilmaista, että koirilla on neljä jalka ja ihmisillä kaksi jalkaa, autolla taas ainakin kolme pyörää.

- Ominaisuudet voidaan määritellä transitiivisiksi (esimerkiksi esisän isä on myös esi-isä), käänteisiksi (esimerkiksi aviomies on vaimon kääteisominaisuus), funktionaalisiksi (esimerkiksi jokaisen ihmisen äiti on joku tietty ihminen) tai kääntäen funktionaalisiksi (esimerkiksi henkilön sosiaaliturvatunnuksesta voidaan päätellä yksikäsitteisesti henkilö, mutta äidistä ei).

Kuvausloogikoista on hyötyä sekä ontologiaa kehitettäessä että sitä käytettäessä. Kehitysvaiheessa formaalisti määritellyistä käsitteistä voidaan tarkistaa, että nämä ovat loogisesti konsistentteja eivätkä johda virhepäätelmiin. Määrittelyistä voidaan myös johtaa automaattisesti mahdollinen moniperintäinen luokkahierarkia (subsumption hierarchy), mikä helpottaa ylläpitotyötä. Kun ontologia on valmis, voidaan sitä hyödyntää esimerkiksi etsittäessä annetunlaisia yksilöitä, pääteltäessä yksilön kuulumista johonkin luokkaa yms. Ontologiaperustaiseen päättelyyn on käytettävissä valmiita päättelykoneita, kuten Protege-2000 editorin OWL-apuohjelmistoon (plug-in) sovitettavissa olevat FACT(++), RACER ja Pellet[25]. Ongelmana on, että niihin kuvauslogiikan päättelykoneisiin voida vielä nykyisin yhdistää luontevasti sääntöperustaista päättelyä, jota kuitenkin tarvitaan yleensä sovelluksissa. Loogiset säännöt muodostavat semanttisen webin ”teknologiakakussa” ontologia-kerroksen päällä olevan seuraavan kerroksen ja ovat vilkkaan tutkimus- ja kehitystyön kohteena mm. W3C:ssa[26].

Yksi hankaluus tässä työssä on, että kuvauslogiikat poikkeavat tietojenkäsittelyssä perinteisesti käytetystä ns. logiikkaohjelmoinnista ja Prolog-tyyppisistä kielistä, vaikka molempien looginen perusta onkin predikaattilogiikassa. Kuvauslogiikoissa ei tehdä ns. suljetun maailman oletusta (closed world assumption), jonka mukaan kaikki mitä ei voida kumota on totta. Oletus on hyödyllinen mm. tietokantajärjestelmissä ja monissa käytännön tilanteissa. Toinen merkittävä ero on, että kuvauslogiikat eivät käytä ns. yksikäsitteisten nimien oletusta (unique name assumption). Tämä merkitsee sitä, että kuvauslogiikoissa kaksi resurssia, esimerkiksi kaksi yksilöä, joilla on eri tunniste, eivät välttämättä ole eri yksilöitä. Näin voidaan esimerkiksi päätellä, että Iltatähti ja Aamutähti ovat sama planeetta Venus. Käytännössä yksikäsitteisten nimien oletus voidaan kuitenkin tehdä usein – tuntematon yksilö voidaan esimerkiksi ilmaista muuttujalla – ja oletus on käytössä tietokantajärjestelmissä ja Prolog-kielessä.

OWL-kieli on standardoitu kolmessa muodossa. OWL Light laajentaa RDF(S)-kieltä käyttökelpoisemmaksi mm. luokkakohtaisten ominaisuusrajoitteiden osalta. Seuraava aste OWL DL on kevytversiota monipuolisempi mm. loogisten konnektiivien avulla luotujen uusien luokkien osalta. OWL DL on kuitenkin edelleen laskennallisesti ratkeava kuvauslogiikka ja sille on saatavilla tehokkaita päättelykoneita. Vapain OWL:n muoto on OWL Full. Se on täysin yhteensopiva semanttisen webin alemman tason RDFS-kielen kanssa, mutta valitettavasti OWL Full ei ole ratkeava, ts. ei ole olemassa algoritmia, joka äärellisessä ajassa voisi ratkaista mielivaltaisen OWL Full kielellä esitetyn tehtävän. Todennäköisesti OWL Full tulee kuitenkin olemaan laajasti käytetty sen ilmaisullisen joustavuuden takia. Eri tavoin esitetylle tiedolle voidaan kehittää käytännössä erilaisia, sovelluskohtaisia päättelykoneita, vaikkei yleistä OWL Full -päättelijää voidakaan toteuttaa.

[25] http://protege.stanford.edu/plugins/owl/api/ReasonerAPIExamples.html
[26] http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/

3.9 Lopputulos

Kevyen ontologisoinnin lopputuloksena syntyy yhtenäinen, kaikki käsitteet systemaattisesti kattava luokkahierarkia, jossa on eroteltu yksilöt luokista ja jossa ominaisuuksia voidaan periä läpi koko luokkahierarkian. Käsitteillä on yksilöivät URI-tunnisteet, joiden määrittelyt löytyvät webistä RDF-muotoisina. Osa-kokonaisuussuhteet on erotettu ja näistä on mahdollisesti koostettu omat paikka-, aika- tms. ontologiansa. Tuloksena syntyneisiin taksonomiatyyppisiin ontologioihin voidaan eri sovelluksissa alkaa luoda tarkempia ontologisia kuvauksia esimerkiksi kehysmallin avulla.

Muunnoksen aikana ei ole syytä hävittää tarpeettomasti tesauruksessa olevia semanttisia suhteita tai muuta informaatiota, joilla on luonnollisesti oma arvonsa. LT, ST, RT ym. suhteet voidaan jättää sellaisenaan ontologiaan ja hyödyntää niitä siltä osin kuin se on mahdollista. Esimerkiksi MuseoSuomi-hankkeessa ontologioissa säilytettiin kaikki MASA-tesauruksen informaatio ja suhteet ja käytettiin assosiatiivisia RT-suhteita eri tavoin hyväksi tietosisältöjä toisiinsa yhdistettäessä. Tarkemman semanttisen mallin laatimiseen esimerkiksi kehysmallin pohjalta ei ollut käytettävissä aikaa tai resursseja.

4 Yhteenveto

Tesauruksia on perinteisesti laadittu ihmisten, erityisesti tiedon indeksoijan tarpeita silmällä pitäen. Tällöin termien semantiikan kuvaus on voitu jättää osittain ihmisen tulkinnan varaan. Yhä enenevässä määrin tesauruksia käyttävät kuitenkin tietokonesovellukset. Uusi tärkeä sovellusalue on mm. semanttinen web ja web-palvelut. Koska tietokoneelta puuttuu ihmisen piilotietämys ja tulkintakyky, täytyy termien merkitys määritellä sanastoissa aiempaa täsmällisemmin ontologioina. Artikkelissa esitettiin yksikertainen malli perinteisen tesauruksen muuttamiseksi semanttisen webin ontologiaksi. Mallia on kehitetty mm. Museoalan ontologian kehitystyössä ja ontologisoinnin tuloksia on pilotoitu menestyksellisesti useissa semanttisissa portaalihankkeissa. Mallia testataan parhaillaan FinnONTO-hankkeessa mm. Yleisen suomalaisen asiasanaston muutostyössä yksinkertaiseksi ontologiaksi.

Kiitokset

Kiitokset Katri Seppälälle, Suvi Kettulalle ja Eeva Kärjelle kommenteista artikkelin aiempaan luonnokseen liittyen. YSA-YSO-muunnoshankkeeseen ovat eri tavoin osallistuneet ja siihen vaikuttaneet: Juha Hakala, Tuula Haapamäki, Miikka Junnila, Tomi Kauppinen, Eeva Kärki, Peter Lindroos, Ville Komulainen, Tuukka Ruotsalainen, Mirva Salminen, Katri Seppälä, ArttuValo, Kim Viljanen ja Anu Ylikangas.

Viitteet

J. Aitchison, A. Gilchrist, D. Bawden (2000): Thesaurus construction and use: a practical manual. Europa Publications, London.

D. Fensel (2003): Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, Springer-Verlag, Berlin, 2001. 2nd Edition, Springer-Verlag, Berlin.

E. Hyvönen, A. Valo, V. Komulainen, K.i Seppälä, T. Kauppinen, T. Ruotsalo, M. Salminen, and A. Ylisalmi (2005a): Finnish National Ontologies for the Semantic Web - Towards a Content and Service Infrastructure. Proceedings of International Conference on Dublin Core and Metadata Applications (DC 2005), Madrid, Spain, short papers, 2005.

E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, .K Viljanen, S. Saarela, M. Junnila, and S. Kettula (2005b): MuseumFinland -- Finnish Museums on the Semantic Web. Journal of Web Semantics, Vol. 3, No. 2.

R. L: Leskinen (toim.) (1997). Museoalan asiasanasto. Museovirasto, Helsinki.

S. Staab, R. Studer (toim.) (2003) Handbook on Ontologies. Springer-Verlag, Berlin.

J. Sowa (2000): Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, California.

H. Suonuuti (2001). Guide to Terminology. NordTerm Publication 8, Tekniikan Sanastokeskus TSK, Helsinki.

C. Fellbaum (1998): WordNet. The MIT Press, Massachusetts.