Ontologian arviointi OntoClean-menetelmällä

Katariina Nyberg

[1.12.2008, Espoo]

TEKNILLINEN KORKEAKOULU KANDIDAATINTYÖN TIIVISTELMÄ
Informaatio- ja luonnontieteiden tiedekunta

Tekijä: Katariina Nyberg		
Työn nimi: Ontologian arviointi OntoClean-menetelmällä		
Sivumäärä: 23	Päiväys: 1.12.2008	Espoo
Tutkinto-ohjelma: Informaatioverkostot		
Vastuuopettaja: Stina Immonen		
Työn ohjaaja: Eero Hyvönen		

Semanttinen web on nykyaikainen visio internetistä, jonka sisältämä tieto on löydettävissä helposti, koska se on esitetty sellaisessa muodossa, jota kone ymmärtää. Semanttisen webin myötä on mahdollista luoda yhteensopivuutta palvelujen kesken, jotka eivät muuten olisi yhteensopivia, sekä suorittaa hakuja sanoilla, joiden merkitys kone ymmärtää.

Webin sisältämä tieto tehdään semanttiseksi kun siihen liitetään metatietoa. Tämän metatiedon merkitys on esitetty niin, että konekin ymmärtää sen ja osaa käsitellä sitä. Metatiedon merkitys on muutettu koneymmärrettävään muotoon ontologioiden avuilla. Ontologia on kokoelma käsitteitä ja niiden välisiä suhteita. Ontologioita voidaan luoda jo olemassa olevista asiasanastoista koneellisesti ja käsin.

Tässä kandidaatintyössä esitellään semanttisen webin ja ontologiatyön periaatteita ja tutkitaan OntoClean-menetelmää, jonka avulla ontologioita arvioidaan. Tämän lisäksi OntoCleanin mukainen ylätason ontologia, DOLCE-ontologia, esitellään. Teknillisen korkeakoulun semanttisen laskennan tutkimusryhmässä kehitetty Musiikkialan ontologia kuvaillaan ja sitä arvioidaan OntoClean-menetelmällä.

Nicola Guarino ja Christopher Welty ovat kehittäneet OntoClean-menetelmän ja tämä työ pohjautuu artikkeleihin, jotka he ovat kirjoittaneet siitä. Nicola Guarino on ollut mukana myös OntoClean-menetelmällä kehitetyn DOLCE-ontologian tekemisessä. Tämän lisäksi työ pohjautuu kirjallisuuteen, joka käsittelee semanttisen webin periaatteita ja asiasanastojen ontologiatyötä. Kandidaatintyön kirjoittaja on tutkimusapulaisena työstänyt musiikkialan ontologiaa MUSO:a. Tämän ontologiatyön tulokset ja OntoClean-menetelmän soveltaminen MUSO:on esitellään kirjoitelmassa.

MUSO noudattaa hyvin OntoCleanin periaatteita. Kun OntoClean-menetelmällä tarkistettiin MUSO:a havaittiin, että rekilaulut-käsite on virheellisesti kansanlaulut-käsitteen alakäsitteenä, vaikkakin rekilaulut-käsitteen määritelmässä lukee, että rekilaulu on kansanlaulu. OntoClean-menetelmä antaa hyvän pohjan ontologiatyön tarkistamiselle. Se kiinnittää huomion ontologisoinnin yleisiin virheisiin ja antaa vakaan pohjan ontologian sisältämien hierarkiaratkaisujen arvioinnille. Menetelmä ei yksinään riitä ontologiatyön tekemiseen, mutta sen periaatteet on hyvä sisäistää ontologiatyötä tehdessä.

Avainsanat: ontologiat, semanttinen web, metadata, asiasanastot, ontologisointi, ontologiatyö, OntoClean, DOLCE, MUSO	Julkaisukieli: suomi

Sisällysluettelo

1 Semanttinen web

Semanttinen web tarkoittaa verkkosisältöjen merkityksiä ymmärtävää tietoverkkoa. Internet-verkon sisältämä tieto on semanttinen, kun se esitetään sellaisessa muodossa, jota kone voi lukea ja tulkita. Tämä saavutetaan siten, että tietoon on liitetty metatietoa, joka selittää tietojen yhteyttä muuhun tietoon. Pelkkä tietoon liitetty metatieto ei kuitenkaan riitä semantiikan luomiseksi. Myös metatieto täytyy olla esitetty sellaisessa muodossa, jota kone pystyy lukemaan ja josta käy ilmi metatiedolla liitetyn tiedon merkitys. Käsitteistö, jota käytetään metatietona, esitetään ontologioissa. Ontologioista käy ilmi käsitteiden välisiä suhteita, mikä luo merkitystä webiin.

Toukokuussa 2001 Tim Berners-Lee, James Hendler ja Ora Lassila julkaisivat Scientific Americanissa artikkelin, joka oli otsikoitu The Semantic Web [1]. Tämä artikkeli maalasi ja lanseerasi tieteellisessä yhteisössä kuvan nykyisen webin laajennuksesta, jossa sovellukset keskustelevat toisten sovelluksien kanssa vaivattomasti ja kaivavat internetin loputtomista tietosyövereistä käyttäjälle hyödyllistä tietoa. Tämän lisäksi artikkelissa väitettiin, että nykyisen internet-verkon kehitys semanttiseksi webiksi johtaa ihmiskielen laajempaan kehitykseen.

Tässä kandidaatintyössä käsitellään niitä tapoja, millä tietoa ja sen merkitystä voidaan esittää koneymmärrettävässä muodossa. Tarkastelun kohteena on tiedon kuvailussa käytetyt asiasanastot ja niiden ontologisointi [2, s.2]. Lopuksi esitellään OntoClean-menetelmää [3], jolla ontologioiden loogisuutta voidaan tarkistaa. OntoClean-menetelmää noudattaen on luotu DOLCE-ontologia [4], jonka pohjalle voi luoda omia ontologioita. Sekä menetelmää että ontologiatutkimusta esitellään ja niiden hyödyllisyyttä arvioidaan. Kandidaatintyö pohjautuu lähdeluettelosta viitattuun kirjallisuuteen ja kirjoittajan omaan työkokemukseen Teknillisen korkeakoulun Semanttisen laskennan tutkimusryhmässä.

2 Älykäs yhdistely

2.1 Sovellukset keskustelevat toistensa kanssa

Semanttisen webin tarkoitus on lisätä tekoälyä koneille. Tekoälyllä ei tarkoiteta elokuvista tuttuja sympaattisia tai ihmisiä tappavia robotteja [1, s. 3] vaan sitä, että tieto on esitetty sellaisessa muodossa, jota kone pystyy käsittelemään. Internet-palvelimella tai teknisissä päätteissä toimivat sovellukset on usein ohjelmoitu eri kielillä, tavoilla ja periaatteilla. Elleivät sovellukset noudata samoja standardeja tai niitä ole erityisesti kehitetty yhteensopiviksi, ne eivät pysty toimimaan toistensa kanssa. Berners-Lee et al. [1] maalailevat kuvan Petestä, jonka kaikki ääntä tuottavat laitteet hiljentyvät, kun hän vastaa puhelimeensa. Tämä johtuu siitä, että Pete on antanut kännykälleen yleisen ohjeen, että aina kun hän vastaa puhelimeen, ne laitteet, joissa on säädin äänenvoimakkuudelle, hiljennetään. Nykyteknologialla olisi mahdotonta luoda tällainen yleinen sääntö ja toiminnallisuus, vaan jokaiselle äänentoistolaitteen äänen voimakkuutta olisi säädettävä erikseen.

Sovellukset eivät puhu samaa kieltä ja eivät siksi pysty kommunikoimaan toistensa kanssa. Jos ne pystyvät suorittamaan toiminnallisuuksia yhteistyössä toistensa kanssa, niin silloin ne noudattavat yhteistä protokollaa, joka on erikseen asennettu molempiin sovelluksiin. Lisäämällä koneymmärrettävää merkitystä sille, mitä sovellukset tarvitsevat toimiakseen, saamme sovellukset kommunikoimaan eri toteutustavoista huolimatta. Kun yhteensopivaa rajapintaa sovellusten välillä ei ole, se täytyy luoda. Jokainen sovellus vastaanottaa jonkinlaista tietoa (syöte, *engl. input*) ja palauttaa toisenlaista tietoa (tuloste, *engl. output*). Jotta sovellusta voidaan käyttää semanttisen webin tarkoitusten mukaisesti, on kuvailtava minkälaista tietoa sovelluksen tarvitseman syötteen ja tulosteen on oltava ja missä muodossa ne annetaan. Kuvailun täytyy olla siinä mielessä koneymmärrettävää, että sovellus osaa hyödyntää ja käyttää tätä tietoa automaattisesti. Ihmisen ei siis tarvitse erikseen jokaista sovellustyyppiä ja käyttötapausta varten ohjelmoida haluttua toiminnallisuutta.

Teknillisen korkeakoulun Semanttiset web-palvelut -kurssin [5] luennoilla mainittiin esimerkki tulostimesta lentokentällä. Ennen lentoa matkustajan on tulostettava tärkeä dokumentti, jotta hän voi lukea ja käsitellä sitä lentomatkallaan. Hänen kannettava päätteensä etsii lentokentän alueelta tulostimia. Nämä voivat olla eri valmistajien laitteita ja se muoto, missä ne vastaanottavat tulostettavaa tietoa, vaihtelee. Kaikki tulostimet eivät välttämättä edes vastaanota tavallisten

matkustajien syötteitä, vaan ne on varattu virkailijoiden käyttöön. Matkustajan langaton pääte tarvitsee listan kaikista laitteista, joissa on tulostimen toiminnallisuus. Jokainen tulostava laite on semanttisen webin tekniikalla merkitty koneymmärrettävästi tulostavaksi laitteeksi. Semanttisesti on merkitty myös se, missä muodossa kukin laite vastaanottaa tulostettavaa dokumenttia. Myös laitteen fyysinen sijainti lentokentällä ja onko matkustajalla lupaa käyttää laitetta, pystytään kertomaan. Matkustajan henkilökohtainen laite pystyy etsimään lähimmän käytettävissä olevan tulostimen, lähettämään dokumentin sille, etsimään lentokentän pohjakartan lentokentän yleisestä palvelusovelluksesta ja ohjaamaan matkustajan lentokentän pohjakartalla tulostimen luokse.

Sovellukset eivät puhu samaa kieltä eikä niillä ole mahdollisuuksia ymmärtää toisiaan. Semanttisen webin tekniikalla on tarkoitus luoda yhteys sovellusten välillä ja esittää niiden tuottama, tallentama ja vastaanottama tieto sellaisessa muodossa, jota sovellukset pystyvät käsittelemään.

2.2 Oikeanlaiset hakutulokset

Internet ei sisällä ainoastaan ihmisten kirjoittamia sovelluksia, vaan se on alkuperäisimmässä muodossaan sisältänyt ihmisten luomia dokumentteja. Nämä dokumentit sisältävät monenlaista tietoa monenlaisessa muodossa [1, s 4.]. Tietosisällöt on esitetty ihmisymmärrettävässä muodossa. Tämä on jokaiselle internetin käyttäjälle mielekästä. Internetin tarkoitus on toimia tiedotusvälineenä, jolla mahdollistetaan ihmisten välinen kommunikaatio.

Viestinnässä internetistä puhutaan englannin termillä *masses-to-masses medium*. Ihmismassat kommunikoivat ihmismassojen kanssa. Jotta kaikki tieto olisi kaikkien käytettävissä, on internetin tietosisältöä rikastettava niin, että myös koneet ymmärtävät tietosisällön merkityksen. Yksittäisen ihmisen on mahdotonta päästä selville internetin massiivisesta tietosisällöstä. Etsiäkseen oikeanlaista tietoa, ihmisen on turvauduttava hakukoneisiin, kuten Google-hakukoneeseen. Google etsii tietoa dokumenteista hakemalla niistä käyttäjän antamia hakusanoja. Se asettaa dokumentit tärkeysjärjestykseen sen mukaan, kuinka suosittu dokumentti on. Dokumentin suosiota mitataan mm. sillä, kuinka monta suoraa yhteyttä (hyperlinkki, puhekielellä linkki) johtaa muilta sivustoilta kyseessä olevalle sivustolle. Sivustojen suosio kasvaa sillä, että suositut sivustot linkittyvät sille.

Perinteinen hakutulos asiasanojen mukaan ei kuitenkaan aina johda haluttuun tulokseen. Hakusanoja vertaillaan muihin sanojen niiden pelkän leksikaaliseen ulkomuodon mukaan. Teknillisen korkeakoulun semanttisen laskennan professori Eero Hyvönen esitti Teknillisen korkeakoulun Semanttinen web -kurssin luennolla [6] esimerkin siitä, että hakusana ”politiikka” voi

antaa hakutulokseksi sivun, joka sisältää lauseen: ”Tämä sivu ei käsittele politiikkaa.” Hakutulokset riippuvat sivustojen suosiosta. On mahdollista kasvattaa keinotekoisesti omien sivujensa suosiota luomalla sivuverkoston, joka sisältää sivuja, jotka toimivat pelkkinä linkkilistoina.

Nykyiset hakukoneet palvelevat internetin selailijaa tyydyttävästi, mutta hakutulosten osuvuutta voidaan parantaa rikastamalla webiä semantiikalla. Semanttinen hakukone ymmärtää hakusanojen merkitystä ja pystyy tarvittaessa pyytämään tarkennusta käytettyyn hakusanaan: ”Tarkoitatko hakusanalla ’nokia’ Nokia-yhtiötä vai Nokia-kaupunkia?”

Semanttiselle hakukoneelle tai henkilökohtaiselle digitaaliselle apulaitteelleen voi antaa tehtäviä, kuten ”Etsi kotipaikkakuntani sairaalasta äidilleni hoitoaika, joka sopii kalenteriini”. Tässä esimerkissä yhdistyvät semanttinen tiedonhaku ja sovellusten yhteisymmärrys. Hakukone etsii hakijan paikkakunnan kaikki sairaalat, siivilöi niistä ne, jotka tarjoavat haluttua hoitoa, ja palauttaa ne sairaalat ja niiden tarjoamat hoitoajat sen mukaan, miten hakijan kalenterissa on vapaita aikoja [1, s. 2]. Paras hoitoaika voidaan valita sen mukaan, kuinka lähellä se on hakijan kotia eikä suosion mukaan. Ainoastaan suosion mukaan ehdotettu hoitoaika voisi sijaita täysin väärässä paikassa, jopa eri maassa. Hakuun voidaan liittää myös oikeanlainen suosituskriteeri, esimerkiksi lääketieteellisen aikakausilehden arviointi. Semanttinen hakukone etsii oikeanlaista tietoa ja semanttinen sovellus ottaa yhteyden muihin semanttisiin internet-sovelluksiin.

3 Tarkoituksenmukaisen tiedon löytäminen

3.1 Hakusanat

Google-hakukoneeseen saa lisäominaisuutena Google Suggestin[1], joka ehdottaa hakusanoille täydennystä. Kun hakusanakenttään kirjoitetaan keskeneräistä sanaa, kentän alle ilmentyy lista mahdollisista hakusanoista ja niiden yhdistelmistä. Tämä auttaa epävarmaa tiedonhakijaa löytämään sellaisen hakusanayhdistelmän, jota hän todennäköisesti tarvitsee. Täydennysmekanismi myös torjuu kirjoitusvirheitä. [2, s. 1.] Valitettavasti se ei tue vielä muita kieliä kuin englantia. Muun muassa YouTube-sivun hakukenttä käyttää samaa mekanismia.

Täydennysmekanismi helpottaa hakutoimenpiteitä ja on tämän takia askel sellaista internetiä kohti, jossa käyttäjällä on mahdollisuus löytää tarkoituksenmukaista tietoa. Tämä mekanismi pohjautuu sanojen kirjainmerkkeihin. Lista hakusanaehdotuksista muodostuu siitä, kuinka suosittuja hakusanat ovat, eli kuinka paljon muut käyttäjät ovat käyttäneet niitä. On mahdollista manipuloida suosiota kirjoittamalla toistuvasti niitä hakusanayhdistelmiä, jotka johtavat halutuille sivustoille. Tällä tavalla kone tulkitsee hakusanat suosituksi ja nostaa niiden sijaintia ehdotusten listalla.

Google tutkii tilastollista tietoa, jota se on kerännyt sen käyttäjien syöttämästä tiedosta. Yrityksessä kehitetään algoritmeja, jotka mittaavat kuinka suosittuja sivustot ovat ja mitä hakusanoja tai niiden yhdistelmiä käytetään eniten. Googlen tietosysteemit seuraavat massojen virtaa, ei itse tietoa. Kuin sopulilauma tämä tie joskus johtaa väärään suuntaan. Tarvitaan älykästä hakumenetelmää, joka pystyy massojen virtauksesta huolimatta löytämään oikeanlaista tietoa.

3.2 Tiedon esittäminen

Semanttisen merkityksen luomiseksi tarvitaan tiedon esittämistä (*engl. Knowledge Representation, KR*) [7]. KR on tekoälyn tutkimuksen piirissä syntynyt järjestelmä, jossa tietoa on punottu rakenteeseen ja näin syntyneen rakenteellisen tiedon avulla pystytään tekemään johtopäätöksiä. Perinteisesti KR on vaatinut keskitettyä järjestelmän, jossa jokainen osapuoli rajoittuu käyttämään vain sellaisia termejä, jotka on rakenteellisesti kuvattu. [1, s. 5]

Internet-verkko koostuu itsenäisistä solmukohdista (*engl. hubs*), jotka joko toimivat tai eivät. Solmukohdat ovat yhteydessä toisiinsa ja luovat täten verkon. Perinteiselle KR-systeemille internet

[1] http://www.google.com/webhp?complete=1&hl=en

luo haasteen, koska se ei ole kokonaisuus, joka on suljetusti hallittavissa. Järjestelmästandardien kanssa pystytään internetissäkin luomaan jonkin verran yhtenäisyyttä. Standardien laatiminen on hidas ja raskas prosessi eikä siksi ole mielekästä jättää sitä ainoaksi vaihtoehdoksi. Semanttinen web perustuu jo olemassa olevaan maailmanlaajuiseen kuvailukieleen, jonka avulla tietoa voidaan esittää kohdejärjestelmästä ja sen kuvailijoista riippumatta.

XML-kuvailukieleen (*eXternal Mark-up Language*) pohjautuva RDF-kuvailukieli (*Resource Description Framework*) [8] on työkalu, jolla tietoa kuvaillaan semanttisessa webissä. Kieli koostuu peruslauseista, joissa on subjekti, predikaatti ja objekti. Näitä lauseita kutsutaan kolmikoiksi ja ne esittävät kaikki ne yhteydet, joita erilaisille yksilöillä on. Kun tietoa kuvaillaan lukuisilla kolmikoilla, niiden merkitys saadaan sellaiseen muotoon, jota deterministinen tietokone pystyy laskennallisesti käsittelemään. Koneet eivät kykene ymmärtämään ja tulkitsemaan ihmisen tavoin, jolloin kaikki tieto, jota tietokoneen on tarkoitus käsitellä, täytyy olla loogisesti pitävää.

Tulkinnan sijaan koneet, sovellukset tai hakukoneet, muodostavat päättelyketjuja. Näitä voitaisiin palauttaa ihmisluettavassa muodossa. Ihminen, joka on antanut semanttiselle sovellukselle tehtävän, voisi myöhemmin tarkistella päättelyn kulkua ja arvioida sitä omalla ymmärryksellään [1, s. 2.].

Kolmikon osiin, subjektiin, predikaattiin ja objektiin, voidaan liittää globaali yksilöivä tunniste. Yksilöivänä tunnisteena käytetään URI-osoitteita (*Universal Resource Identifier*) [9]. Ne näyttävät verkkosivujen osoitteilta. Niissä on aluenimi, esimerkiksi *http://www.yritys.fi/*, ja paikallinen tunnus *#toimitusjohtaja*. Aluenimi varmistaa sen, että kuvailtu yksilö, jonka paikallinen nimi voi olla terminä yleinen, omaa yksilöllisen tunnisteen. Aluenimi voi olla sama kuin organisaation rekisteröimä kotisivun osoite. Tällöin URI:lla *http://www.yritys.fi/#toimitusjohtaja* merkityn käsitteen alkuperä on selvästi ilmaistu. [2] Tunniste ei itsessään välttämättä viittaa olemassa olevaan kotisivuun, vaikka onkin ilmaistu samanmuotoisena.

3.3 Asiasanastot

Kirjastot ja muut laitokset, joilla on isot tietokannat esineistä ja dokumenteista, kuten museot, joutuvat yhdenmukaistamaan sen tavan, miten esineet ja dokumentit kirjataan. Koska säilytettävä tieto täytyy olla löydettävissä, on mielekästä, että käytetään yhteisiä asiasanoja, kun kuvaillaan jotain säilytettävää esinettä tai dokumenttia. Tätä kutsutaan indeksoinniksi ja sen periaatteista on hyötyä myös silloin, kun halutaan esittää tietoa koneymmärrettävässä muodossa.

Organisaatiot, kuten museot ja kirjastot, ylläpitävät tietojärjestelmiä, jotka sisältävät paljon indeksoitavaa tietoa. Indeksointiin käytetään asiasanastoja. Asiasanastot sisältävät organisaation käyttöön tarkoitettuja käsitteitä, joilla pystytään merkitsemään tietojärjestelmien tallentamia tietoja.

Yleisen suomalaisen asiasanaston YSA:n [10] käyttöä suositellaan muun muassa Teknillisen korkeakoulun kandidaatinseminaarissa. Asiasanastot sisältävät termejä, jotka on linkitetty toisiinsa viiden eri suhteen mukaan: laajempi termi (LT) ja sen käänteinen suhde suppeampi termi (ST), rinnakkaistermi (RT) sekä suositeltava synonyymi (KÄYTÄ) ja sen käänteinen suhde ei-suositeltava synonyymi (KORVAA). Tämän lisäksi termiin voidaan liittää huomautus, joka kuvailee termin merkitystä ja käyttöä. [2, s.2-4]

Asiasanastot tarjoavat yhtenäisen termistön, jolla indeksointia pystytään suorittamaan. Myös synonyymien etsiminen ja käyttö helpottuvat. LT-suhteen avulla pystytään suppeampi termi sitomaan osaksi laajempaa termiä, jolloin yksityiskohtaisesti indeksoitu tieto on nopeammin löydettävissä.

Semanttinen web hyödyntää termien välisiä suhteita. Hakukone saa tietoa siitä, mihin käsitteisiin hakusana liittyy ja mitkä kaikki eri sanat tarkoittavat samaa kuin hakusana. Sanastot eivät ole kuitenkaan esitetty sellaisessa muodossa, että koneet pystyisivät tekemään loogisesti pitäviä päätelmiä niistä. Sanastojen sisältämiä suhteita voidaan toki hyödyntää laskennallisessa päättelyssä, mutta se voi johtaa virheelliseen tulokseen. Hyvönen (2005) esittelee *Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita?* - artikkelissaan esimerkin kahdesta laajempi termi -suhteesta, joiden merkitykset eroavat toisistaan huomattavasti: sairaalat-termin laajempi termi on terveydenhuoltolaitokset. Sairaala on tietynlainen versio terveydenhuoltolaitoksesta ja sillä on samat ominaisuudet kuin terveydenhuoltolaitoksella. Komeetat-termin laajempi termi on aurinkokunnat. Komeetta ei ole tietyntyyppinen aurinkokunta eikä sillä ole samanlaisia ominaisuuksia kuin aurinkokunnilla, esimerkiksi omaa planeettasysteemiä. Komeetta kuitenkin sisältyy aurinkokuntaan. Tämä eroavaisuus näiden kahden suhteen välillä on ihmiselle täysin ymmärrettävissä, sillä ihminen osaa tulkita tämän tiedon ja omaa tarpeeksi piilotietoa ymmärtääkseen näiden kahden samannimisen suhteen merkityksen eroavuudet. Kone ei kykene samaan, joten sille on tarkennettava, että ensimmäisessä suhteessa on kyseessä ylä-/alakäsite -suhteesta ja toisessa kyseessä on sisältyvyyssuhde. [2] Sanastojen sisältämien termien välisiä suhteita ei ole määritelty tarpeeksi tarkoiksi, jotta laskennallisesti ja täysin ilman tulkintaa toimiva kone pystyisi niitä käsittelemään.

Sanastoissa termit tunnistetaan niiden leksikaalisen nimen mukaan. Samalla termillä voi olla kuitenkin monenlaista leksikaalista versiota. Vieraskielisillä sanoilla, jotka on alun perin kirjoitettu muilla kuin latinalaisilla kirjaimilla, on olemassa monta eri kirjoitustapaa. Nämä eri tavat kirjoittaa sama sanat eivät siis ole synonyymejä keskenään, vaan ilmentymiä samasta asiasta. Tästä syntyy tarve sille, että tietokone osaa tunnistaa nämä käsitteet samana yksilönä. Aikaisemmin jo esitelty URI-tunniste tuo ratkaisun tähän. Siinä yksilölle on annettu tunniste ja sen kaikki eri kirjoitusvariantit pystytään yhdistämään tähän tunnisteeseen RDF-kielen kolmikoiden avulla.

3.4 Ontologiat

Asiasanaston muuttaminen koneymmärrettävään muotoon kutsutaan ontologisoinniksi. Ontologia on filosofian oppi olemisesta. Se tutkii asioiden olemusta ja niiden välisiä suhteita [11]. Tietojenkäsittelytieteessä ontologia on kokoelma käsitteitä, jotka on yhdistetty toisiinsa erilaisten suhteiden avulla. Useimmille käsitteille on olemassa jokin yläkäsite, jonka kaikki ominaisuudet käsite perii. Näitä käsitteitä kutsutaan luokiksi ja niiden hierarkkista suhdetta hyponymiaksi. Hyponymia esitetään kolmikoissa (subjekti . predikaatti . objekti) seuraavasti:

```
http://www.yso.fi/#alakäsite. rdfs:subClassOf . http://www.yso.fi/#yläkäsite
```

Hierarkkisen rakenteen lisäksi on mahdollista liittää termeihin asiasanastojen tavoin myös muita suhteita. Sisältyvyyssuhde voidaan ilmaista esimerkiksi Dublin Core -standardin [12] predikaatilla dcterms:isPartOf. Sisältyvyyssuhteita on monenlaisia. Sisältyvyyttä voi olla se, että jokin kaupunginosa kuuluu kaupunkiin, tai että käsi on osa ihmisen kehoa. Sisältyvyyssuhde voi ilmaista sitä, mistä materiaalista termin kuvaava käsite on tehty. Meronymisten suhteiden välillä on kuitenkin eroja johtuen siitä, että materiaalin ja tuotteen suhde on tuotteelle ehdoton, kun taas ihminen voi olla olemassa myös ilman toista kättä. [2, s. 10]

Neulontapuikot liittyvät villalankakerään, mutta kumpikaan termi ei sisälly toiseen eikä ole hierarkkisessa suhteessa toisiinsa. Ne kuitenkin voidaan liittää toisiinsa ontologiassa assosiatiivisella suhteella. Assosiatiivisten suhteiden tarkka merkitys käsitteille ei ole yksiselitteinen ja varmaa on vain jonkinlaisen suhteen olemassaolo kyseisten käsitteiden välillä. Assosiatiivisten suhteiden eri tyyppien määrä on suurempi kuin sisältyvyyssuhteiden, sillä termit voivat linkittyä toisiinsa usealla tavalla. Jotta koneen päättelyä pystytään johtamaan paremmin, assosiatiivisia suhteita voidaan kuvailla kehyksellä, jossa assosiatiivista tilannetta voidaan kuvailla monimuotoisemmin. Voidaan asettaa toimija, väline, tulos, kohde, paikka, ympäristö, aika ja ajankohta tälle assosiatiiviselle tilanteelle [2, s. 13]. Tilanteelle luodaan siis kehys, jossa siihen

linkitetyt termit on jaoteltu niiden rooliensa mukaan. Vaikka tämä on selkeämpi tapa esittää termien riippuvuussuhteita, voi kehyksien luominen olla raskasta. Sitä suositellaankin tehtäväksi sovelluskohtaisesti [2, s. 12].

3.5 Suomalaiset ontologiat

YSA on Teknillisen korkeakoulun semanttisen laskennan tutkimusryhmässä ontologisoitu Yleiseksi suomalaiseksi ontologiaksi (YSO) [13, s.1]. Se on osa FinnONTO-projektin[2] kehittämistyötä, jonka tarkoituksena on tehdä Suomesta semanttisen webin edelläkävijä. YSO:n lisäksi on olemassa lukuisia erityisalojen ontologioita, kuten valokuvausalan ontologia (VALO), taideteollisuusalan ontologia (TAO), museoalan ontologia (MAO) ja Viikin tiedekirjaston ontologia (AFO). Nämä on koottu KOKO-ontologiaan, jota voi selailla tutkimusryhmän kehittämällä ONKI-selaimella[3].

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

[3] http://www.yso.fi/onki/koko

4 Menetelmä ontologian arviointiin

4.1 OntoClean-menetelmä

Ontologiatyö voidaan tehdä käsin tai automaattisesti. Usein se tehdään puoliautomaattisesti niin, että automaattisia ontologiaratkaisuja tarkistetaan ja täydennetään käsin. Asiasanaston ontologisointi voidaan automatisoida, mutta ihmisen on silti tarkistettava ja arvioitava tulos. Guarino ja Welty arvioivat työn vaativan taitoa ja tarkkuutta [14, s. 1]. Ontologian sisältämän tiedon on oltava loogisesti pitävää, jotta tietokone pystyy käsittelemään sitä laskennallisesti. OntoClean-menetelmällä pyritään luomaan tapa, jolla vältetään niin sanottuja mallintamisen sudenkuoppia (*engl. modeling pitfall*) [3, s.152].

Ihminen ei käsittele tietoa saman lailla kuin tietokone. Ihmisen päättely on tulkinnallista. Tietokone on laskukone ja sen kaikki päättelyt ovat determinististä. Ihminen tietää, että valtameri on vettä, tällöin ontologiassa valtameren yläkäsite voisi olla vesi. Tämä ei kuitenkaan pidä paikkansa, sillä valtameri koostuu vedestä [3, s.156]. Valtameri ontologiakäsitteenä ei peri niitä ominaisuuksia, joita vedellä ontologiakäsitteenä on. Tulkinnallisen ymmärryksen ja deterministisen laskennan eroavaisuuksien takia ontologiatyöhön tarvitaan työkaluja, joilla ihminen pystyy arvioimaan ontologian sisältämien suhteiden loogista pitävyyttä. Työkaluja voidaan käyttää tarkistamaan esimerkiksi tehtyjä päätöksiä. Seuraavaksi esitellään OntoClean -menetelmää, jonka ominaisuuksilla pyritään välttämään yleisimpiä sudenkuoppia, joihin ontologiatyössä voidaan horjahtaa.

4.2 Liiteominaisuudet

OntoClean on menetelmä, jolla pyritään arvioimaan ontologiaratkaisuja. Nicola Guarino ja Christopher Welty ovat julkaisseet yhdessä artikkeleita, jossa he esittelevät menetelmää ja soveltavat sitä ontologiatyössä [mm. 3,14]. Näissä artikkeleissa esitellään liiteominaisuuksia (*engl. meta properties*), joilla pyritään kartoittamaan ontologian sisältämän käsitteen luonnetta ja arvioimaan sen periytyvyyssuhteita. OntoClean sisältää sääntöjä, jotka kontrolloivat käsitteiden periytyvyyttä. Tietyillä liiteominaisuuksilla varustetut käsitteet saavat periä ainoastaan tietynlaisia käsitteitä. OntoClean-menetelmässä käytetään liiteominaisuuksia epäloogisen hierarkian muodostumisen estämiseksi.

OntoClean-menetelmää esittelevässä teoksessa *An overview of OntoClean* [3] käytetään ominaisuus-sanaa (*engl. property*) silloin, kun puhutaan yhden käsitteen merkityksestä toiselle käsitteelle. Tämä ei tarkoita samaa asiaa kuin ominaisuus, josta puhutaan RDF-kielessä [3, s.152]. OntoCleanin ominaisuuden luonnetta ja käyttötarkoitusta kuvaillaan siihen liitetyillä liiteominaisuuksilla.

Esimerkiksi *ihminen*-luokka voidaan yhdistää *Petteri*-ilmentymään ja luoda ominaisuus, jolla ilmaistaan, että *Petteri*-ilmentymä on tyyppiä *ihminen*:

```
http://www.yso.fi/#Petteri . rdf:type . http://www.yso.fi/#ihminen
```

OntoCleanissä puhutaan ominaisuuksista kun tarkoitetaan jonkun erityisen luokan merkitystä siihen liitettäville yksilöille. Esimerkiksi omena-luokka vastaa omenana olemista [3, s.152]. Yleensä luokista luodaan ilmentymiä, mutta jotkut luokat ovat sen tyyppisiä, että niitä ei voida instanssioida. Nämä luokat voivat toimia ainoastaan jonkun ilmentymän roolina tai ominaisuutena. Tämän takia tässä työssä puhutaan yleisesti luokista tai käsitteistä ja niihin liitettävistä yksilöistä. OntoCleanissä määrätään liiteominaisuuksia luokille ja nämä liiteominaisuudet kertovat siitä tavasta, millä yksilö voidaan liittää ko. luokkaan.

OntoCleanissä mielletään ilmentymät luokan jäseninä [3, s.152]. Ilmentymä ei ole sama asia kuin alaluokka. Esimerkiksi Petteri ei ole ihmisen alaluokka, sillä ihmiset voidaan mieltää eläinlajiksi, joka on ollut olemassa nykyisessä muodossaan jo pari sataa tuhatta vuotta. Petteri ei ole eläinlaji eikä hän ole ollut olemassa nykyisessä muodossaan kuin korkeintaan noin 100 vuotta.

Liiteominaisuuksia on useita. OntoClean-metodia kehitetään yhä ja siksi tässä kandidaatintyössä keskitytään kolmeen keskinäiseen liiteominaisuuteen: *rigidity, identity* ja *unity*.

Rigidity (R)

Yhtä OntoClean-metodin liiteominaisuuksista kutsutaan englannin kielen termillä *rigidity (R).* Suomenkieleen käännettynä se tarkoittaa lujuutta tai jäykkyyttä. R-ominaisuus kuvailee käsitettä, joka on välttämätön siihen liittyvälle yksilölle. Ellei yksilö ole yhdistetty tähän käsitteeseen, yksilö ei ole olemassa. Ihminen-luokka on Petterille välttämätön ja tämän takia Petteri luodaan Ihminen-luokan instanssiksi. Jos hän lakkaa olemasta tyyppiä ihminen, hän lakkaa olemasta kokonaan. Tätä käsitteen välttämättömyyttä kuvaillaan merkinnällä +R. Joskus käsite voi olla välttämätön vain joillekin siihen liittyville yksilöille. Tätä liiteominaisuutta ilmaistaan merkillä -R (*engl. non-rigidity*). Esimerkiksi Ihmemaa Oz:in maailmassa aivojen omistaminen on välttämätön ihmisille

mutta ei variksenpelättimille. Aivojen omistaminen voidaan merkitä liiteominaisuudella -R tarinan kaikille toimijoille [3, s. 153]. Luokka, joka ei ominaisuutena ole välttämätön millekään siihen yhdistettävälle yksilölle, merkitään ~R (*engl. anti-rigidity*). Opiskelija-luokka on sellainen käsite, johon voidaan yhdistää ~R-liiteominaisuus, sillä yksilön olemassaolo voi jatkua senkin jälkeen kun hän on lakannut olemasta opiskelija [3, s. 153].

R-suhde on erittäin voimakas liiteominaisuus, sillä se kertoo paljon käsitteen hyödynnettävyydestä ontologiassa. Välttämättömät käsitteet, jotka on merkitty +R-liiteominaisuudella, muodostavat ontologian perusrungon (*engl. backbone*) [3, s. 163]. Näistä käsitteistä voidaan luoda suoraan ilmentymiä, sillä käsitteen ominaisuudet ovat välttämättömiä ilmentymien olemassaololle. Tällöin ei tapahdu sitä, että yksilö, joka on merkitty käsitteen ilmentymäksi, lakkaisi ilmentämästä käsitettä, mutta jatkaisi olemassaoloaan. Opiskelijaa ei ole järkevää merkitä opiskelija-luokan ilmentymäksi, vaan ihminen-luokan ilmentymäksi. Opiskelija-luokka ilmaisee eräänlaista väliaikaista roolia, jonka ihminen voi ottaa käyttöön.

Käsitteet, jotka eivät ole yksilöiden ominaisuuksina välttämättömiä, voidaan sisällyttää kuitenkin ontologiaan. Niiden tarkoitus on rikastaa ontologiaa. Nämä käsitteet voivat olla Guarinon ja Weltyn mukaan rooleja (*engl. roles*), ajallisesti muuttuvia käsitteitä (*engl. phased sortals*) ja jotain tiettyä laatua ilmaisevia käsitteitä (*engl. attributions*). Jälkimmäisintä ei kuitenkaan suositella käytettäväksi ontologian luokkana [3, s. 166]. Sen sijaan laadulle tulisi luoda ontologiaan yleinen luokka, jonka ilmentymän avulla tiettyä laatua voidaan kuvailla. Esimerkiksi punaiselle ei kannata luoda omaa luokkaa, vaan tulisi luoda väri-luokka, jonka ilmentymä punainen väri on.

Identity (I, O)

Ontologian käsitteen identiteettiarvoa kuvailee liiteominaisuus *identity*. Tällä liiteominaisuudella pyritään määrittelemään, ovatko kyseessä olevaan käsitteeseen liittyvät yksilöt erotettavissa toisistaan jonkun yhteisen identifioivan kriteerin mukaan [3, s. 155]. Ontologiassa voi olla entiteetti-luokka, joka voi olla monen erilaisen luokan yläkäsite. Luokan alakäsitteillä ei ole olemassa mitään yhteistä määritelmää, jonka mukaan ne voitaisiin erottaa toisistaan [3, s.158]. Tällöin olio-käsite merkitään määreellä -I. Identiteetin yhteinen kriteeri on ominaisuus, jonka käsitteen alakäsitteet perivät. Yleisesti se esitetään merkillä +I. Joskus käsitteellä on kuitenkin oma identifioiva kriteeri, jota se ei ole perinyt yläkäsitteeltään. Tällöin käytetään merkkiä +O. [3, s. 155] Identiteettiä ilmaisevalla liiteominaisuudella ei ole *anti-identity*-variaatiota. Guarino ja Welty siteeraavat W. V. Quinea, joka 1969 julkaistussa Ontological Relativity and Other Essays -

teoksessa ilmaisee: *"no enity without identity"*, eli ei entiteettiä ilman identiteettikriteeriä. Tämä tarkoittaa sitä, että ei ole olemassa käsitettä, jos sitä ei pystytä edes jollakin määritelmällä nimeämään ja erottamaan ko. käsitteeksi.

Unity (U)

Sen sijaan että pyrittäisiin erottamaan käsitteet toisistaan yhteisen (+I/O) tai erilaisten (-I) identiteettikriteerin mukaan, voidaan määritellä, ovatko käsitteeseen liittyvät yksilöt yhdenmukaisia keskenään. Tämän *unity*-liiteominaisuuden myötä pyritään myös erottamaan, ovatko käsitteeseen liittyvät yksilöt kokonaisuuksia itsessään vai eivät. Esimerkiksi Juridinen henkilö -käsite voidaan liittää yksityishenkilöille ja yrityksille. Yksityishenkilöt ja yritykset eivät ole yhdenmukaisia toistensa kanssa, joten ne merkitään *non-unity* (-U) liiteominaisuudella. Valtameri-luokkaan liittyvät yksilöt ovat kaikki yhdenmukaisesti valtameriä, joten ne merkitään +U-liiteominaisuudella. Vesi, joista valtameret koostuvat, ei käsitä minkäänlaista yhdenmukaisuutta siihen liitettävien yksilöiden kesken ja kaiken lisäksi se on materiaali, joka ei ole käsitettävissä kokonaisuudeksi. Sen kohdalla ei voi tehdä minkäänlaista erottelua sille, mitä osia sillä on. Tällaiset käsitteet, joihin liitettävät yksilöt eivät ole selkeitä kokonaisuuksia merkitään *anti-unity* (~U) -liiteominaisuudella. Esimerkiksi vesi-luokka voidaan mieltää virheellisesti valtameri-luokan yläluokaksi. Seuraavassa luvussa käsitellään sitä, miten OntoCleanillä voi tarkistaa periytyvyyden oikeanlaisuutta liiteominaisuuksien avulla.

4.3 Periytyvyys

Liiteominaisuudet määräävät luokalle periytyvyyden sääntöjä, jotka takaavat, ettei ontologia sisällä epäloogisia alaluokkasuhteita. Epäloogiset alaluokkasuhteet johtavat aikaisemmin mainittuihin sudenkuoppiin. OntoClean-metodi sisältää neljä sääntöä [3, s. 156]:

1. Yläluokan ollessa ei-välttämätön (~R) alaluokankin täytyy olla niin (~R).
2. Kun yläluokan ilmentymät voidaan kaikki erottaa yhteisellä identiteettikriteerillä (+I/+O), myös sen alaluokat on erotettava tällä tai omalla identiteettikriteerillään (+I/+O).
3. Kun yläluokan kaikki ilmentymät mielletään yhdenmukaisiksi (+U), sama täytyy päteä myös sen kaikkien alaluokkien ilmentymille (+U).
4. Kun yläluokan kaikki ilmentyvät eivät ole itsessään kokonaisuuksia (~U), ei myöskään sen alaluokkien ilmentymiä voi mieltää kokonaisuuksiksi (~U).

Näillä päättelysäännöillä voidaan esimerkiksi huomata sellainen epälooginen virhe ontologian hierarkiassa, jossa valtameri on veden alaluokkana. Vesi-luokan yksilöt eivät ole kokonaisuuksia, vaan ne ovat abstrakteja määreitä (~U). Valtameret ovat kokonaisuuksia ja ne voidaan mieltää yhteenkuuluviksi (+U). Neljäs periytyvyyssääntö pätee tässä. Näillä kahdella määreellä on assosiatiivinen suhde, jossa sijainti koostuu maantieteellisistä alueista. Koostumus ei merkitse hierarkiaa.

Auto koostuu autonosista, joista yksi on moottori. Moottorin yläluokaksi voisi siis merkitä autonosa. Tämä on kuitenkin väärin, sillä moottori voi hyvin toimia myös veneen moottorina, jolloin se ei ole enää autonosa. Tällaista selkokielistä tarkistusta ei kuitenkaan aina välttämättä keksi, jolloin on tärkeää tarkistaa hierarkiasuhteen loogista pitävyyttä liiteominaisuuksien avulla. Autonosa on luokka, joka ei ole välttämätön siihen liitettäville yksilöille. Yksilö, joka toimii autonosana, voi olemassaolonsa aikana päätyä esimerkiksi rap-taiteilijan kaulakoruksi, jolloin se ei enää ole autonosa. Yksilö itsessään ei kuitenkaan tämän takia lakkaa olemasta. Kyseessä on rooli, jonka yksilö voi ottaa väliaikaisesti käyttöön, eli autonosa on ei-välttämätön käsite (~R). Moottori sen sijaan on ominaisuutena välttämätön sen ilmentymille, sillä jos ne lakkaavat olemasta moottoreita ne lakkaavat olemasta kokonaisuudessaan. Myös toimintakyvytön moottori on silti moottori. Moottori merkitään liiteominaisuudella +R ja ensimmäinen periytyvyyssäännön mukaan autonosa ei ole moottorin yläluokka. Rooli ei merkitse hierarkkista suhdetta. [14, s. 4]

5 DOLCE-ontologia

OntoClean-menetelmään nojautuen on kehitetty DOLCE-ontologia, joka on välttämättömistä käsitteistä (+R-liiteominaisuus) koostuva ylemmän tason ontologia [4, s.4]. Sen tarkoitus on luoda yleinen pohja jokaiselle ontologialle, jota luodaan. Siinä on se hyöty, että eri tahojen kehittämät ontologiat voidaan yhdistää helposti, jos ne noudattavat samanlaista ylemmän tason ontologiaa.

DOLCE määrittelee perustaluontoiset yläkäsitteet, joiden alaisuuteen erikoistuneiden sanastojen käsitteet voidaan järjestää. YSO, joka muodostaa KOKO-ontologian juuren, noudattaa DOLCE-ontologian kolme peruskategoriaa: abstrakti (*Abstract*), muuttuva (*Perdurant/Occurence*) ja pysyvä (*Endurant*) [6, s. 4]. Pysyvät ja muuttuvat -luokkien ero juontuu siitä, miten nämä suhtautuvat ajan kulumiseen. Karkeasti sanottuna pysyvät-luokan ilmentymä voi ominaisuuksiltaan muuttua ajan myötä, mutta on ajan kulusta huolimatta sama ilmentymä. Muuttuvat-käsitteen ilmentymä on olemassaololtaan sitoutunut aikaan. Oletetaan, että on olemassa kaksi ajasta riippuvaa ominaisuutta, jotka eivät voi vallita samaan aikaan. Pysyvän käsitteen ilmentymään voidaan liittää kummatkin ominaisuudet (eri aikoina), ilman että ilmentymän identiteetti muuttuu. Muuttuvat-käsitteen ilmentymä muuttuu toiseksi ilmentymäksi, kun siihen on liitetty nämä erilaiset ominaisuudet. Esimerkiksi ihminen on pysyvät-luokan ilmentymä. Keskustelu, joka hänellä on toisen ihmisen kanssa, on muuttuvat-luokan ilmentymä. [6, s. 3].

DOLCE-ontologian ja OntoClean-metodin jaot ja periaatteet muodostuvat sen mukaan, miten ihminen hahmottaa maailman. Se on sidoksissa sovittuihin käsityksiin maailman jaosta. OntoClean-menetelmällä pyritään luomaan ontologioita, jotka ovat loogisia ja soveltuvat laskennalliseen päättelyyn. Menetelmän liiteominaisuudet kuitenkin pohjautuvat ihmisten käsitykseen maailman rakenteesta. [6, s. 2] Tämä tuntuu hieman ristiriitaiselta, sillä pyritäänhän OntoClean-menetelmällä pääsemään eroon ihmisen tulkinnan virheistä. Ristiriitaa ei kuitenkaan ole, jos pohdimme sitä, mihin ontologioita tarvitaan. Niin kuin aikaisemmin määriteltiin, ontologioiden avulla pyritään luomaan käsitteistö, jonka avulla tiedon merkitystä voidaan kertoa tietokoneelle. Tietokoneelle on kerrottava, mitä kuvailtu tieto merkitsee ihmiselle, jotta tietokoneen päättelyketjuista on jotakin hyötyä ihmiselle. OntoClean-menetelmä ei noudata tarkkaa loogista hierarkiarakennetta sen takia, että se pystyisi aukottomasti selittämään maailman, vaan siksi, että se pystyisi esittämään ihmisen käsitystä maailmasta loogisesti pitävällä tavalla.

6 Musiikkialan asiasanaston työstäminen ontologiaksi

6.1 Automaattisen ontologioiden yhdistelyn tarkistaminen

Musiikkialan asiasanasto (MUSA) sisältää 900 käsitettä, jotka on tarkoitettu musiikin indeksointiin [15]. Koska se noudattaa YSA:n periaatteita ja rakenteita [15], oli mielekästä ontologisoida MUSA osaksi YSO-ontologiaa. Ontologioiden yhdistely tapahtui siten, että tehtävää varten kirjoitettu ohjelma muutti asiasanat ontologian käsitteiksi ja loi niiden välille subClassOf-suhteen laajempi termi (LT) -suhteen avulla (ks. luku 3.3). Jokaista LT-suhdetta kohden lisättiin ontologiassa yksi kolmikko kuvailemaan tätä periytyvyyttä. Tämän jälkeen käsitteiden nimiä vertailtiin YSO:n käsitteiden nimiin ja samannimisille käsitteille luotiin ekvivalenssisuhde. Tuloksena syntynyt ontologia tunnetaan nimellä MUSO.

Kesällä 2008 sain tehtäväkseni tarkistaa ja tarvittaessa korjata MUSO:a. Ontologiaeditori Protégé-ohjelmalla[4] kävin sitä läpi ja arvioin ekvivalenssi- ja hierarkiasuhteita. MUSO ei ole ontologia, jota voisi järkevästi esittää yksin, koska se ei sisällä omia abstrakteja yläkäsitteitään. Sen sijaan se on osa YSO:a ja täydentää tätä musiikkialan asiasanoin. Ekvivalenssisuhteet ovat tärkeitä sen takia, että niiden avulla saadaan MUSO-käsite ripustettua YSO:on. Jos MUSO-käsitettä ei ole YSO:ssa, sille täytyy löytää yläkäsite, joka on joko ripustettavissa YSO:on tai jo valmiiksi YSO-käsite.

Useimmissa tapauksissa ekvivalenssisuhteet pitävät paikkansa: YSO ja MUSO sisältävät 485 samaa käsitettä. Todettuani ekvivalenssisuhteen oikeaksi tarkistin YSO-käsitteen paikan hierarkiassa. Esimerkiksi laulajat olivat YSO:ssa virheellisesti näyttämötaiteilijoiden alaisuudessa. Laulajat ovat muusikoita siinä missä viulistitkin ja käsitteen oikea yläkäsite on muusikot. Jouduin tekemään kymmenkunta hierarkiamuutosta YSO:on. Muutamassa tapauksessa pystyin luomaan ekvivalenssisuhteen kahden synonyymisen käsitteen välille: continuo tarkoittaa samaa kuin basso continuo.

Muutamat YSO:n samannimiset käsitteet tarkoittavat muuta kuin samannimiset käsitteet musiikkitieteissä. Esimerkiksi intonaatio merkitsee YSO:ssa äänenkorkeuden vaihtelua puheessa ja sen yläluokkia ovat kielelliset ilmiöt ja sisäiset ominaisuudet. Musiikkitieteissä intonaatio tarkoittaa kolmea eri asiaa: sävelpuhtautta, pianojen ja urkujen soinnillisten ominaisuuksien säätämistä sekä tietyn katolisen sävellystyypin alkulaulua. Tällöin ekvivalenssisuhde oli purettava. Intonaatio-

[4] Stanford Center for Biomedical Informatics Research 2008, Protégé, (http://protege.stanford.edu/)

käsitteiden tapauksessa täytyi eritellä saman asiasanan taustalla olevat käsitteet, jotta jokaiselle eri merkitykselle olisi oma URI-tunnisteensa. Yhteensä MUSO:n asiasanojen taustalla olevien käsitteiden tarkempi erittely tuotti 46 uutta käsitettä.

MUSO:ssa on 473 käsitettä, joille ei löytynyt ekvivalenttia käsitettä YSO:sta. Tämä ei ollut ongelma, jos niiden yläkäsitteille löytyi ekvivalentteja YSO:sta. Jos sopivaa yläkäsitettä ei ollut MUSO:ssa, se täytyi löytää YSO:sta, jotta käsite saataisiin ripustettua osaksi ontologiaa. Usein sopiva yläkäsite löytyi. Joissakin tapauksissa oli mielekästä luoda YSO:n yläkäsitteelle tarkentava alakäsite. Esimerkiksi sisäiset ominaisuudet -käsitteen alakäsitteeksi luotiin musiikilliset sisäiset ominaisuudet -käsite. Uusia tarkentavia käsitteitä luotiin kolme.

Ekvivalenssi- ja hierarkiasuhteiden tarkistuksen suoritin sen tietämyksen pohjalta, jonka olen hankkinut musiikkitieteen opinnoissani ja Teknillisen korkeakoulun Semanttisen webin kurssilla [6]. Tämän lisäksi sain ohjausta hierarkiarakenteista Semanttisen laskennan tutkimusryhmän YSO:n kehittäjältä Katri Seppälältä. Hänen kanssa kävin lukuisia keskusteluita käsitepiirteiden periytyvyydestä.

6.2 MUSO-ontologian jälleenarviointi OntoClean-menetelmällä

Ontologiatyöni jälkeen tutustuin OntoClean-menetelmään ja arvioin ontologiatyöni uudelleen käyttäen OntoCleanin liiteominaisuuksia. Kävin ontologiaa ensin läpi ja valitsin sen jälkeen muutaman osapuun, joiden hierarkiaratkaisuja halusin tarkistaa. Koin jokaisen U, I ja R-liiteominaisuuden erikseen pohtimisen työlääksi ja minulla oli vaikeaa keksiä itselleni kysymyksiä, joiden avulla voisin arvioida, minkälaisen liiteominaisuuden kullekin käsitteelle keksin. Liiteominaisuus, jonka arvioinnin koin selkeimmäksi on välttämättömyydestä kertova R-liiteominaisuus. Sitä voidaan arvioida pohtimalla, lakkaako yksilö olemasta, jos sitä ei enää liitetä ko. käsitteeseen.

OntoCleanin mukaan hierarkiasuhde on väärä, jos ei-välttämätön käsite (~R) perii käsitteen, joka on kokonaan tai osittain välttämätön (+/-R, ks. Luku 4.3). Kesti kauan, kunnes löysin MUSO:sta tällaisen virheen: Kansanlaulut-käsitteen alakäsite on rekilaulut ja yläkäsite laulut (ks. kuva 1). Laulut-käsitteen alaisuudessa on musiikillisia tuotoksia, jotka on tarkoitettu laulettavaksi. Kaikista ei kuitenkaan voi olettaa, että ne ovat välttämättömästi vain lauluja. Laulut-käsite on osittain välttämätön ja sille liitetään -R-liiteominaisuus. Kansanlaulut-käsite voidaan liittää niille laulettavaksi tarkoitetuille musiikillisille tuotoksille, jotka mielletään kansanmusiikiksi.

Kansanmusiikki on määritelmä, joka on annettu sen käsittämille musiikkiteoksille vasta jälkeenpäin. Se on kuulonvaraisesti siirtynyttä kansanperinnettä [16, kansanmusiikki], joten teoksien syntyhetkellä niitä ei käsitetty kansanmusiikiksi, vaan pelkäksi musiikiksi. Kansanlaulut-käsite on ei-välttämätön siihen liitettäville käsitteille, eli siihen liitetään ~R-liiteominaisuus. Rekilaulut ovat määritelmän mukaan:

> 1800-luvulla yleistynyt suomalainen, alkuaan yksisäkeistöinen, kahdesta keskenään loppusoinnullisesta säeparista muodostunut lyyrissävyinen kansanlaulu. Ensimmäinen säepari sisältää tavallisesti yleisluontoisen mieleen johdatuksen, yleensä luontoaiheisen, johon jälkisäepari liittää laulajan omakohtaisen kokemuksen. [16, rekilaulut]

Käsite on erittäin tarkasti määritelty ja jos on olemassa teos, joka sopii määritelmään, se tunnistetaan rekilauluksi. Rekilaulut-käsite on täten välttämätön käsite (+R). Rekilaulut-käsite ei voi olla kansanlaulut-käsitteen alakäsite OntoClean-menetelmän mukaan, joka määrittelee, että kaikki ei-välttämättömät käsitteet voivat periä ainoastaan ei-välttämättömiä käsitteitä (ks. Luku 4.3). Kansanlaulut-käsite on rooli, joka voidaan antaa laululle. Rekilaulun määritelmässä on mainittu, että rekilaulu on kansanlaulu. Tämä ei tarkoita sitä, että rekilaulu on kansanlaulun alakäsite. Se on luonnollisen kielen epätarkka tapa sanoa, että rekilaulu mielletään kansanlauluksi, eli sillä on sellainen rooli nykyajan musiikkitieteissä.

```
yso:yso-käsitteet -U-I-R
└yso:pysyvä -U-I+R
 └yso:henkiset tuotokset -U+I+R
  └yso:kulttuuriset tuotokset -U+I+R
   └yso:musiikilliset tuotokset -U+I+R
     └yso:sävellykset (muso) +U+O+R
      └muso:laulut (yso) +U+O-R
       └muso:kansanlaulut (yso) +U+I~R
        └muso:rekilaulut +U+I+R
```

Kuva 1Kansanlaulut ja rekilaulut -luokkien yläluokat YSO:ssa

6.3 OntoClean-menetelmän hyödyllisyys

OntoClean-menetelmä paljastaa hyvin niitä loogisia virheitä, joita syntyy kun tulkintaan kykenevä ihminen luo ontologioita. Liiteominaisuuksilla määritellään jokainen käsite sen mukaan, mikä niiden merkitys on muille käsitteille ja minkälaisiksi ne mielletään. Omassa kokeilussani havaitsin hyvin nopeasti, että liiteominaisuuksien keksiminen jokaiselle käsitteelle on erittäin työlästä. Mielestäni olisi ajan haaskausta, jos jokaiselle jo luodun ontologian käsitteelle liitettäisiin kolme

liiteominaisuutta. Se on uuvuttavaa työtä, joka vaatii paljon keskittymistä ja jossa voi uupumisen takia syntyä monta ajatteluvirhettä.

En suosittelisi OntoClean-menetelmää koko ontologian tarkistamiseen, koska menetelmä ei ota kantaa ylä- ja alakäsitteiden varsinaisiin ominaisuuksiin. Se on menetelmä, jolla arvioidaan jo tehtyjen hierarkiapäätösten loogista pitävyyttä ja antaa yhden selkeän tavan arvioida tiettyjä hierarkiaratkaisuja. Se soveltuu arviointimenetelmäksi niille ratkaisuille, joista on hieman epävarma. Ontologiatyötä tekevälle on hyödyllistä tutustua OntoClean-menetelmään ja sisäistää ne yleiset hierarkiavirheet, joita OntoCleanillä pyritään välttämään. Tällöin ontologiaa luodessaan hän voi käydä mielessään läpi OntoCleanin periaatteita ja testata ongelmakohdissa liiteominaisuuksilla hierarkian oikeanlaisuutta.

7 Jälkipuhe

Tämä kirjallisuuskatsaus esittelee ontologiatyön monia vaiheita. Se keskittyy kuitenkin ainoastaan ontologiatyön viimeiseen vaiheeseen, ontologian arviointiin, esittelemällä OntoClean-menetelmää. Kattavampi työ olisi voinut esitellä myös yksityiskohtaisemmin automaattista ja käsintehtyä ontologiatyötä.

Tämä kandidaatintyö paljasti sen, että OntoClean-menetelmää voidaan käyttää vain yhteen pieneen mutta tärkeään osaan ontologiatyöstä. Menetelmä ei tarjoa tapoja muuttaa käsitteistöjä tietokoneymmärrettävään muotoon. Sen tarjoamat tarkistusmuodot ovat hyviä paljastamaan jo tehtyjen ratkaisujen laatua, mutta eivät itsessään anna hierarkiaratkaisuja.

DOLCE-ontologia on hyvä pyrkimys luomaan yhteinen pohja kaikille ontologioille. Sen ja OntoCleanin noudattaminen ontologioissa varmasti helpottaa semanttisen webin sovelluksia, jotka voivat luottaa siihen, että hierarkiaratkaisut on tehty aina saman periaatteen mukaan. Tällöin sovellukset voivat esimerkiksi käydä hierarkiapuuta läpi yhdellä tietynlaisella tavalla ja odottaa samanlaisia tuloksia ontologialta kuin ontologialta. Esimerkiksi roolit löytyvät aina omasta kohdastaan puusta ja kaikki ilmentymät on luotu luokista, jotka voidaan olettaa varmasti välttämättömiksi.

OntoClean ei kuitenkaan ratkaise sitä ontologioiden välistä ongelmaa, joka syntyy, kun ontologiat on luotu eri kielillä. OntoCleanin mukaan ontologia koostuu käsitteistä sellaisen hierarkian mukaan, joka kuvailee, miten ihminen hahmottaa maailman. Eri kielillä on hieman eri tapa ilmaista ja käsittää maailmaa ja tämän takia myös eri kielillä laaditut ontologiat eroavat periaatteiltaan.

Tämä kandidaatintyön onnistui esittelemään semanttista webiä yleisölle, joka ei välttämättä tiedä sen takana olevasta tekniikasta mitään. Se esitteli selkeästi suomeksi OntoClean-menetelmän periaatteita ja sitä, miten menetelmä toimii. Näiden lisäksi työ dokumentoi MUSO:n ontologiatyötä ja sovelsi OntoCleaniä MUSO:on.

8 Lähdeluettelo

1 Tim Berners-Lee, James Hendler and Ora Lassila (2001): The Semantic Web, Scientific American

2 Eero Hyvönen (2005): Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita? TKK, Espoo

3 Nicola Guarino, Christopher Welty (2004): An Overview of OntoClean, Trento/Hawthrone (sisältyy julkaisuun S. Staab, R. Studer (toim.) (2003): Handbook on Ontologies. Springer-Verlag, Berlin)

4 Aldo Gangemi et al. (2002): Sweetening Ontologies with Dolce, Rooma/Padova

5 Teknillinen korkeakoulu (syksy 2008): AS-75.3600 Semanttiset web-palvelut L, (https://noppa.tkk.fi/noppa/kurssi/as-75.3600/luennot, 27.11.2008)

6 Teknillinen korkeakoulu (kevät 2008): AS-75.2500 Semanttinen web L, (https://noppa.tkk.fi/noppa/kurssi/as-75.2500/etusivu)

7 J.R. Sowa (2000):, Knowledge Representation: Logical, Philosophical and Computational Foundations, Brooks-Cole, Pacific Grove

8 W3C World Wide Web Consortium: RDF, (http://www.w3.org/RDF/, 27.11.2008)

9 RFC3986 (2005), (http://www.ietf.org/rfc/rfc3986.txt)

10 Helsingin yliopiston kirjasto (2000): Yleinen suomalainen asiasanasto (http://vesa.lib.helsinki.fi/ysa/)

11 Wikipedia - fi (10/2008): Ontologia, (http://fi.wikipedia.org/wiki/Ontologia)

12 Dublin Core Metadata Initiative (2008) (http://dublincore.org/)

13 Hyvönen et al. (2007): Yleinen suomalainen ontologia YSO, TKK, Espoo

14 Nicola Guarino, Christopher Welty (2002): Evaluating Ontological Decisions with OntoClean, Communications of the ACM, New York

15 Semantic Computing Research Group (2008): www.yso.fi Documentation Wiki, MUSA, (http://www.yso.fi/wiki/index.php?title=Musiikin_asiasanasto, 27.11.2008)

16 MUSA Onki-selaimella, (http://www.yso.fi/onki/musa/, 27.11.2008)