TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY

Robin Lindroos

Paikkatiedon ontologiapalvelu

Diplomityö, joka on jätetty opinnäytteenä tarkastettavaksi
diplomi-insinöörin tutkintoa varten.

Espoon Otaniemessä 30.5.2008

Työn valvoja: professori Eero Hyvönen

TEKNILLINEN KORKEAKOULU **DIPLOMITYÖN TIIVISTELMÄ**

Tekijä, Diplomityön nimi

Robin Lindroos
Paikkatiedon ontologiapalvelu

Päivämäärä: 30.5.2008 Sivumäärä: 74

Tiedekunta

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Professuuri

AS-75 Viestintätekniikka

Työn ohjaaja

professori Eero Hyvönen

Tiivistelmäteksti

Tämä diplomityö käsittelee menetelmiä, joilla paikkatietoaineistoja muunnetaan ontologiseen muotoon sekä esittelee palvelun, ONKI-Paikan, jolla ontologisessa muodossa olevaa paikkatietoa voidaan tuottaa, ylläpitää ja hakea. Palvelu perustuu paikkatiedon mallintamiseen Suomalaisen paikkaontologian SUO:n mukaisesti.

Työ koostuu neljästä vaiheesta. Ensimmäisessä vaiheessa selvitetään menetelmä, jolla SUO-ontologia populoidaan paikkainstansseilla. Erityistä huomiota kiinnitetään paikkojen uniikkien tunnisteiden, URI:en luomiseen. Toisessa vaiheessa selvitetään, miten ontologian populointivaiheessa tuotetut paikkojen RDF-kuvaukset on tallennettava. Kolmannessa vaiheessa ratkotaan ontologisessa muodossa olevan paikkatietoaineiston suuren määrän tuomia ongelmia muun muassa kehittämällä paikkatiedon RDF-varastolle indeksointitietokanta nopeita hakuja varten. Neljännessä vaiheessa kehitetään rajapinta hakujen suorittamista varten sekä hakurajapintaa hyödyntävä graafinen, selaimessa toimiva käyttöliittymä.

Työ on tehty osana FinnONTO-projektia, jossa kehitettiin suomalaisiin olosuhteisiin räätälöityjä semanttisen webin ontologioita sekä näitä hyödyntäviä palveluita.

HELSINKI UNIVERSITY OF TECHNOLOGY **ABSTRACT OF MASTER'S THESIS**

Author, Name of the Thesis

Robin Lindroos

An Ontology-Based Service for Geographic Information

Date: 30.5.2008 Number of pages: 74

Faculty	Professorship
Faculty of Electronics, Communications and Automation	AS-75 Media Technology
Supervisor professor Eero Hyvönen	

Abstract

This master's thesis focuses on methods for converting materials containing geographic information into an ontology-based form. As a result of this thesis a service called ONKI-Geo has been developed for creating, handling and fetching geographic data in an ontlogy-based form. The service is based on representing geographic information in a form specified by the Finnish geo-ontology SUO.

The work consists of four phases. The first phase focuses on creating a method for populating the SUO-ontology with place instances. Special focus is targeted on the creation of unique identifiers, URIs, for the places. In the second phase a method is developed for storing the RDF descriptions created when populating the SUO-ontology with place instances. The third phase focuses on solving the problems caused by the large amount of RDF/XML data created and stored in the RDF database. A method is created for indexing the RDF data which will enable fast searches. In the fourth phase an interface is developed for searching and fetching geographic information. A browser-based graphical user interface that utilizes the search interface is also created.

This work has been done as part of the FinnONTO-project that focused on developing national semantic web ontologies and services based on the ontologies.

Alkusanat

Tämä diplomityö tehtiin osana FinnONTO-projektia Semanttisen laskennan tutkimusryhmässä Teknillisessä korkeakoulussa. Haluan etenkin kiittää professori Eero Hyvöstä mahdollisuudesta saada olla mukana tämän innovatiivisen ja inspiroivan tutkimusryhmän työssä. Kiitokset myös paikkapajan muille jäsenille Riikka Henriksonille ja Tomi Kauppiselle. Erityinen kiitos myös erityisen mukaville huonetovereilleni Eetu Mäkelälle, Jussi Kurjelle ja Reetta Sinkkilälle.

Helsingissä 30.5.2008

Robin Lindroos

Sisällysluettelo

1 JOHDANTO

Pakkatiedon hyödyntäminen on yhä enenevissä määrin arkipäivää lähes kaikille. Ajaessamme autoa käytämme navigaattoreita. Ennen ajoon lähtöä voimme tarkistaa tieolosuhteet[1], jotta tietäisimme kiertää ne tiet, joilla esiintyy häiriöitä. Internetissä voimme hakea aikatauluja joukkoliikenteen kulkuvälineille ilmoittamalla haluttu reitti paikasta A paikkaan B[2]. Voimme nälän tullen hakea läheltä ravintolaa kartan avulla[3]. Voimme jopa selailla koko maailmaa karttana ja satelliittikuvina maailmanlaajuisen karttapalvelun[4] avulla.

Mutta miten hakea nopein reitti lähimpään kiinalaiseen ravintolaan siten, että reitti ei kulje tietöiden hidastamien tieosuuksien kautta? Ainoa keino olisi ensin hakea lähin ravintola ja ravintolan osoite. Tämän jälkeen pitäisi hakea lyhin reitti ravintolan osoitteeseen ja verrata reittisovelluksen ehdottamaa reittiä Tiehallinnon häiriötiedostuksiin. Käyttäjälle tämä tarkoittaa monen erillisen palvelun käyttöä ja tiedon siirtämistä hakujen välillä palvelusta toiseen. Käyttäjälle vaivattominta olisi pystyä tekemään suoraan ravintolaoppaassa haku, joka palauttaa lyhimmän reitin ravintolaan ottaen automaattisesti huomioon tieosuudet, joilla esiintyy häiriöitä.

Ongelmana on se, etteivät eri palvelut välitä tietoa ja tarkemmin ottaen juuri paikkatietoa keskenään. Yksi syy voi olla se, etteivät palvelut yksinkertaisesti ole suunniteltu toimimaan keskenään. Toinen ehkä edellisenkin syyn taustalla oleva syy voi olla se, ettei paikkatiedon välittämiselle ole mitään yksinkertaista ja yhtenäistä tapaa, jota voitaisiin soveltaa kaikissa eri palveluissa. Paikkatiedon määrittely voi olla hyvinkin sovelluskohtaista sovelluksen käyttötarkoituksesta ja paikkatiedon tarkkuudesta riippuen. Jotta palvelut voisivat keskustella keskenään, tarvitaan palveluiden välille yhteinen kieli, jolla paikkatietoa välitetään. Tämän yhteisen kielen pitää olla mahdollisimman skaalautuva, jotta se sopisi monen eri tasoisen paikkatiedon kuvailemiseen.

Pitäisi olla olemassa palvelu, joka tarjoaisi eri sovelluksille ja palveluille paikkatietoa ja näille sovelluksille yhtenäisen tavan viitata paikkoihin. Paikkatieto ja viittaus paikkatietoon pitää olla samalla tavalla tallennettu kaikissa sovelluksissa tai vähintäänkin samassa muodossa haettavissa. Paikkatietoa voidaan siis ajatella eri resurssien metatietona, jolla resurssit eri sovelluksissa ja palveluissa linkittyisivät verkon kautta yhteen.

1.1 Tutkimuksen tavoitteet ja rajaus

Tässä diplomityössä selvitetään miten tällainen paikkatietoa jakava palvelu voitaisiin toteuttaa semanttisen webin teknologioita hyödyntäen, tarkastellaan paikkatietoa yleisellä tasolla ja lopuksi esitellään tämän diplomityön tuloksena kehitetty paikkatiedon ontologiapalvelu, ONKI-Paikka. Tällä palvelulla paikkatiedon kohteille eli paikoille

1 Tiehallinto, ajantasainen liikennetiedottaminen, http://alk.tiehallinto.fi/alk/

2 Reittiopas, http://www.reittiopas.fi/ tai Matka.fi, http://www.matka.fi/

3 Eat.fi, http://www.eat.fi/

4 Google Maps, http://maps.google.com/

luodaan kuvaukset semanttisen webin tekniikoita käyttäen ja luodaan paikoille globaalisti uniikit tunnisteet, joilla paikkoihin ja niihin liittyvään tietoon voidaan viitata sovelluksesta riippumatta yhteisellä tavalla.

Työssä selvitetään ensin, mitä paikkatieto on ja millä tavalla eri palvelut sitä yleensä hyödyntävät. Tärkeä osa työtä on selvittää, miten paikoille voidaan määritellä globaalisti yksilölliset tunnisteet, joiden avulla paikkoihin viitataan muissa semanttisen webin resursseissa. Tämän pohjalta kehitetään palvelu, joka semanttisen webin teknologioita hyödyntäen mahdollistaa paikkatiedon tallentamisen, ylläpidon ja hakemisen. Palvelun täytyy toimia webissä, jotta siitä olisi mahdollisimman laaja hyöty eri sovelluksissa ja jotta palvelu olisi globaalisti saavutettavissa. Paikkatiedon ylläpitoon ja hakemiseen kehitetään rajapinnat, joiden kautta sekä koneet että ihmiset pystyvät palvelua käyttämään.

Palvelun toiminta rajataan tässä työssä koskemaan paikkatiedon muuntamista ontologiseen muotoon sekä ontologisessa muodossa olevan paikkatiedon hyödyntämistä muissa semanttista webiä käyttävissä sovelluksissa. Paikkatietoaineisto rajataan käsittämään nimettyjä paikkoja, kuten ihmisen määrittelemiä maantieteellisiä ja hallinnollisia alueita maanosista kuntiin, ja maantieteellisiä kohteita kuten järviä, saaria ja soita.

Palvelun tärkeimpänä kohderyhmänä on määritelty eri resurssien annotointia suorittavat tahot, joilla on tarvetta lisätä resurssien metatietoihin viittauksia paikkatietoon. Annotoinnilla tarkoitetaan metadatan eli tietoa kuvailevan tiedon luontia. Annotoijien tehtävänä on etenkin oikean paikan tunnisteen löytäminen käytettäväksi metadatana jonkin resurssin kuvaamisessa. Esimerkiksi ravintoloiden hakupalvelun ylläpitäjät voisivat lisätä ravintolan metadataan ravintolan sijaintia kuvaavan tiedon, jossa käytettäisiin hyväksi paikan semanttisen webin tunnistetta. Tällaista tehtävää varten palvelun on tarjottava nopea ja helppokäyttöinen työväline paikkojen hakemiseen.

1.2 Tutkimuksen rakenne

Tutkimus aloitetaan luvussa 2 selvittämällä yleisellä tasolla, mitä paikkatieto on ja miten sitä hyödynnetään etenkin paikkoihin viitattaessa. Luvussa 3 tutustutaan semanttisen webin teknologioihin ja niiden hyödyntämiseen resurssien metatiedon kuvaamiseen ja hallintaan. Neljännessä luvussa tutkitaan miten paikkatieto ja semanttinen web voidaan yhdistää paikkatietoa tarjoavan palvelun pohjaksi. Luvussa 5 selvitetään paikkatiedon lähteitä ontologiapalvelun aineiston pohjaksi. Tämän jälkeen esitellään luvussa 6 edellisissä luvuissa tehtyjen selvitysten ja johtopäätösten perusteella kehitetty paikkatiedon ontologiapalvelu ONKI-Paikka. Lopuksi arvioidaan tuloksia ja selvitetään palvelun jatkokehitystarpeita.

2 PAIKKATIETO

Tämän työn tuloksena kehitettävä ONKI-Paikka perustuu paikkatiedon esittämiseen ja hakemiseen. Paikkatieto on tässä sovelluksessa nimettyjä maantieteellisiä paikkoja, joihin yleensä viitataan paikan nimellä tai disambiguoinnin tarpeessa myös laajemman paikan nimellä, jossa em. paikka sijaitsee. Tässä luvussa selvitetään, mitä paikkatieto tarkalleen ottaen on ja mitä käytäntöjä, suosituksia ja standardeja sen esittämiselle ja tallentamiselle on olemassa. Luvussa selvitetään myös, kuinka yleisesti nämä suositukset ja standardit ovat käytössä esimerkiksi Suomessa.

2.1 Paikkatiedon esittäminen

Sanastokeskus TSK:n laatima geoinformatiikan sanasto[5] määrittelee paikkatiedon ”tiedoksi kohteista, joiden paikka Maan suhteen tunnetaan”. Paikalla viitataan yleensä koordinaattien ja tietyn koordinaattijärjestelmän avulla ilmoitettuun sijaintiin Maan suhteen joko suoraan tai välillisesti toisen paikan sijainnin kautta. Ennen paikkatieto jaettiin sijainti- ja ominaisuustietoihin. Nykyisen käsityksen mukaan sijaintitieto on yhden tyyppinen ominaisuustieto paikalle. Muita ominaisuustietoja ovat paikantavat, ajoittavat ja temaattiset ominaisuustiedot. Paikantava ominaisuus on esimerkiksi kohteen osoite tai kiinteistötunnus, joka epäsuorasti määrittelee paikan sijainnin. Ajoittava ominaisuus voisi esimerkiksi olla kunnan perustamisvuosi. Temaattisilla ominaisuuksilla voidaan esimerkiksi kertoa, milloin jokin rakennus on rakennettu, kuinka monta kerrosta rakennuksessa on tai minkä värinen rakennus on. (Sanastokeskus TSK 2005)

2.1.1 Sijaintitiedot

Sijaintitiedot kuvailevat paikkatietokohteen geometriaa tai topologiaa. Geometria ilmoittaa kohteen muodon ja koon geometrisen primitiivin kautta. Primitiivejä ovat piste, viiva (tai jana) ja alue (tai monikulmio). Pistemäinen kohde voi olla esimerkiksi rakennus tai jokin maastokohde kuten hiidenkirnu. Viivamaisia kohteita ovat esimerkiksi tiet ja joet. Kunnat, järvet ja metsät voidaan puolestaan kuvata aluemaisina kohteina. Topologian avulla määritellään kohteiden sijaintisuhteet suhteessa toisiinsa. Usein viiva- ja aluemaisille kohteille on tarve määritellä myös yksittäinen koordinaattipiste, jolla kohteen sijainti voidaan ilmoittaa. Esimerkiksi luvussa 5.3 esiteltävässä Maanmittauslaitoksen Paikannimirekisterissä ilmoitetaan joen koordinaattipisteeksi joen suun koordinaatit.

Sijaintitieto esitetään joko kaksi- tai kolmiulotteisten koordinaattien avulla. Koordinaattitieto määrittelee kohteen sijainnin jossakin tunnetussa koordinaattijärjestelmässä, joka koostuu koordinaatistosta ja datumista. Koordinaatisto määrittelee akseliston, jonka datumi kiinnittää Maahan määrittelemällä koordinaatiston origon ja orientaation. Käytetty koordinaattijärjestelmä ja koordinaattipisteiden arvot määrittelevät siten yksiselitteisesti maantieteellisen kohteen sijainnin Maan suhteen. Suomessa on vuodesta 2006 alkaen ollut käytössä kansallinen EUREF-FIN koordinaattijärjestelmä, joka on realisaatio ETRS89-

5 Geoinformatiikan sanasto, http://www.tsk.fi/fi/info/GeoinformatiikanSanasto.pdf

järjestelmästä[6] (European Terrestrial Reference System). ETRS89 on koordinaattijärjestelmä, joka on luotu Euraasian mannerlaatan Euroopan puoleiselle kiinteälle osalle ja kiinnitetty mannerlaattojen sijaintiin vuonna 1989. ETRS89-järjestelmän realisoimiseksi kehitettiin EUREF89-koordinaatisto, jonka Suomelle tarkennettu versio EUREF-FIN on määritelty JUHTA:n suosituksessa JHS 154 (2006).

Suomessa julkisen hallinnon eri sektoreiden tietohallintoa koskevat standardit ja suositukset määrittelee Julkisen hallinnon tietohallinnon neuvottelukunta (JUHTA). JUHTAn toimesta on määritelty myös julkisen hallinnon suositukset JHS 153 (2006) ja JHS 154, joissa määritellään ETRS89-järjstelmään liittyvät yksityiskohdat kuten Suomessa käytettävät koordinaatit. Muun muassa Maanmittauslaitos noudattaa näitä suosituksia ja vuodesta 2006 eteenpäin laitoksen tuottamat kartat esittävät koordinaatit suositusten mukaisessa EUREF-FIN-muodossa.

WGS84 eli World Geodetic System 1984 on maailmanlaajuisesti yleisesti käytössä oleva koordinaattijärjestelmä. WGS84 on alun perin Yhdysvaltojen puolustushallinnon karttalaitoksen NIMA:n[7] määrittelemä järjestelmä, joka on dokumentoitu NIMA:n julkaisussa TR8350.2 (2000). Järjestelmän koordinaatisto perustuu muun muassa GPS-satelliittien ratatietoihin. Määritelmässä todetaan, että ETRF89, johon myös EUREF-FIN perustuu, voidaan pitää identtisenä WGS84-järjestelmän kanssa. Maanmittauslaitoksen mukaan ETRS89 yhtyy WGS84-järjestelmään alle metrin tarkkuudella[8]. Näin pieni ero ei käytännössä näy kartoissa, jolloin Suomessa käytettävä EUREF-FIN-karttakoordinaatisto voidaan pitää samana kuin WGS84.

2.1.2 Muut ominaisuustiedot

Maantieteelliseen kohteeseen liittyy usein muitakin kuin sijaintiin liittyviä tietoja. Näitä tietoja kutsutaan yleensä ominaisuus- tai attribuuttitiedoiksi. Ominaisuustiedot jaetaan usein neljään eri kategoriaan: yksilöivät, paikantavat, ajoittavat sekä kuvailevat tiedot.

Paikoille annettavat yksilöivät tiedot voivat olla nimi tai jokin muu ihmisen antama tunnus paikalle. Yksilöivät tiedot helpottavat etenkin ihmisten välisessä kommunikaatiossa kohteeseen viittaamista. Pelkät paikannimet ovat kuitenkin harvoin täysin yksilöiviä. Mikään ei estä sitä, että useammalla paikalla voisi olla sama nimi. Maanmittauslaitoksen paikannimirekisterin mukaan ainoastaan noin kolmasosa paikkojen nimistä on yksilöllisiä, eli vain yhdelle paikalle on annettu sama nimi. Myöhemmin luvuissa 4.3 ja 4.4 esitetään miten paikoille voidaan muodostaa täysin yksilölliset tunnukset semanttisen webin tarpeisiin. Siinä paikoille luodaan yksilölliset URI-tunnisteet, joita käytetään tiettyyn paikkaan viittaamiseen etenkin semanttisessa webissä.

Ihmisten välisen kommunikaation helpottamiseksi paikoille voidaan myös antaa helpommin ymmärrettävät paikantavat tiedot kuten osoitteet. Edellä esitetyt sijaintitiedot ovat myös paikantavia, mutta ihmisten väliseen kommunikaatioon koordinaatit ovat liian

6 European Terrestrial Reference System 89, http://etrs89.ensg.ign.fr/

7 National Imagery and Mapping Agency, http://www.nima.mil/

8 http://www.maanmittauslaitos.fi/Tietoa_maasta/Kartoitus/Koordinaatti_ja_korkeusjarjestelmat/

yksityiskohtaisia ja sen takia epäkäytännöllisiä. Osoite on usein jaettu hierarkiatasoihin, joiden avulla ihmisen on helpompi ymmärtää missä paikka sijaitsee. Voidaan esimerkiksi sanoa, että paikka sijaitsee Suomessa, tai tarkemmin jossain Suomen kunnassa ja siinä kunnassa olevassa katuosoitteessa. Paikoille voidaan myös antaa ajoittavia tietoja, kuten milloin paikka, kuten kunta, on perustettu tai lakkautettu (Kauppinen et al. 2008). Näiden lisäksi paikoille voi antaa kuvailevia tietoja, jotka antavat tietoa esimerkiksi paikan fysikaalisista tai tilastollisista ominaisuuksista, kuten kasvillisuudesta tai väkiluvusta.

2.2 Paikkatiedon tallentaminen

Paikkatiedon tallentamisen ytimenä on yleensä jonkinlainen paikkatietojärjestelmä (engl. geographic information system, GIS). Paikkatietojärjestelmä määrittelee yleensä tarkasti, miten paikkatieto tallennetaan ja esitetään. Monissa sovelluksissa paikkatieto ei kuitenkaan tarvitse olla niin yksityiskohtaisesti ja tarkasti esitetty kuin paikkatietojärjestelmissä. Esimerkiksi kirjastojen ja museoiden järjestelmissä riittää usein valtio ja kaupunki, jossa jokin teos tai esine on tehty.

Esimerkiksi Museovirasto on laatinut yksityiskohtaiset ohjeet siitä, miten paikkatieto tallennetaan esineiden tietoihin[9]. Pääasiallisesti paikkatieto tallennetaan paikkojen nimillä, ja tarkenteena on yleensä laajempi alue, jossa paikka sijaitsee, kuten kunta tai valtio. Pääasiallisesti esitysmuoto perustuu paikkojen nimiin sekä paikkojen hallinnolliseen ja topologiseen hierarkiaan, joiden avulla paikkoja diambiguoidaan. Paikkojen nimet tallennetaan ainoastaan suomen kielellä, jonka takia voisi olettaa, että paikkatiedon välittäminen kansainvälisten järjestelmien kesken on melkein mahdotonta. Luvussa 4.2 selvitetään tarkemmin tämän paikkatiedon yksinkertaistetun esitysmuodon ongelmia.

Oikeassa paikkatietojärjestelmässä tieto on tallennettu huomattavasti hienorakeisemmin. Open Geospatial Consortium on määritellyt paikkatiedon esittämiselle XML-pohjaisen esitysmuodon nimeltä GML, eli Geography Markup Language. GML on kehitetty yhteistyössä kansainvälisen standardointiorganisaation ISO:n, ISO/TC 211 Geographic information / Geomatics -komitean kanssa. Vuonna 2007 GML hyväksyttiin ISO:n standardiksi numerolla ISO 19136 (2007). GML:n pääasiallinen käyttötarkoitus on olla yhtenäinen mallintamis- tallentamis- ja siirtomuoto maantieteelliselle informaatiolle etenkin verkossa toimivien eri sovellusten välillä. Standardia käytetään muun muassa JPEG2000 kuvaformaatissa kuvaan liittyvän paikkatiedon tallentamiseen ja esittämiseen (OGC 2006).

Maanmittauslaitos noudattaa ensisijaisesti standardeja ja julkisen hallinnon suosituksia, joten myös GML on laitoksen sovelluksissa käytössä. JUHTA:n suosituksessa JHS 162 (2007) määritellään paikkatietojen mallinnus tiedonsiirtoa varten ja GML on tässä suosituksessa valittu paikkatiedon esittämismuodoksi ISO:n standardien mukaisesti.

Paikkatiedolle löytyy myös uusia mallinnustapoja, joita tällä hetkellä kehitetään. Yksi tällainen paikkatiedon esitysmuoto perustuu semanttisen webin teknologioihin. Paikkatiedon mallinnus voidaan tehdä paikkatiedon semantiikan lähtökohdista. Tällöin rakennetaan niin sanottuja ontologioita, joilla määritellään aihealueen käsitteistö ja sen

9 http://www.nba.fi/fi/paikkatiedot_tallennus

luokitus. Paikkatieto ja sen käsitteet määritellään ensin tarkasti, jonka jälkeen ontologiaa populoidaan instansseilla, eli yksilöillä, jotka ovat tiettyjen käsitteiden ilmentymiä. Tämän diplomityön perustana on paikkatiedon mallintaminen ja tallentaminen ontologioiden avulla. Diplomityön lähtökohtana on jo olemassa oleva paikkatieto-ontologia, joka tämän työn tuloksena populoidaan paikkainstansseilla eri paikkatiedon lähdeaineistoista. Ontologioista kerrotaan yleisellä tasolla enemmän luvussa 3.2 ja paikkatieto-ontologioista enemmän luvussa 4.1.

3 SEMANTTINEN WEB

Tässä luvussa käydään ytimekkäästi läpi semanttisen webin tekninen perusta, jolle tämä diplomityö rakentuu. Semanttiseen webiin liittyy useita määrityksiä, suosituksia ja standardeja, joiden tunteminen on oleellista tässä työssä tehtyjen ratkaisujen ymmärtämiseksi. Suurin osa teknologioista, joihin semanttinen web perustuu, on W3C:n[10] sekä IETF:n[11] käsialaa. Näitä teknologioita käyttäen, voidaan luoda WWW:n sisällön semanttinen kuvailu.

3.1 Perusteet

Kun WWW sai alkunsa 90-luvun alkupuolella, oli uuden tekniikan hyöty siinä, että saatiin mahdollisuus esittää informaatiota ja etenkin dokumentteja koko maailman kattavana verkkona. Perustana oli HTML-dokumentti, jonka syntaksin avulla oli mahdollista luoda sisältöä webiin ja luoda linkkejä muuhun sisältöön. Sisällön määrän kasvaessa syntyi hakukoneita, joilla valtavaksi kasvaneesta HTML-dokumenttien ja tiedostojen verkosta pystyi hakemaan haluamaansa tietoa. Nykypäivän hakukoneiden kanssa on kuitenkin ongelmia. Usein saadaan valtava määrä hakutuloksia, joista vain murto-osa on merkityksellisiä. Haku perustuu usein melkein täysin kirjaimellisesti juuri siihen sanaan, jota haussa käytettiin, eikä hakukone osaa palauttaa tuloksia esimerkiksi sanan synonyymien perusteella. (Antoniou & van Harmelen 2004)

Nykyinen WWW on suunniteltu ihmisten eikä koneiden luettavaksi (Hyvönen 2001). Jotta WWW-dokumenttiviidakkoon olisi mahdollista saada jonkinlainen järjestys on etenkin W3C:n johdolla kehitetty järjestelmä nimeltä semanttinen web[12]. Semanttinen web koostuu useasta eri spesifikaatiosta ja W3C:n suosituksesta, jotka yhdessä muodostavat perustan aineistojen ja resurssien semanttiselle kuvailulle. Tämän diplomityön kannalta keskeisimmät neljä teknologiaa semanttisen webin mahdollistamiseksi ovat URI, RDF, RDF Schema ja OWL.

URI, eli Uniform Resource Identifier (Berners-Lee et al. 2005) on kehitetty resurssien globaalin nimeämiskäytännön perustaksi. RDF, eli Resource Description Framework (Manola & Miller 2004) on standardi tietomalli ja syntaksi resurssien kuvailemiseksi. RDF Schema (Brickley & Guha 2004) on kehitetty standardiksi, jolla on mahdollista määritellä malli resurssien metatietojen kuvaamiseksi. OWL, eli Web Ontology Language (Dean & Schreiber 2004) on ontologioiden määrittelykieli.

URI on koko semanttisen webin verkkometaforan perusta. Kaikelle tiedolle voidaan luoda yksilöllinen tunniste, jolla tietoon voidaan viitata. Viitattavaa semanttisessa webissä kutsutaan resurssiksi. Resurssi voi olla mikä tahansa yksilöitävä fyysinen kohde, abstrakti käsite tai asia, johon voidaan viitata. Samalla tavalla kun ihminen on kielessään määritellyt kaikelle omat sanat, on myös semanttisessa webissä määriteltävä kaikelle ikään kuin sanat, URI:t, joiden avulla semanttisen webin laajuudessa voidaan asioihin viitata.

---

10 World Wide Web Consortium, http://www.w3.org/

11 Internet Engineering Taskforce, http://www.ietf.org/

12 W3C Semantic Web Activity, http://www.w3.org/2001/sw/

3.2 Ontologiat

Antoniou & van Harmelen (2004) mukaan termi ontologia juontaa juurensa filosofiaan. Ontologia on filosofian ala, joka tutkii olemassaolon luonnetta ja sitä mitä asioita on olemassa ja miten niitä voidaan kuvailla. Havainto siitä, että maailma koostuu kohteista tai olioista, joita voi ryhmitellä abstrakteihin luokkiin yhteisten piirteiden perusteella, on alunperin ontologiasta peräisin oleva malli.

Nykyisin tämä luokka- ja oliomalli on hyvin yleinen muun muassa ohjelmointikielissä ja muuallakin tietojenkäsittelytieteessä. Myös ontologia on terminä siirtynyt filosofeilta tietojenkäsittelijöiden käyttöön. Etenkin semanttisen webin yhteydessä ontologia tarkoittaa nykyään mallia, joka formaalilla tavalla kuvaa johonkin tiettyyn, rajattuun aihealueeseen liittyvää käsitteistöä. Yksinkertaistettuna ontologia koostuu rajatusta listasta käsitteitä sekä näiden käsitteiden välisistä suhteista. Aivan kuten oliot voidaan ryhmitellä luokkiin, voidaan myös luokat ryhmitellä ylemmän tason abstrakteihin luokkiin. Tästä syntyy luokkien ja alaluokkien hierarkia, joka on ontologioissa yleinen rakenne. Esimerkiksi abstrakti luokka *eläimet* voitaisiin jakaa alaluokkiin *nisäkkäät*, *matelijat*, *linnut, hyönteiset* jne. Linnut taas voitaisiin jakaa vaikkapa alaluokkiin *lokit*, *tikat*, *hanhet* jne. Lopputuloksena olisi eläinkuntaa kuvaava ontologia, jossa kaikki eläimet on luokiteltu ja luokkien väliset sukulaissuhteet on määritelty luokkahierarkian kautta.

Samalla tavalla voidaan myös maantieteellinen käsitteistö mallintaa formaalisti ontologian avulla, jossa alalla käytetyt termit ja niiden suhteet on tarkasti määritelty. Tämän diplomityön pohjana on Suomalainen paikkaontologia SUO (Henriksson et al. 2008), jota esitetään tarkemmin luvussa 4.1. Ontologia ei itsessään yleensä määrittele luokille yksilöitä, joita myös kutsutaan olioiksi tai instansseiksi. SUO-ontologia määrittelee abstraktin luokan nimeltä *paikka*, jonka alaluokat määrittelevät ontologian paikkatyyppihierarkian. Hierarkiasta löytyy muun muassa hallinnollisia alueita kuten *valtio* ja *kunta*. Tämän työn yksi tavoite on luoda SUO-ontologian määrittelemille paikkatyypin luokille paikkainstansseja käyttäen lähteenä nimettyjen paikkojen paikkatietoaineistoja.

3.3 FinnONTO-projekti

Tämä diplomityö on tehty osana FinnONTO-projektia[13] Semanttisen laskennan tutkimusryhmässä[14]. FinnONTO-projektin päämääränä oli kehittää suomalaisen webin ontologiaperustainen sisältöinfrastruktuuri. Käytännössä projektissa kehitettiin useita uusia ontologioita, ontologiapalveluita sekä niitä hyödyntäviä sovelluksia (Hyvönen et al. 2007). Tämän diplomityön tuloksena syntynyt ontologiapalvelu ONKI-Paikka, on yksi SUO-ontologiaa mutta myös SAPO-ontologiaa hyödyntävä sovellus. SAPO, eli Suomen ajallinen paikkaontologia (Kauppinen et al. 2008) on myös FinnONTO-projektissa kehitetty ontologia, jolla voidaan kuvailla paikoille ajan myötä tapahtuvia muutoksia, kuten hallinnollisten alueiden yhdistymisiä tai muita rajamuutoksia.

13 FinnONTO, http://www.seco.tkk.fi/projects/finnonto/

14 Semantic Computing Research Group, http://www.seco.tkk.fi/

4 PAIKKATIETO SEMANTTISESSA WEBISSÄ

Tässä luvussa selvitetään, miten edellisessä kahdessa luvussa esitetyt paikkatieto ja semanttinen web voidaan yhdistää paikkatietopalvelun perustaksi. Luvussa esitellään paikkatietoa määritteleviä ontologioita sekä tapoja, joilla paikkoihin, etenkin nimettyihin paikkoihin, voidaan viitata eri järjestelmissä. Lisäksi selvitetään, miksi paikkoihin viittaaminen on usein ongelmallista, ja esitetään miten semanttinen web voi tuoda ratkaisun tähän ongelmaan.

4.1 Maantieteelliset ontologiat

Kuten jokaisella tieteellisellä ja muullakin alalla, on myös maantieteessä oma sanastonsa ja termistönsä. Jotta maantieteellistä informaatiota voitaisiin muuntaa ontologiseen muotoon, tarvitaan alalle kehitetty ontologia.

Maantieteellisiä ontologioita tutkitaan ja kehitetään muun muassa Ateenan teknillisessä korkeakoulussa Geospatiaalisen ontologian tutkimusryhmässä[15]. Yhdysvaltain ilmailu- ja avaruushallinto NASA kehittää omia maantieteellistä ontologioita nimellä SWEET (Semantic Web for Earth and Environmental Terminology)[16]. Näiden ontologioiden avulla NASA pyrkii luomaan yhteisen kielen, jolla maantieteellistä informaatiota välitetään eri projektien välillä. Ison-Britannian maanmittauslaitos, Ordnance Survey, kehittää myös omia ontologioitaan paikkatiedon mallintamiseen[17]. Nämä ontologiat keskittyvät jokainen tiettyyn osa-alueeseen maanmittaustoiminnassa. Esimerkiksi rakennuksiin, vesistöihin ja hallinnollisiin alueisiin liittyvillä tiedoilla on omat ontologiansa Ordnance Surveyn semanttisissa järjestelmissä.

Myös W3C:llä on toimintaa maantieteellisen informaation mallintamisen saralla. W3C Geospatial Incubator Group[18] kehittää ja selvittää, miten nykyisen webin sisältämä maantieteellinen informaatio voitaisiin hyödyntää tehokkaammin luomalla geospatiaalinen ontologia tarkoitusta varten. Ontologia toisi osaksi ratkaisun siihen, miten nykyisen webin sisältöä voisi laajentaa ja tarkemmin luokitella maantieteellisellä informaatiolla, vaikkapa uusia hakumahdollisuuksia varten. Jos palveluiden maantieteellisestä sijainnista on saatavilla tarkat tiedot, voisi esimerkiksi olla mahdollista hakea kaikki itseään lähellä olevat palvelut. Voisi myös olla mahdollista hakea kaikki tiettyä paikkaa koskevat uutiset kaikista verkossa toimivista uutispalveluista, joilla sisältö on laajennettu geospatiaalisella lisäinformaatiolla.

Esseessään *John Wilkinsin analyyttinen kieli (El idioma analítico de John Wilkins)* vuodelta 1953, Jorge Luis Borges kirjoittaa:

---

15 Geospatial Ontology Research Group, http://ontogeo.ntua.gr/

16 Semantic Web for Earth and Environmental Terminology (SWEET), http://sweet.jpl.nasa.gov/

17 Ordnance Survey Ontologies, http://www.ordnancesurvey.co.uk/oswebsite/ontology/

18 W3C Geospatial Incubator Group, http://www.w3.org/2005/Incubator/geo/

> *”These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled Celestial Emporium of Benevolent Knowledge. On those remote pages it is written that animals are divided into: (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's-hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance”*

Listalla on tuskin mitään oikeaa todellisuusperää, mutta se onkin tarkoitettu hupaisalla tavalla havainnollistamaan sitä ajatusmaailman ja kulttuurin eroa, mihin länsimaiset tutkimusmatkailijat törmäsivät matkoillaan Aasiaan 1600 ja 1700-luvuilla. Tämä on hyvä esimerkki siitä, miten kulttuuririippuvaisia sanojen ja käsitteiden luokittelu ja semantiikka voi olla. Sama pätee myös maantieteellisten ontologioiden kehittämisessä. Ontologian rakenne ja sanasto perustuu usein kulttuurisin eroihin ja omasta näkökulmasta tehtyihin luokituksiin (Kavouras et al. 2005).

Maantieteellisiä ontologioita on useassa hankkeessa yritetty yhtenäistää esimerkiksi luomalla ontologioille yhteiset yläkäsitteet. Esimerkiksi hallinnolliset aluejaot voivat eri valtioiden välillä vaihdella huomattavan paljon. Yhdysvalloissa valtion sisäinen hallinnollinen aluehierarkia alkaa osavaltioista[19]. Suomessa sen sijaan hallinnollinen aluehierarkia alkaa lääneistä[20]. Osavaltio ja lääni eivät kuitenkaan hallinnollisina alueina vastaa toisiaan, joten maantieteellisissä ontologioissa ne eivät voisi olla ekvivalentteja luokkia. Ekvivalentti tarkoittaa tässä täydellisesti vastaavaa tai samanarvoista. Yhdysvalloissa osavaltiolla on esimerkiksi lainsäädännöllistä valtaa, kun taas Suomen lääneillä tällaista ei ole. Tämän takia osavaltiot ja läänit eivät hallinnollisina eliminä vastaa toisiaan.

Myös valtion sisällä voi paikallinen aluehallinto vaihdella esimerkiksi osavaltiosta toiseen. Yhdysvalloissa osavaltiot jaetaan 48:ssa osavaltiossa hallinnollisiin alueisiin nimeltä *county,* kun taas Luisianan osavaltiossa tätä aluehallinnon hierarkiatasoa kutsutaan nimellä *parish* ja Alaskassa nimellä *borough*[21].

Myös aluehallinnon hierarkiatasojen määrä vaihtelee maasta toiseen, minkä takia hallintotasoja on vaikeaa tai jopa mahdotonta verrata toisiinsa. Muun muassa tästä syystä Suomen maantieteellisten paikkojen ja hallinnollisten alueiden kuvailuun sopivaa ontologiaa ei ollut valmiina saatavilla, joten sellainen piti kehittää. Suomalainen paikkaontologia SUO (Henriksson et al. 2008) on osana FinnONTO-projektia kehitetty suomalaisiin maantieteellisiin olosuhteisiin sopiva ontologia. Tämän työn tuloksena kehitetty ontologiapalvelu, ONKI-Paikka, on kehitetty SUO-ontologiaa hyödyntäen.

---

19 CIA - The World Factbook -- United States, https://www.cia.gov/library/publications/the-world-factbook/geos/us.html

20 CIA - The World Factbook -- Finland, https://www.cia.gov/library/publications/the-world-factbook/geos/fi.html

21 Geographic Areas Reference Manual, http://www.census.gov/geo/www/garm.html

4.2 Paikkoihin viittaaminen

Semanttisen webin yksi lähtökohta on määritellä mahdollisimman formaalilla tavalla eri käsitteiden semanttiset suhteet toisiinsa (Manola & Miller 2004). Ongelma kuitenkin on usein käsitteiden merkityksen vaihtelu eri kielissä ja kulttuureissa. Maantieteellisten kohteiden määritteleminen ja niihin viittaaminen on tästä syystä usein kulttuuririippuvaista.

4.2.1 Paikan nimi

Useimmilla maantieteellisillä kohteilla on jonkinlainen ihmisen antama nimi. Nimeä käytetään lähes aina, kun puheessa tai kirjoitetussa tekstissä viitataan johonkin paikkaan. Nimillä viittaaminen on ihmiselle luonteva tapa puhua asioista. Puheessa ja tekstissä käy yleensä aina asiayhteydestä ilmi, mistä paikasta tarkalleen ottaen on kyse, jos saman nimisiä paikkoja sattuisi löytymään useita.

Paikkojen nimet ovat usein hyvin kulttuuririippuvaisia. Monilla paikoilla voi olla useita nimiä riippuen siitä, kuka tai ketkä paikalle on nimen antanut. Suomi on kaksikielisenä maana hyvä esimerkki siitä, kuinka paikoilla voi olla useita virallisia nimiä eri kielillä. Itse asiassa Suomessa virallisia paikkojen nimiä löytyy Lapissa jopa kolmella eri saamen kielellä: pohjoissaame, inarinsaame ja koltansaame (KOTUS 2007). Eri kielillä kommunikoidessa käytetään siis usein eri nimiä samoille paikoille. Tällöin käytetty kieli on tärkeä osa asiayhteyttä, jonka perusteella ihminen pystyy disambiguoimaan, mistä paikasta on kyse.

Maantieteellisiin kohteisiin ja paikkoihin viittaaminen kohteen nimellä on yleensä ongelmallista monestakin syystä. Jos asiayhteydestä ei erikseen käy ilmi mistä paikasta on kyse, voi pelkkä paikan nimi olla riittämätön määrite. Suurin ongelma on se, että saman nimisiä paikkoja löytyy yleensä useita. Luvussa 5.3 esitettävän Maanmittauslaitoksen Paikannimirekisterin aineistoa läpikäymällä voidaan todeta, että esimerkiksi Pyhäjärvi-nimisiä maantieteellisiä kohteita löytyy Suomesta vähintään 47 kappaletta. Olisi siis ongelmallista viitata Pyhäjärveen pelkällä paikan nimellä. Jos tarkemmasta asiayhteydestä ei ole tietoa, olisi mahdotonta tietää mihin näistä 47:stä Pyhäjärvestä viitataan.

Paikkojen nimet voivat myös muuttua ajan myötä (Kauppinen et al. 2008). Esimerkiksi hallinnollisissa alueissa tapahtuvat muutokset, kuten kuntaliitokset, muodostavat yleensä uusia paikkoja, joille saatetaan antaa aivan uudet nimet. Paikan ensisijainen nimi voi myös muuttua, jos kaksikielisen kunnan kielisuhteet muuttuvat. Muissa kielissä, kuten englannin kielessä, viitataan Suomessa oleviin paikkoihin sillä paikan nimellä, joka on kunnan virallisen enemmistökielen käytössä (KOTUS 2006). Esimerkiksi käytetään nimeä Ekenäs eikä Tammisaari, koska kaupungin enemmistökieli on ruotsi. Jos kielisuhteet muuttuvat ja Tammisaaren enemmistökieleksi tulisi suomi, muuttuu myös englanninkielisissä teksteissä viittaus tähän kaupunkiin. Paikan nimi ei siis ole ajan suhteen pysyvä.

Pelkällä paikan nimellä ei siis voida täydellä varmuudella päätellä mistä paikasta on kyse. Tästä syystä ei aina ole mahdollista viitata maantieteelliseen kohteeseen pelkällä paikan nimellä. Esimerkiksi semanttisessa webissä ei koneellisesti voi päätellä, mistä paikasta on kyse, ellei ole tiedossa jotain tarkempia määritteitä, jonka perusteella paikka voitaisiin

disambiguoida.

4.2.2 Alue- ja paikkahierarkia

Yleensä kohteeseen voidaan viitat yksiselitteisesti nimeämällä isompi kokonaisuus, jossa kohde sijaitsee. Voitaisiin siis viitata esimerkiksi Ulvilassa olevaan Pyhäjärveen. Tämä on hyvin yleinen tapa tehdä viittauksia paikkoihin ja se on käytössä muun muassa Suomen kirjastoissa, joissa julkaisujen kuvailuissa noudatetaan kansainvälisiä bibliografisen kuvailun ISBD-sääntöjä (Saur 2007). ISBD, eli International Standard Bibliographic Description on Kansainvälisen kirjastoseurojen liiton IFLA:n[22] laatima säännöstö erilaisten julkaisujen kuvailuun. Paikkatiedon, kuten kustannus- tai julkaisupaikan kuvaus tehdään ensisijaisesti pelkällä paikan nimellä. Mikäli paikan tunnistamisen kannalta on tarpeellista, lisätään paikan nimen perään suluissa tarkenteena esimerkiksi maan tai valtion nimi. Jos kustantajan tunnistamisen kannalta on tarpeellista, lisätään paikan nimen perään suluissa kustantajan koko osoite.

Tämäkin lähestymistapa paikkoihin viittaamisessa on varsin ongelmallinen koneellisen disambiguoinnin näkökulmasta. Ihminen kyllä osaa yleensä tulkita suluissa olevan tiedon tyypin, tiedon rakenteen ja merkityksen takia, ja pystyy siksi tulkitsemaan ISBD:n paikkamerkintöjä. Kone sen sijaan ei voi ilman lisätietoja tulkia onko suluissa valtion, läänin tai kunnan nimi tai osoite. Ongelmallista tässä paikkamerkinnässä on myös se, että suluissa oleva tarkenne voi paikka- ja aluehierarkiassa olla pääpaikan ala- tai yläpuolella. Jos esimerkiksi viittauksessa lukee ”Helsinki (Suomi)” on tarkenteena alue, jossa pääpaikka sijaitsee, ja viittaus kohdistuu loogisesti Suomessa olevaan kaupunkiin nimeltä Helsinki. Mutta jos viittauksessa lukee ”Helsinki (Kirkkokatu 6)”, on tarkenne paikan sisällä oleva pienempi alue tai paikka, ja loogisesti viittauksen kohde onkin paikka Kirkkokatu 6 eikä Helsinki. Koneellisen tulkinnan kannalta tämä paikkojen merkintätapa on siis varsin ongelmallinen. Olisi hyvin vaikeaa luoda koneelle säännöstö, jonka perusteella se voisi tulkita suluissa olevan tarkenteen semantiikan oikein.

Yleensä kuitenkin viittauksissa aluetarkenteena käytetään paikkahierarkiassa ylempänä olevaa paikkaa. Puhuttaessa Ulvilassa olevasta Pyhäjärvestä, on aivan selvää, mistä järvestä on kyse, sillä Ulvilassa on olemassa vain yksi järvi nimeltään Pyhäjärvi. Näin ei kuitenkaan aina ole. Mikään ei takaa sitä, ettei jossakin paikkahierarkiassa olisi olemassa useita saman nimisiä paikkoja suoraan ylemmän hierarkiatason alla. Toisin sanoen paikkahierarkiassa voi olla sisaruksia, joilla on samat nimet. Esimerkiksi Inarin kunnassa on jopa neljä järveä, joiden nimi on Pyhäjärvi. Jotta voitaisiin yksiselitteisesti viitata johonkin näistä neljästä Pyhäjärvestä, tarvitaan siis vielä tarkempi tapa viitata johonkin maantieteelliseen kohteeseen.

Myös alue- ja paikkahierarkiaviittauksissa käytetään paikan nimeä, jonka takia myös tämä viittaustapa kärsii samoista ongelmista kuin viittaaminen pelkällä paikan nimellä. Paikan nimi voi muuttua, mutta myös alue- ja paikkahierarkiassa voi tapahtua muutoksia (Kauppinen et al. 2008). Esimerkiksi kuntaliitosten yhteydessä paikkahierarkiassa ylempänä oleva paikka voi muuttua täysin toiseksi paikaksi. Vuonna 2007 esimerkiksi

22 International Federation of Library Associations and Institutions, http://www.ifla.org/

Toijala ja Viiala yhdistyivät Akaan kaupungiksi[23]. Jos paikkaviittauksissa on ennen vuotta 2007 käytetty Toijalaa tai Viialaa, pitää viittauksissa vuodesta 2007 lähtien käyttää Akaan kaupunkia.

4.2.3 Paikkatyyppi

Maantieteellisistä kohteista puhuttaessa on yleensä tarpeellista erotella, minkä tyyppisestä kohteesta on kyse. Usein paikan nimestä käy jo ilmi, minkä tyyppisestä paikasta on kyse. Esimerkiksi Pyhäjärvi on varsin helppo todeta vesistökohteeksi. Tämä tosin on jälleen selvä vain ihmisten välisessä kommunikaatiossa, sillä tiedämme, mitä sana järvi tarkoittaa. Kone sen sijaan ei tätä päättelyä pysty tekemään ellei sille ole erikseen opetettu sanojen semantiikkaa ja sen perusteella pystyisi päättelemään, että *järvi*-sanan sisältävät paikannimet viittaavat vesistökohteisiin. Tämäkin päättely olisi kuitenkin vain suuntaa-antava, sillä kaikki *järvi*-nimiset paikat eivät suinkaan ole vesistökohteita. Suomen kaupunki nimeltä Pyhäjärvi ei ole vesistökohde vaan hallinnollinen alue.

Paikkatyypin käyttäminen paikkaviittauksen tarkenteena voisi siis tietyissä tilanteissa toimia paikan disambiguoivana tarkenteena. Pyhäjärven kaupungista puhuttaessa on selvä mistä paikasta on kyse, sillä Suomessa on ainoastaan yksi tämän niminen kaupunki. Paikkatyyppi ei kuitenkaan auta Inarissa olevien Pyhäjärvien diambiguoimisessa. Kaikki neljä Inarin Pyhäjärveä ovat samantyyppisiä vesistökohteita. Myöskään tyyppi paikan tarkenteena ei siis takaa täysin varmaa viittausta tiettyyn paikkakohteeseen.

4.2.4 Koordinaatit

Viitattaessa Inarin kunnassa oleviin Pyhäjärvi-nimisiin maantieteellisiin kohteisiin pitäisi siis viittaukseen edelleenkin lisätä jokin tarkenne, sillä paikan nimi, paikan sijainti aluehierarkiassa ja paikkatyyppi eivät pystyneet näitä neljää Pyhäjärveä erottamaan toisistaan. Kuten luvussa 2.1 kerrottiin, jaetaan paikkatiedot usein sijainti- ja ominaisuustietoihin. Kaikki yllä luetellut paikan määrittävät tiedot ovat ominaisuustietoja, jotka eivät määrittele paikan maantieteellistä sijaintia. Paikan sijainnista aluehierarkiassa pystytään toki päättelemään suunnilleen missä paikka sijaitsee jos tiedetään missä aluehierarkiassa ylempänä oleva alue maantieteellisesti sijaitsee. Tämä ei kuitenkaan auta Inarissa olevien Pyhäjärvien disambguoimisessa.

Paikoille pitäisi siis lisätä sijaintitieto esimerkiksi maantieteellisten koordinaattien avulla. Koordinaatit ovat kaikille maantieteellisille kohteille lähes yksilöllisiä, koordinaattien tarkkuudesta riippuen. Lisäämällä koordinaatit Inarissa sijaitsevien Pyhäjärvien viittauksiin voidaan lopulta yksiselitteisesti viitata johonkin tiettyyn Pyhäjärveen Inarissa.

Koordinaattien ongelma on se, etteivät ne yksinään kerro paikasta juuri mitään muuta kuin että missä ne maapallolla sijaitsevat. Tarkkuudesta riippuen voi saman koordinaattipisteen kohdalla sijaita useita maantieteellisiä kohteita, jonka takia pelkästään koordinaattipisteen perusteella paikkaan viittaaminen on ongelmallista. Esimerkiksi jonkin hallinnollisen

23 Valtioneuvoston päätös Toijalan kaupungin ja Viialan kunnan lakkauttamisesta ja uuden Akaan kunnan perustamisesta (524/2006), http://www.finlex.fi/fi/laki/alkup/2006/20060524

alueen keskipiste saattaa hyvinkin osua samaan kohtaan jonkin maantieteellisen kohteen keskipisteen kanssa. Tällöin pelkästä koordinaattipisteestä ei voisi päätellä viitataanko esimerkiksi kuntaan tai järveen, joka sattuu sattumalta sijaitsee kunnan sisällä juuri samassa kohdassa kuin kunnan maantieteellinen keskipiste. Viitattaessa paikkoihin koordinaateilla, tarvitaan siis myös koordinaattien lisäksi jokin tarkentava määrite, jotta viittaus olisi täysin yksiselitteinen.

4.2.5 Sijainti- ja ominaisuustiedoista erillinen tunniste

Paikkaviittauksissa ihanteellisinta olisi tunniste, joka ei jättäisi lainkaan varaa tulkinnalle. Jotta yllä mainittuja paikkatietoja voisi käyttää viittauksissa, pitäisi ne kaikki lisätä viittauksiin. Tällainen tunniste, joka muodostuu useasta paikkaan liittyvästä erillisestä tiedosta on kuitenkin varsin ongelmallinen monestakin syystä. Paikan nimi olisi välttämätön osa tällaista yhdistelmätunnistetta, mutta yllä esitettyjen paikan nimiin liittyvien ongelmien takia, nimen käyttäminen pysyvissä paikkaviittauksissa olisi hankalaa. Paikkaviittauksessa käytettävän tunnisteen tulisi olla mahdollisimman muuttumaton ajan suhteen, jotta viittauksissa ei myöhemmin jäisi varaa tulkinnalle.

Semanttisen webin perusteknologiat, kuten RDF (Manola & Miller 2004) tarjoavat tähän ongelmaan ratkaisun. Kuten luvussa 3.1 kerrottiin, on yksi semanttisen webin kulmakivistä resurssien nimeäminen uniikilla tunnisteella, URI:lla (Berners-Lee et al. 2005). Tämän tunnisteen avulla voidaan viitata resursseihin yksiselitteisesti koko webin laajuudessa. Myös maantieteellisiin kohteisiin viittaaminen onnistuu URI:en avulla. URI voi olla paikalle täysin muista paikan ominaisuuksista riippumaton tieto, jonka takia URI ei kärsi edellä mainittujen paikkaominaisuuksien ongelmista paikkaan viittaamisen yhteydessä.

Iso haaste URI:en kanssa on kuitenkin se, että harvoilla paikoilla on missään valmiiksi määritelty URI. Tämä ei kuitenkin ole ylitsepääsemätön ongelma, sillä sellainen voidaan paikoille luoda varsin helposti. Seuraavissa luvuissa käydään läpi paikkojen URI:en luomiseen liittyviä haasteita ja esitellään lopuksi ONKI-Paikan URI-tunnisteiden luomiseen kehitetty säännöstö sekä perustelut tälle ratkaisuille.

4.3 URI-tunnisteet maantieteellisille kohteille

Koska semanttinen web on suhteellisen nuori ja vielä alkuvaiheissa oleva käsite ja teknologia, ei kaikelle vielä ole ehditty luomaan tunnisteita, joita voisi semanttisen webin viittauksissa käyttää. Tämä on myös totta suurimmalle osalle maantieteellisiä paikkoja. Yksi semanttisen webin vahvuuksista on kuitenkin se, että kuka tahansa voi laajentaa sitä ja lisätä siihen resursseja aivan kuten kuka tahansa voi laajentaa WWW:tä lisäämällä siihen omia dokumentteja. Näin ollen paikkaresursseja voidaan vapaasti lisätä yksinkertaisesti luomalla paikoille URI-tunnisteet, joita muut semanttisen webin käyttäjät voivat hyödyntää viittauksissaan. Semanttinen web, siihen perustuvat teknologiat ja rakenteet luovat kehyksen globaalille tunnistejärjestelmälle, jota voidaan hyödyntää myös paikkaresurssien yksilöimiseen. Tunnisteiden luomiseen on kuitenkin oltava yhteisesti käytössä oleva säännöstö, joka estäisi tai vähentäisi mahdollisuutta sille, että eri tahojen toimesta määritellään samoille paikoille useita erilaisia URI-tunnisteita.

URI-tunnisteissa on kuitenkin ihmiselle tiettyjä ongelmia. Kuten paikan nimi tai nimet, on myös URI ihmisen antama ominaisuustieto paikalle. Pelkästään paikkaa tarkastelemalla ei pystytä toteamaan, mitkä paikan nimet tai paikan URI ovat, tai edes sitä, onko paikalle jo jossakin määritelty URI. Jotta URI-tunnisteisiin perustuvia paikkaviittauksia voitaisiin helposti hallita ja käyttää, pitää niille luoda oma hallintaympäristönsä ja rekisterinsä. Myös uusien URI-tunnisteiden luomiseen pitää olla olemassa selkeä, mahdollisimman suoraviivainen ja yksinkertainen menetelmä. Seuraavaksi esitellään miten URI-tunnisteiden luominen paikoille on toteutettu ONKI-Paikassa.

4.4 URI-tunnisteet ONKI-Paikassa

Semanttisen webissä resurssien tunniste, URI, koostuu nimiavaruudesta ja lokaalinimestä muodossa:

<nimiavaruus><lokaalinimi>,

jossa <nimiavaruus> on jokin webin laajuudessa globaalisti uniikki URI, ja <lokaalinimi> nimiavaruuden alla uniikki tunniste, jonka syntaksiin perehdytään luvussa 4.4.2. Tämä juontuu semanttisen webin kuvailuteknologioiden, kuten RDF:n XML-syntaksista, jossa jokainen XML-dokumentin elementin nimi itse asiassa on URI, mutta normaalisti esitetty pelkällä lokaalinimellä lyhyemmän syntaksin takia (Bray et al. 2006). Suunniteltaessa URI-tunnistetta paikoille, on ensin päätettävä nimiavaruus tai nimiavaruudet, joiden alle SUO-ontologian paikkainstanssit osoitetaan. Tämän jälkeen pitää päättää minkälaisella lokaalitunnisteella paikkainstanssit nimetään nimiavaruuden alla. Sekä nimiavaruuden että lokaalinimen muotoon liittyy useita teknisiä seikkoja, joiden perusteella ONKI-Paikan palvelussa sijaitsevien paikkainstanssien URI-tunisteiden nimeämiskäytäntö päätetään.

Tärkein periaate paikan URI-tunnisteen luomisessa on se, että URI:n pitäisi ajan myötä olla mahdollisimman staattinen ja muuttumaton. Ideaalitapauksessa kerran luotu URI paikalle ei koskaan muutu. Tämä on myös alkuperäinen idea URI:lle, eikä alun perin hyvin suunnitellulle URI:lle ole koskaan tarvetta tehdä muutoksia (Berners-Lee 1998). Tämä on myös otettu tärkeimmäksi periaatteeksi paikkojen URI-tunnisteiden määrittämisessä.

4.4.1 Nimiavaruus

Koska ONKI-Paikka on aineistolähtöinen palvelu, eli paikkatietoa voi lisätä palveluun eri aineistoista, on URI-tunnistetta päätettäessä pohdittava, pitääkö paikan alkuperäisaineisto jollain tavalla näkyä URI:ssa. Yksi vaihtoehto olisi käyttää jokaiselle aineistolle omaa nimiavaruutta, jolloin paikan URI:sta kävisi selvästi ilmi se, mistä aineistosta paikkainstanssi on peräisin. Maanmittauslaitoksen Paikannimirekisteristä peräisin olevien paikkainstanssien nimiavaruus voisi esimerkiksi olla http://www.maanmittauslaitos.fi/onto/pnr/.

Toinen vaihtoehto on se, että nimiavaruus on sovelluskohtainen ONKI-Paikalle, jolloin nimiavaruudeksi luonnollisesti voisi ottaa SUO-ontologian nimiavaruus http://www.yso.fi/onto/suo/. Koska jokainen ONKI-Paikassa oleva paikkainstanssi on myös SUO-ontologian instanssi, olisi varsin luontevaa, että kaikilla paikoilla olisi sama nimiavaruus.

Nimiavaruuden valintaa tehtäessä löytyy myös teknisempiä seikkoja pohdittavaksi. IETF:n[24] määrittelemä standardi RFC3986 (Berners-Lee et al. 2005) määrittelee URI-tunnisteen geneeriseksi muodoksi seuraavan:

<scheme>:<hierarchical_part>[?<query>][#<fragment>],

jossa hakasuluissa olevat osat ovat valinnaisia. URI:n <scheme> on skeema, joka määrittelee spesifikaation, jonka mukaan resurssin nimeäminen tapahtuu. URI:n hierarkkinen osa <hierarchical_part> on varsinainen resurssin nimen osan, josta yleensä ilmenee resurssin omistaja tai hallinoiva taho. Esimerkiksi tässä diplomityössä käytettyjen URI-tunnisteiden hierarkkinen osa alkaa domainnimellä www.yso.fi, jonka perään lisätään polku resurssin nimeen domainin alla. URI:n <query> sisältää valinnaisia ei-hierarkkisia tunnistetietoja resurssille, joiden avulla tunnistetta voidaan tarkentaa. URI:n <fragment> osa on valinnainen tarkenne, jolla viitataan johonkin resurssin osaan. Tarkastellaan seuraavaksi kahta URI:a tämän standardin valossa:

1. http://www.yso.fi/onto/suo/id001
2. http://www.yso.fi/onto/suo#id001

Molemmat voisivat olla esimerkkejä jonkin paikkainstanssin URI:sta, jossa lokaalinimi on id001. Erona näissä on nimiavaruus, joka ensimmäisessä esimerkissä on http://www.yso.fi/onto/suo/ ja jälkimmäisessä http://www.yso.fi/onto/suo#. Semanttisen webin tasolla molemmat nimiavaruusvaihtoehdot ovat teknisesti yhtä oikeita. On täysin sovelluksen kehittäjän päätettävissä kumpaa vaihtoehtoa ryhtyy käyttämään. Internetin protokolla-tasolla kuitenkin näillä nimiavaruuksilla on yksi merkittävä ero. URI-standardi määrittelee, että fragmenttiosaa ei välitetä asiakassovelluksesta palvelimelle. Jos selaimella haetaan osoite http://www.yso.fi/onto/suo#id001, välittyy URI:stä palvelimelle vain ennen #-merkkiä oleva osa. URI:n fragmenttiosa on RFC3986-standardissa määritelty käytettäväksi vain asiakassovelluksessa.

URI, eli Uniform Resource Identifier, on nimensä mukaan vain tunniste, joka ei välttämättä osoita Internetissä mihinkään varsinaiseen sisältöön. URL, eli Uniform Resource Locator, sen sijaan on aina osoitin tai osoite johonkin tiedostoon (Berners-Lee et al. 2005), joka itsessään saattaa sisältää linkkejä toisiin tiedostoihin. Tällä tavalla syntyy dokumentti-tiedostojen välinen verkko. URL on yksi URI:n muoto. Toinen URI:n muoto on URN, jolla voidaan yksikäsitteisesti nimetä jokin tieto (Moats 1997). Semanttisen webin yhteydessä URI ei ole määritykseltään osoitin tiedostoon vaan johonkin semanttisen webin resurssiin, jota ei Internetin kautta pystytä siirtämään paikasta toiseen. Esimerkiksi semanttisessa webissä oleva paikkainstanssi on viittaus reaalimaailmassa olevaan paikkaan, eikä sitä tietenkään voida liikuttaa Internetissä bittivirtana. Semanttisen webin perusteknologiat tarjoavat kuitenkin jokaiselle resurssille esitysformaatin, jota voi siirrellä Internetissä bittivirtana, nimittäin resurssin RDF-kuvaus (Manola & Miller 2004).

Sauermann et al. (2006) kutsuvat ylhäällä esitettyä ensimmäistä URI-versiota 303-URI:ksi ja jälkimmäistä hash-URI:ksi, koska englanniksi #-merkki on nimeltään hash. 303 tulee HTTP/1.1-protokollan (Fielding et al. 1999) pyyntöjen statuskoodeista, jossa tällä koodilla

24 Internet Engineering Task Force, http://www.ietf.org/

tuleva vastaus palauttaa asiakassovellukselle haetulle osoitteelle uudelleenohjausta varten toisen osoitteen. Sauermann et al. (2008) esittelevät ideana sen, että jokaista semanttisen webin URI:a kohden voisi olla olemassa URL, joka viittaa resurssia kuvaavan RDF/XML-tiedostoon. URI siis pysyisi puhtaasti viittauksena resurssiin ja sitä vastaava URL olisi viittaus resurssia kuvaavaan tiedostoon. Tällä tavoin eri viittausten merkitys myös pysyisi loogisesti erillään. Onhan RDF-kuvauksen sisältävä tiedosto myös semanttisen webin resurssi.

Tällainen järjestely on mahdollista vain 303-URI:lla, joka voi hyödyntää HTTP-protokollan uudelleenohjausmekanismia. Ideana olisi siis se, että semanttisen webin resurssien nimiavaruudella olisi vastaava nimiavaruus resursseja kuvaaville RDF-tiedostoille. Jotta lokaalinimi säilyisi mukana uudelleenohjauksessa, pitää se nimenomaan olla osana 303-URI:a. ONKI-Paikan paikkainstansseilla voisi siis olla nimiavaruus http://www.yso.fi/onto/suo/, jota vastaisi RDF-tiedostojen nimiavaruus http://www.yso.fi/rdf/suo/. Kun palvelin saa käsiteltäväkseen esimerkiksi osoitteen http://www.yso.fi/onto/suo/id001, ohjaisi se asiakassovelluksen (esimerkiksi selaimen) osoitteeseen http://www.yso.fi/rdf/suo/id001, josta löytyisi RDF-kuvaustiedosto paikalle, jolla on lokaalinimi id001.

Myös hash-URI:n avulla olisi mahdollista hakea RDF-kuvaustiedosto paikkainstanssille. Silloin tosin palvelin joutuisi palauttamaan kaikkien nimiavaruuden alla olevien instanssien RDF-kuvaukset yhdessä tiedostossa, jotta asiakassovellus voisi URI:n fragmenttiosan perusteella poimia tiedostosta halutun instanssin kuvauksen. Paikkatietopalvelimessa tämä olisi kuitenkin mahdottomuus paikkainstanssien valtavan määrän takia. ONKI-Paikan instanssimäärä on useita miljoonia, jolloin yhden, kaikki instanssit sisältävän RDF-tiedoston kooksi tulee useita gigatavuja.

Koska ONKI-Paikan yhdeksi suunnittelulähtökohdaksi on otettu mahdollisuus hakea paikkainstanssin RDF-kuvaus URI:n kautta, on tässä sovelluksessa ainoa ratkaisu käyttää 303-URI-tunnisteita. Kuten tämän luvun alussa todettiin, on eri aineistoista peräisin oleville paikkainstansseille mahdollista antaa omat nimiavaruudet. Yksinkertaisuuden vuoksi on päätetty kaikille ONKI-Paikan paikkainstansseille käyttää yhteistä SUO-ontologian nimiavaruutta. Tällöin on helpompaa ylläpitää koko RDF-aineistoa yhdellä palvelimella yhden osoitteen kautta.

4.4.2 Paikkojen lokaalinimet

Lokaalinimen syntaksia määriteltäessä on otettava huomioon se, mitä muita lokaalinimiä saman nimiavaruuden alla tulee olemaan. SUO-ontologian nimiavaruudessa sijaitsee myös kaikki ontologian omien luokkien nimet. Luokkien lokaalinimet muodostetaan luokan suomenkielisestä nimestä, josta kirjaimet on muutettu pienaakkosiksi, diakriittiset merkit[25] on poistettu kirjaimista ja välilyönnit on korvattu alaviivoilla. Luokkien lokaalinimet on määritelty RDF Schema-määrityksen *rdfs:label*-ominaisuudella (Brickley & Guha 2004). Esimerkiksi http://www.yso.fi/onto/suo/hallinnollinen_alue on hallinnollisen alueen luokan

25 Diakriittiset merkit ovat perusaakkosiin yhdistettäviä merkkejä, jotka yhdessä muodostavat uuden aakkosen. Esimerkiksi diakriittinen merkki trema (¨) yhdistettynä a-kirjaimeen muodostaa ä-kirjaimen.

URI ja http://www.yso.fi/onto/suo/laani on läänin luokan URI. Tulevaisuudessa uusien luokkien ja instanssien lisääminen ei saa johtaa nimikonflikteihin nimiavaruuden lokaalinimissä. Instanssien nimet pitää siis valita sellaisiksi, etteivät ne missään tapauksessa voi saada samaa nimeä kuin jokin ontologian luokka.

Lokaalinimen muodostamiseen löytyy monia vaihtoehtoja. Nimen voisi muodostaa esimerkiksi paikan nimen perusteella. Silloin kuitenkin joudutaan taas samojen ongelmien eteen kuin paikkojen viittausten kanssa. Pohdittaessa viittausten muotoa päädyttiin ratkaisuun, jossa paras vaihtoehto olisi löytää viittausformaatti, joka ei perustu paikan sijainti- tai ominaisuustietoihin. Vaihtoehdoiksi jää jokin juokseva numerointi ONKI-Paikkaan lisättäville paikkainstansseille, tai jokin paikka-aineiston sisäiseen tunnisteeseen perustuva nimeäminen.

Juoksevan numeroinnin hyvä puoli on se, että se on täysin kaikista muista paikan ominaisuustiedoista erillinen tieto. Numeeristen tunnisteiden antaminen olisi myös helppoa, kunhan vain jatkuvasti pitää kirjaa siitä, mikä on viimeksi jollekin paikkainstanssille annettu numero. Tässä onkin juoksevan numeroinnin huonoin puoli, sillä silloin paikka-aineistojen lisäys paikkatietopalveluun pitäisi aina tapahtua keskitetyn palvelun kautta, jossa myös pidetään kirjaa numeroinnin kulusta. Tämä ei sovellu hajautetuille palveluille, jossa paikkainstansseja pitää voida luoda erillisiin aineistoihin, jotka kokonaisuudessaan tuodaan palvelun käyttöön.

Toinen huono puoli juoksevassa numeroinnissa tulee eteen silloin, kun jo tuotuja aineistoja päivitetään. Pitäisi olla järjestelmä, jolla paikka-aineistosta tuotu paikkatieto yhdistetään paikkapalvelussa olevaan paikkainstanssiin. Ainoa ratkaisu tähän olisi ylläpitää listaa tai muuta linkitystä, jossa paikka-aineiston sisäinen tunniste yhdistetään ontologiapalvelussa olevaan paikkainstanssin URI:in. Tällaisen listan ylläpitäminen olisi täysin mahdollista, mutta toisi koko ontologiapalveluun ylimääräisen tason hallittavaksi.

Juoksevan numeroinnin ylläpidon sijaan olisi mahdollista käyttää paikka-aineistojen sisäisiä tunnisteita. Jotta eri aineistoista peräisin olevista tunnisteista ei tulisi päällekkäisyyksiä, voisi lokaalinimessä käyttää esimerkiksi paikka-aineistokohtaista etuliitettä. Näin ontologiapalvelussa keskitetysti hallittavaksi tiedoksi jäisi pelkästään paikka-aineisto-kohtainen tunniste, jota käytettäisiin lokaalinimen etuliitteessä. Päivityksiä olisi helppo ajaa ontologiapalveluun, sillä paikkainstanssin URI olisi helposti muodostettavissa paikka-aineistolle annetun tunnisteen ja aineiston sisäisen paikkatunnisteen yhdistelmänä.

Semanttisessa webissä instanssin nimen syntaksi on sama kuin elementin nimellä XML-dokumenteissa. W3C:n suosituksen mukaan elementin nimi on määritelty seuraavalla tavalla: *”A Name is a token beginning with a letter or one of a few punctuation characters, and continuing with letters, digits, hyphens, underscores, colons, or full stops, together known as name characters. Names beginning with the string 'xml', or with any string which would match (('X'|'x') ('M'|'m') ('L'|'l')), are reserved for standardization in this or future versions of this specification.”* (Bray et al. 2006).

Tämän säännön mukaan päätettiin muodostaa ontologiapalvelun paikkainstanssin nimi seuraavalla tavalla:

<aineistotunniste>_<paikkatunniste>,

jossa <aineistotunniste> on ontologiapalvelussa annettu, kirjaimella alkava tunniste ja <paikkatunniste> on aineiston sisäisesti käytössä oleva tunniste, joka on tarvittaessa muunnettu sellaiseen muotoon, että paikkainstanssin nimi on yllä esitetyn XML-nimisäännön mukainen. Myöhemmin luvussa 5 esitetään tarkemmin, miten ONKI-Paikassa aineistoille annetaan tunnisteet ja miten eri aineistojen sisäiset tunnisteet muunnetaan yllä esitetyn XML-nimisäännön mukaiseen muotoon, jos sille on tarvetta. Aineistot, joissa sisäistä tunnistetta ei ole tai sitä on mahdotonta muodostaa, käytetään paikkainstansseille aineiston sisäistä juoksevaa numerointia. Tällöin aineiston ylläpitäjälle jää vastuu siitä, että juokseva numerointi toimii oikein myös silloin, kun aineistoa päivitetään.

Lokaalinimen luomisessa pyritään ONKI-Paikassa kolmeen päämäärään:

1. Nimi on mahdollisimman paljon erillinen paikan sijainti- tai ominaisuustiedoista, sillä nämä tiedot voivat ajan myötä muuttua paikkainstanssilla.
2. Nimi on riippumaton ulkoisista laskureista, eli tunniste olisi luotavissa paikan omien (meta)tietojen perusteella. Ulkoinen laskuri vaikeuttaisi paikkatietojen RDF-muunnosta hajautetussa järjestelmässä.
3. Nimi olisi mahdollisesti myös ihmisille sellaisessa ymmärrettävässä muodossa, että URI:a katselemalla voisi päätellä mistä paikasta tai muusta instanssista on kyse. Tämä voi olla ristiriidassa päämäärän 1 kanssa, mutta jos instanssille löytyy staattinen ominaisuus, joka ei instanssin elinkaaren aikana tule muuttumaan, voi sitä käyttää URI:n luomisessa.

Lokaalinimien luominen on erityisen haastavaa hallinnollisille alueille, joiden maantieteellinen kattavuus voi ajan myötä vaihdella. Otetaan esimerkkinä Suomi vuonna 2008 ja Suomi vuonna 1930. Useimmille olisi aivan selvää että molemmissa tapauksissa puhutaan samasta valtiosta, Suomesta. Semanttisen webin näkökulmasta asia ei kuitenkaan ole niin yksinkertainen. Sotien yhteydessä luovutettujen alueiden takia Suomi ei ole maantieteellisen kattavuuden osalta enää sama kuin vuonna 1930. Itsenäisenä valtiona, poliittisena toimijana muiden valtioiden joukossa, Suomi on kuitenkin edelleenkin sama valtio kuin vuonna 1930.

Semanttisessa mielessä poliittinen toimija Suomen valtio on siis sama entiteetti tänään kuin vuonna 1930. Mutta maantieteellisenä ja geospatiaalisena entiteettinä Suomi ei olekaan enää sama. Tämän takia maantieteellisessä ontologiassa Suomen valtiolla vuonna 1930 ja Suomen valtiolla vuonna 2008 pitää olla eri URI:t ja lokaalinimet. Tämä aineiston sisällä tapahtuva ajallinen muutos on jollain tavalla huomioitava myös paikkainstanssin URI:ssa.

ONKI-Paikassa paikkojen maantieteellisen kattavuuden muuttuminen ajan myötä on huomioitu lisäämällä valinnainen aikamääre lokaalinimeen paikan olemassaolon aloitusajasta, jolloin lokaalinimestä muodostuu seuraavanlainen:

<aineistotunniste>_<paikkatunniste>[_<aikamääre>]

Tässä <aikamääre> on valinnainen lisä jo aiemmin esitettyyn lokaalinimeen. Aikamääreen muoto on VVVVKKPP, jossa VVVV on vuosiluku, KK on kuukausi kahdella numerolla ja

PP on päivä kahdella numerolla. Esimerkiksi Moskovan välirauhan alueluovutusten jälkeisen Suomen lokaalinimi on tämän säännön mukaan A0003_FI_19440919. Suomen rajat muuttuivat vielä vuonna 1947, jolloin Inarin kunnassa sijainnut Jäniskosken-Niskakosken alue myytiin Neuvostoliitolle 15. heinäkuuta[26]. Nykyinen Suomen lokaalinimi on siis A0003_FI_19470715. Valtioiden URI-tunnisteista tarkemmin luvussa 5.2.3.

4.4.3 Muiden paikkatietoresurssien lokaalinimet

SUO-ontologiassa on paikkojen lisäksi myös muita yksilöitäviä resursseja. Esimerkiksi paikkojen koordinaattipisteet ovat resursseja, joille on luotava URI:t. Paikka-aineistoille on varattu aineistokoodit, jotka alkavat kirjaimella A. Paikka-aineistoista louhittavat muut resurssit, jotka eivät ole paikkoja, saavat aineistokoodin, joka alkaa kirjaimella B. Koordinaattipisteet muodostavat oman resurssiaineistonsa, jonka aineistokoodiksi on valittu B0001. Nämä B-aineistot eivät ole varsinaisia aineistoja kuten paikka-aineistot, sillä aineistojen resursseista ei ONKI-Paikan toteutuksessa löydy erillisiä RDF-tiedostoja kuten paikoista. Nämä paikoista erilliset aineistot luodaan lähinnä sen takia, että niiden sisältämien resurssien URI-tunnisteiden luominen olisi jollain tavalla järjestelmällistä.

Koordinaattipisteiden URI:t muodostetaan yksinkertaisesti yhdistämällä koordinaattien edustaman koordinaattijärjestelmän tunnus ja koordinaattipisteet:

<järjestelmätunnus>_<koordinaatti1>_<koordinaatti2>[_<koordinaatti3>],

jossa <järjestelmätunnus> on käytetyn koordinaattijärjestelmän EPSG-tunnus[27] ja koordinaatit 1-3 ovat koordinaattijärjestelmässä käytetyt arvot koordinaattipisteen määrittelemiseksi. Kolmiulotteisissa koordinaattijärjestelmissä <koordinaatti3> edustaa yleensä korkeutta esimerkiksi merenpinnan yläpuolella. ONKI-Paikan sisäisesti käytössä olevan WGS84-koordinaattijärjestelmän EPSG tunnus on 4326. Tässä koordinaattijärjestelmässä <koordinaatti1> on leveysaste, <koordinaatti2> pituusaste ja <koordinaatti3> korkeus.

Kaikki koordinaattiarvot ilmoitetaan desimaaleina, joissa on käytetty pistettä desimaalilukujen erottimena. Tällä tavalla saadaan uniikit URI:t jokaiselle koordinaattipisteelle. Pisteen tarkkuus, eli desimaalien määrä koordinaateissa muuttaa myös koordinaattipisteen URI:a. GEOnet Names Server -aineisto, joka esitetään luvussa 5.5, antaa Helsingille koordinaatit 60.175556 leveysaste ja 24.934167 pituusaste WGS84-koordinaattijärjestelmässä. Korkeudesta merenpinnan yläpuolella ei mainita mitään. Tästä muodostuu kaksiulotteinen koordinaattipiste, jonka lokaalinimi on yllä esitetyn säännön mukaan B0001_4326_60.175556_24.934167.

4.5 URI-viittausten heikkoudet ja vahvuudet

URI-tunnisteilla hyvät ja huonot puolensa, joiden välillä semanttisen webin sovellusten suunnittelijoiden on tasapainoteltava. Jo URI:n luonti saattaa olla vaikeaa, jotta se täyttäisi kaikki pysyvyyden ja luettavuuden vaatimukset. Lähtökohtana URI on kuitenkin varsin mainio viittaustapa laajan tukensa ansiosta. Suuri osa informaatiosta välittyy nykyisin

26 http://www.finlex.fi/fi/sopimukset/sopsteksti/1947/19470009/19470009_2

27 European Petroleum Survey Group, http://www.epsg.org/

verkon yli, ja yhtenä osoitemuotona on juuri URI. Tämän takia verkossa toimivat sovellukset osaavat jo käsitellä URI-tunnisteita verkkoresurssien osoitteina.

4.5.1 Ylläpidettävyys

Sekä ajallisesti että spatiaalisesti staattisille resursseille on varsin yksinkertaista antaa pysyvät URI:t. Kuten Berners-Lee (1998) esittää on URI:n määrittelemisessä otettava huomioon tunnisteen pysyvä luonne. URI luodaan kerran ja sitä voidaan sen jälkeen käyttää tunnisteena resurssille niin kauan kuin resurssiin on tarve viitata. Hyvin suunniteltua URI:a ei siis ole ikinä tarvetta muuttaa, ellei resurssi itsessään muutu, niin että se voidaan luokitella uudeksi resurssiksi, joka vaatii oman URI:nsa.

URI:en ylläpidon kannalta tämä tarkoittaa sitä, että URI:n luomisvaiheeseen on kiinnitettävä erityistä huomiota, sillä kerran luotu ja julkaistua URI:a ei voida enää muuttaa. Tämä on totta myös tämän työn tuloksena luoduille paikkojen URI-tunnisteille. Sen jälkeen kun URI:t on julkaistu ja otettu käyttöön muissa sovelluksissa, kuten annotaatiojärjestelmissä, ei URI:a voida enää ongelmitta muuttaa. Jos kuitenkin myöhemmin ilmenee pakottava tarve muotoilla paikan URI uudelleen on ainoa keino luoda paikalle kokonaan uusi URI ja määritellä se samaksi kuin vanha URI. Tästä enemmän alempaa luvussa 4.5.3.

URI:n pysyvyydellä varmistetaan se, etteivät resurssien annotoinnit lakkaa myöhemmin toimimasta, jos annotoinnissa käytetyn resurssin URI muuttuu. Vanhalle URI:lle pitää aina olla saatavilla uusi URI jonkin mekanismin kautta.

4.5.2 Huono luettavuus

Selvästi huono puoli URI-tunnisteissa on se, etteivät ne välttämättä sisällä ihmiselle suoraan ymmärrettävää informaatiota. URI:t ovatkin luotu koneiden välillä tapahtuvaan kommunikointiin ja semanttisessa webissä olevien resurssien yksilöimiseen. Tunnisteiden ei siis tarvitsekaan sisältää ihmiselle ymmärrettävää informaatiota, koska se ei olekaan niiden tarkoitus. Semanttisen webin sovellusten suunnittelijoilla on kuitenkin vapaus määritellä resursseille sellaisia URI-tunnisteita, jotka sisältävät ihmiselle suoraan ymmärrettävää informaatiota. URI:n lokaalinimi voisi esimerkiksi olla muodostettu resurssin nimestä. Tämä ei kuitenkaan ole suositeltavaa, sillä URI:n tulisi myös olla kieliriippumaton. Jos esimerkiksi kiinalainen loisi resursseille URI:t ja nimeäisivät ne resurssien kiinankielisten nimien mukaan, eivät ne suurimmalle osalle suomalaisista olisi yhtään sen luettavampia kuin jos URI:t olisi muodostettu numeerisista tunnisteista.

4.5.3 Ontologiset suhteet

Yksi URI-viittausten vahvuuksista on niiden käyttö semanttisen webin resurssien nimeämisessä. Tämä mahdollistaa esimerkiksi semanttisten suhteiden muodostamisen URI:lla nimettyjen resurssien välille. Jos URI myöhemmin muutetaan jollekin paikkainstanssille, on se mahdollista liittää vanhaan URI:in esimerkiksi *owl:sameAs*-ominaisuudelle (Dean & Schreiber 2004). Tämä ominaisuus määrittelee sen, että molemmat URI:t viittaavat samaan instanssiin. URI:t saatetaan joskus joutua muuttamaan kokonaiselle paikka-aineistolle,

jolloin on oltava olemassa mekanismi ilmaista annotaatiojärjestelmille että niissä käytetyt vahat URI:t ovat vanhentuneet ja korvattu jollakin toisella.

Näiden syiden takia voidaan URI:en käyttöä myös paikkojen tunnisteina perustella hyväksi ratkaisuksi. Kun paikoille on määritelty URI-tunnisteet, voidaan paikkoihin helposti viitata kaikissa niissä järjestelmissä, joissa URI:a käytetään resurssin tunnisteena.

4.6 Paikkainstanssi

Yksi vaikeimmin ratkottavista pulmista tässä työssä on määritellä se, mikä on paikka-instanssi ja milloin paikkainstanssissa tapahtuvat muutokset (Kauppinen et al. 2008) ovat niin merkittäviä, että lopputuloksena on aivan uusi paikkainstanssi. Kuten luvussa 4.2.5 esitettiin, on semanttisen webin näkökulmasta ainoa staattinen tieto instanssista sen URI. Kaikki muut instanssin ominaisuudet voivat ajan myötä muuttua. Tämä ei kuitenkaan päde semanttisen webin sovellustasolla, jossa jotkin instanssin ominaisuuksien muutokset saattavat sovelluksen näkökulmasta muuttaa instanssin luonnetta niin huomattavasti, ettei sitä enää voida pitää samana instanssina.

Paikkatietosovelluksen kohdalla on määriteltävä, mitkä paikkojen ominaisuudet ovat sellaisia, että niiden arvojen muuttaminen muuttaa paikkainstanssia huomattavalla tavalla. Toisin sanoen on määriteltävä se, mikä tai mitkä ominaisuudet yksilöivät paikkainstanssit toisistaan ajallisessa ulottuvuudessa. Monet paikkatyypit ovat ajan myötä varsin staattisia. Esimerkiksi luonnolliset maantieteelliset kohteet muuttuvat yleensä ihmisen näkökulmasta niin hitaasti, että niitä voidaan pitää ajan suhteen vakioina. Tällaisia kohteita ovat esimerkiksi vuoret tai järvet. Tuhansien tai miljoonien vuosien saatossa nämäkin kohteet muuttuvat, mutta paikkatietosovelluksen elinkaareen suhteutettuna muutos on niin hidas, että sillä ei ole merkitystä.

Sen sijaan ihmisen luomien tai määrittelemien paikkojen ja alueiden elinikä on paikkatieto-sovelluksen näkökulmasta varsin lyhyt. Tällaisten alueiden kohdalla on sovellustasolla huomioitava se, että paikkoihin voi ajan myötä tulla muutoksia. Ihmisen määrittelemät ja usein muuttuvat paikkatyypit ovat etenkin hallinnolliset alueet. Näitä ovat esimerkiksi valtiot ja niiden sisäiset hallinnolliset jaot, kuten läänit ja kunnat. Maailmanlaajuisesti nämä poliittiset ja hallinnolliset jaot ovat jatkuvassa muutostilassa tai ainakin käytännössä on oletettava näin olevan.

Etenkin hallinnollisten alueiden ominaisuudet voidaan karkeasti jakaa maantieteellisiin ja poliittisiin tai kulttuuririippuvaisiin ominaisuuksiin. Esimerkiksi alueen koordinaatit, rajat ja maantieteellinen kattavuus ovat selvästi maantieteellisiä ominaisuuksia. Nämä ominai-suudet vaikuttavat muihinkin paikkainstansseihin esimerkiksi SUO-ontologian suhteen *suo:isPartOf* kautta, jolloin ylemmällä paikkahierarkiatasolla jokin alue perii hierarkiassa alempana olevien paikkojen maantieteellisen kattavuuden. Toisaalta taas alueiden nimet tai hallintomuoto, joka määrittelee paikan tyypin, ovat poliittisia ominaisuuksia, jotka muuttu-essaan eivät välttämättä vaikuta alueen maantieteelliseen kattavuuteen tai paikkahierar-kiaan millään tavalla.

Tämä poliittinen ja maantieteellinen erottelu on avain myös ONKI-Paikan paikka-instanssien määrityksessä. Lokaalinimen määrittelyssä on ajallinen ulottuvuus otettava huomioon URI:en muodostamisessa esimerkiksi hallinnollisille alueille (Kauppinen et al. 2008). Koska URI:n muodostamiseen käytetään etenkin valtioiden kohdalla niiden poliittiseen toimijaan viittaavaa tunnusta, ei sitä voi yksinään käyttää maantieteellisen kohteen URI:n muodostamiseen. Valtioaineistossa lokaalinimi A0003_FI siis viittaa Suomen valtioon kautta aikojen, kun taas viitattaessa Suomeen maantieteellisenä alueena on lokaalinimeen lisättävä aikamääre. Tämä on ensisijaisen tärkeää etenkin annotoinnissa, jossa viitataan maantieteellisiin kohteisiin pelkillä URI-tunnisteilla. URI:n pitää silloin yksilöidä valtio myös maantieteellisen kattavuuden osalta, joka on mahdollista ainoastaan jos jokaisen rajasiirron yhteydessä luodaan uusi valtioinstanssi.

5 LÄHDEAINEISTOT

Paikkatietopalvelun ehkä tärkein komponentti on lähdeaineistot, sillä ilman paikkatietoa, palvelusta ei ole hyötyä kenellekään. ONKI-Paikan lähteiden hankkimisessa pyrittiin maailmanlaajuisesti mahdollisimman kattavaan aineistoon nimetyistä paikoista, olkoon ne luonnollisia muodostelmia tai ihmisen luomia poliittisia tai hallinnollisia alueita. Lähteissä tärkeää on myös tieto siitä, missä kohtaa aluehierarkiaa jokin nimetty paikka sijaitsee. Esimerkiksi tieto siitä, missä valtiossa tai missä valtion sisäisessä hallinnollisessa alueessa jokin paikka sijaitsee, on tärkeä tieto paikasta, mutta myös hyödyllinen tieto muun muassa paikkahakujen rajausehdoissa.

5.1 Aineistot ontologiapalvelussa

ONKI-Paikan aineistot on pyritty hakemaan mahdollisimman virallisista ja laajasti käytetyistä lähteistä. Suomessa paikkatietoa ylläpitää muun muassa Maa- ja metsätalousministeriön alaisuudessa toimiva Maanmittauslaitos[28], joka tuottaa karttoja ja paikkatietoa Suomesta. Muut lähteet Suomesta käsittelevät hallinnollista aluejakoa, joka on määritelty muun muassa laeissa. Maailmanlaajuisesti kattavimmat ja vapaasti käytettävät aineistot ovat Yhdysvaltojen valtiollisten virastojen tuottamia paikkatietoaineistoja.

Yksi huomattavan laaja ja jo semanttisen muodossa oleva paikkatietoaineisto on GeoNames[29], jonka paikkatietopalvelu on vapaasti käytettävissä selaimen kautta. Palvelun pääasialliset lähteet ovat muun muassa Geographic Names Information System[30] ja GEOnet Names Server[31], jotka molemmat ovat lähteitä myös tämän työn tuloksena tehdyssä ONKI-Paikassa ja käsitellään tarkemmin luvuissa 5.5 ja 5.6. GeoNames yhdistelee olemassa olevia aineistoja, mutta toimii tämän lisäksi wiki-tyylisesti, joka tarkoittaa sitä, että kuka tahansa voi lisätä ja muokata palvelussa olevaa paikkatietoa.

Wiki-tyylisen sisällöntuottamisen takia GeoNames-palvelun sisältämää paikkatietoa ei voida pitää virallisena ja täysin luotettavana, sillä palvelussa olevien wiki-tyylisesti lisättyjen paikkatietojen oikeellisuuden tarkistaminen on lähes mahdotonta. Palvelu sisältää myös suurimmaksi osaksi paikkatietoa, joka on jo ONKI-Paikan käytössä suoraan alkuperäisten lähteiden kautta. Näistä syistä GeoNames paikkatietoaineistoa ei ole otettu mukaan yhdeksi aineistoksi ONKI-Paikkaan.

Koska GeoNames on palveluna hyvin suosittu ja jatkuvasti kasvava, on tulevaisuudessa myös tämä huomioitava jollakin tavalla. Koska palvelu on toteutettu semanttisen webin teknologioilla, on jokaisella palvelussa olevalla paikalla URI. Tämän työn jatkokehityksenä voisi olla tarpeen luoda menetelmä, jolla määritellä *owl:sameAs*-suhteita ONKI-Paikan ja GeoNames-palvelun paikkainstanssien välille.

---

28 Maanmittauslaitos, http://www.maanmittauslaitos.fi/

29 GeoNames, http://www.geonames.org/

30 Geographic Names Information System (GNIS), http://geonames.usgs.gov/domestic/index.html

31 NGA GEOnet Names Server (GNS), http://earth-info.nga.mil/gns/html/index.html

5.1.1 Aineistotunniste

Monessa eri paikkatiedon lähdeaineistossa saattaa esiintyä viittauksia samoihin paikkoihin. Paikkojen samanlaisuus ei kuitenkaan aina ole täysin yksiselitteinen asia. Usein paikka-aineistoja on laadittu erilaisista lähtökohdista ja paikkojen luokitukset vaihtelevat aineistosta toiseen. Joissakin aineistoissa luokitukset voivat olla paljon hienojakoisempia kuin toisissa. Esimerkiksi GEOnet Names Server -aineistossa on paikat luokiteltu noin 650:een paikkatyyppiin, kun taas Maanmittauslaitoksen Paikannimirekisterissä, joka esitellään luvussa 5.3, on paikat luokiteltu vain 52:een paikkatyyppiin. Näiden eroavaisuuksien takia aineistot päätettiin pitää erillään ONKI-Paikassa, eikä eri aineistoista löytyviä paikkoja pyritä yhdistämään yhdeksi paikkainstanssiksi ontologiapalvelussa. Eri aineistoista peräisin olevia paikkoja ja niihin liittyviä tietoja pitää myös voida hakea erikseen.

Paikan URI:n muodostamista ja paikkainstanssien alkuperäaineiston merkitsemistä varten tarvitaan aineistoille sisäiset tunnisteet ontologiapalvelussa. Sisäisesti palvelussa riittää, että aineisto tunnistetaan pelkällä numerolla. Aineistoille voisi esimerkiksi antaa järjestysnumeron sitä mukaan kun aineistoja otetaan käyttöön palvelussa. Aineistotunnisteen alussa pitää kuitenkin olla kirjain, sillä tunniste on myös paikkainstanssin lokaalinimen alkuosa, ja kuten luvussa 4.4.2 kerrottiin on lokaalinimen aina alettava kirjaimella. Kirjaimeksi voidaan esimerkiksi valita aakkosten ensimmäinen kirjain A. Myöhemmin on mahdollisuus lisätä myös muita aineistosarjoja, joiden tunniste alkaa jollain toisella kirjaimella tai kirjainyhdistelmällä.

Kirjainta seuraavasta numerosta tehdään numeerinen merkkijono nollatäytöllä (engl. zero padding). Tämä tehdään sen takia, että lokaalinimet ja niillä muodostettavat URI:t olisi helpompi eri sovelluksissa järjestää aakkosjärjestykseen aineistotunnisteen mukaisesti. Oletetaan esimerkiksi, että aineistot järjestysnumeroilla 2 ja 15 halutaan järjestää aakkosellisesti oikeaan järjestykseen niin, että pienimmällä järjestysnumerolla olevat aineistot tulevat ensin. Jos aineistotunnisteet olisivat A2 ja A15, tuottaisi aakkosellinen järjestäminen väärän lopputuloksen. Mutta jos tunnisteet olisivat A0002 ja A0015, menisivät tunnisteet oikeaan järjestykseen.

Nollatäyttöä käytetään myös niissä lokaalinimissä, jotka muodostuvat jonkin aineiston sisäisestä juoksevasta numeroinnista. Jos numerointi alkaa esimerkiksi numerosta 1, tulee myös lokaalinimissä sama aakkosellinen järjestämisongelma eteen, ellei numeerisissa tunnisteissa käytetä nollatäyttöä. Aineistoissa, joissa käytetään sisäisesti tämän tyyppistä juoksevaa numerointia, on ensin määriteltävä aineiston instanssien mahdollinen määrä tulevaisuudessa. Jos instanssimäärä tulee olemaan enintään muutama sata tuhatta, riittä että nollatäytössä muodostetaan kuusinumeroinen merkkijono.

5.1.2 Kielten nimet

Paikoilla on usein eri kielillä eri nimiä, joista yleensä yksi on paikan ensisijainen nimi siinä valtiossa tai valtion osassa, jossa paikka sijaitsee. SUO-ontologiassa paikkojen nimet merkitään SKOS Core-sanaston luokkaominaisuuksilla *skos:prefLabel* ja *skos:altLabel* (Miles & Bechhofer 2008), joiden kieli määritellään *xml:lang*-attribuutin avulla.

Tämän attribuutin arvo on kielikoodi, jonka muodon määrittelee IETF:n BCP 47. BCP, eli Best Current Practise, on suositus, jota tällä hetkellä vastaa dokumentti RFC4646 (Phillips & Davis 2006). RFC, eli Request for Comment, on sarja IETF:n määrittelydokumentteja, joita jatkuvasti, tarpeen mukaan päivitetään. Päivityksen yhteydessä muuttuu myös RFC numero, jolloin BCP-suositus päivitetään käyttämään uusinta RFC:tä. Yksinkertaisimmillaan kielikoodi on kaksikirjaiminen merkkijono, jonka määrittelee standardi ISO 639-1 (2002). Jos kielellä ei ole kaksikirjaimista kielikoodia käytetään standardin ISO 639-2 (1998) määrittelemää kolmikirjaimista kielikoodia. (Ishida 2006)

ONKI-Paikan hakukäyttöliittymän yksi hakujen rajausehdoista on paikan nimen kieli. Käyttöliittymän kielivalintalistassa tarvitaan tämän takia kielille nimet. Muuten listassa näkyisi vain ISO:n määrittelemät kielten kielikoodit. Kielten nimet tarvitaan niillä kielillä, joilla ONKI-Paikan hakukäyttöliittymää on mahdollista käyttää. Alkuvaiheessa käyttöliittymäkielet tulevat olemaan suomi, ruotsi ja englanti. Kielten englanninkieliset nimet on saatavilla ISO 639-standardin määrittelevästä dokumentista. Suomenkieliset kielten nimet on saatavilla Tieteen tietotekniikan keskus CSC:n[32] yhdessä Kotimaisten kielten tutkimuskeskuksen kanssa laatimasta CLDR 1.5-suosituksesta[33], joka määrittelee kaikille ISO 639-2-standardissa oleville kielille suomenkieliset nimet. Vastaava lista kielten ruotsinkielisille nimille löytyy Ruotsin kansalliskirjaston LIBRIS-osaston[34] suosituksesta kielten ruotsinkielisille nimille[35].

Kielten nimet ONKI-Paikan käyttöliittymäkielillä muodostaa oman aineistonsa palveluun. Käytännössä kielten nimiä ylläpidetään tekstitiedostoissa, yksi tiedosto jokaista käyttöliittymäkieltä kohden. Tiedostoissa jokaisen rivin alussa on kielikoodi ja sen jälkeen yhtäläisyysmerkki, jonka perään tulee kielen nimi. Tämä formaatti on Java-ohjelmointikielessä käytetty ominaisuustiedostojen[36] (engl. properties file) formaatti. Ominaisuustiedostoja käytetään muun muassa Java-sovellusten käyttöliittymätekstien määrittämiseen eri kielillä.

5.2 Maailma, maanosat ja valtiot

Paikkatieto on usein monella tavoilla hierarkkista. Paikat on muun muassa maantieteellisten, poliittisten, kulttuuristen tai tilastollisten syiden takia jaettu alueisiin, jotka itsessään voivat olla jaettuja pienempiin alueisiin. Poliittisia jakoja kutsutaan usein hallinnollisiksi alueiksi, joiden ylimpänä tasona on valtiot. Valtioiden sisäiset jaot vaihtelevat suuresti valtiosta toiseen kuten luvussa 4.1 kerrottiin. Tästä syystä yleismaailmallista hallinnollisten alueiden hierarkiaa on vaikea muodostaa maantieteelliseen ontologiaan. Suomessa hallinnollinen aluejako alkaa tällä hetkellä lääneistä, jotka on jaettu maakuntiin, jotka puolestaan on jaettu seutukuntiin, ja seutukunnat kuntiin[37]. Vuonna 2010 läänit tullaan lakkauttamaan

32 CSC - Tieteellinen laskenta Oy, http://www.csc.fi/

33 http://www.csc.fi/sivut/kotoistus/suositukset/cldr-1.5-2007-08-27-kielet.pdf

34 LIBRIS - nationella biblioteksystem, http://www.kb.se/libris/

35 http://oldwww.libris.kb.se/tjanster/katalogisering/formathandbok/sprakkoder.htm

36 http://java.sun.com/docs/books/tutorial/i18n/resbundle/propfile.html

37 http://www.stat.fi/tk/tt/luokitukset/index_alue.html

(YLE 2008), jolloin Suomen sisäinen hallinnollinen aluejako tulee muuttumaan.

5.2.1 Maailma

Riippumatta aluejakotavasta, on kaikkien eri hierarkiajakojen ylätasona aina maailma tai maapallo. ONKI-Paikassa tämä taso muodostaa oman aineistonsa, jolle on päätetty antaa järjestysnumero 1. Maailma-aineistoon kuuluu ainoastaan yksi paikkainstanssi, joten muunnettava aineisto koostuu pelkästään listasta nimiä maailmalle eri kielillä. ONKI-Paikan käyttöliittymää varten tarvitaan maailmalle nimi suomeksi, ruotsiksi ja englanniksi.

Koska aineiston sisäistä paikkatunnistetta ei ole, käytetään juoksevaa numerointia luotaessa paikoille tunnisteita. Samasta syystä kuin aineistotunnisteessa, käytetään myös paikkatunnisteessa nollatäyttöä, jotta lopullinen paikan lokaalinimi olisi aakkosellisesti järjestettävissä myös numeerisesti oikeaan järjestykseen. Saadaan siis ontologiapalvelussa maailmalle lokaalinimi A0001_0000001.

Kuten paikkainstansseja pohtiessa kävi ilmi, voi eri paikoista olla olemassa eri instansseja riippuen siitä näkökulmasta, jolla paikka on määritelty. Voisi olla täysin mahdollista, että joissakin sovelluksissa halutaan eri instansseja maailmalle. Esimerkiksi asuttu maailma ja maantieteellinen maailma, joka vastaisi maapalloa. Asuttu maailmahan ei sisältäisi asumattomia alueita kuten napa-alueet. Näin ollen nämä kaksi maailman instanssia eivät voisi olla samoja. Tästä syystä maailma-aineistoon jätetään tilaa muillekin instansseilla.

5.2.2 Maanosat

Seuraava taso aluehierarkiassa on näkökulmasta riippuen joko maanosat tai valtiot. Maanosat ovat SUO-ontologiassa luokiteltu luontokohteisiin kuuluviksi muodostelmiksi. Maanosat eivät siis ole poliittinen vaan maantieteellinen aluejako, jonka takia valtiot ovat hallinnollisten alueiden hierarkiassa sijoitettu suoraan maailman alle.

Koska ontologiapalvelun hakukäyttöliittymässä haluttiin paikkahaun rajaus alueen mukaan, päätettiin valtiot ryhmitellä myös maanosien alle SUO-ontologian *suo:isPartOf*-ominaisuudella. Valtiot voidaan muutamaa poikkeusta lukuun ottamatta helposti sijoittaa johonkin maanosaan. Esimerkiksi Turkki ja Venäjä sijaitsevat molemmat sekä Euroopassa että Aasiassa. YK:n tilastotoimisto, UNSD[38] on määritellyt alueista ja valtioista listan, joka myös lajittelee valtiot maanosittain (YK 2008). Listassa valtio on määritelty kuuluvaksi vain yhteen maanosaan.

Jaottelu on tehty sillä perusteella, millä maanosalla valtion asutuksen sekä poliittisen historian painopiste sijaitsee. Venäjän tapauksessa suurin osa väestöstä asuu Euroopan puolella, missä myös pääkaupunki Moskova sijaitsee. Myös Venäjän poliittinen historia on selvästi painottunut Eurooppaan eikä Aasiaan. Turkki sen sijaan on samoista syistä selvästi enemmän aasialainen valtio.

Ontologisissa päättelyissä syntyy kuitenkin ongelmia, jos määritellään esimerkiksi Venäjän valtio kuuluvaksi pelkästään Eurooppaan. Paikkarelaatioiden mukaan pätisi silloin päättely,

38 United Nations Statistics Division, http://unstats.un.org/unsd/default.htm

jonka mukaan Mongolian yläpuolella sijaitseva venäläinen Irkutskin kaupunki olisi osa Eurooppaa, koska kaupunki on osa Venäjää, joka on osa Eurooppaa. Tämä ei tietenkään ole oikein, sillä Irkutsk sijaitsee Aasiassa. Tämän takia Venäjä ja Turkki ovat *suo:isPartOf*-suhteessa sekä Eurooppaan että Aasiaan. Tästä seuraa kuitenkin ongelma jos paikka-hierarkiaa tarkastelee ylhäältä alas. Silloin Euroopan osana on Venäjä, jonka osana on Irkutskin kaupunki. Tämä johtaisi virheelliseen päättelyyn että Irkutsk on Euroopassa. Tätä ongelmaa ovat pohtineet Holi & Hyvönen (2006), joiden mukaan *partOf*-suhde voidaan määritellä esimerkiksi todennäköisyytenä. ONKI-Paikan ensimmäisessä versiossa *partOf*-suhdetta käytetään pelkästään hakujen rajausehtojen määrittämiseen ja aluehierarkian visualisoimiseen, jolloin tästä *partOf*-suhteen tulkinnasta ei synny ongelmaa.

YK:n alue- ja valtiolistaa käytetään pohjana myös viralliselle standardille ISO 3166-1 (2006), joka määrittelee valtiot ja niille kaksikirjaimiset, kolmikirjaimiset sekä kolmi-numeroiset koodit[39]. Tämän takia päätettiin käyttää kyseistä YK:n luokitusta pohjana maanosa- ja valtioaineistoihin. Myös muun muassa Tilastokeskus (2007) sekä Julkisen hallinnon tietohallinnon neuvottelukunta suosituksessaan JHS 123 (2005) noudattavat tätä YK:n alueluokittelua ja maanosajakoa. YK:n aineistosta saadaan maanosat sekä niiden englanninkieliset nimet. Aineistoon lisättiin maanosien nimet suomeksi ja ruotsiksi käyttöliittymää varten. YK:n luokittelun mukaan maanosat ovat Aasia, Afrikka, Amerikka, Eurooppa, Oseania ja Antarktis.

Mistään lähteistä ei löytynyt minkäänlaista virallista lyhennettä tai koodia maanosille. Tämä voisi johtua siitä, että eri maanosajakoja löytyy useita. Yhteiselle standardille maanosien jaossa ei ehkä ole ollut tarvetta, koska maanosilla ei poliittisessa aluejaossa ole minkäänlaista merkitystä. Tästä syystä maanosien jako ja nimeäminen jää loppujen lopuksi sovelluksen suunnittelijan tehtäväksi. ONKI-Paikassa pohjaksi on otettu YK:n maanosa-jako, jossa maanosia on siis kuusi. Näille on annettu mahdollisimman helposti ymmärret-tävät kaksikirjaimiset tunnisteet ISO 3166-1 (2006) maakoodistandardin koodikäytännön mukaisesti. Lyhenne on johdettu maanosan englanninkielisestä nimestä. Maanosa-aineiston tunnisteeksi on ontologiapalvelussa määritelty A0002. Maanosat, niiden aineiston sisäiset tunnisteet sekä lokaalinimet on listattu taulukossa 1.

Taulukko 1. *Maanosat ja niiden lokaalinimet SUO-ontologiassa*

Maanosa	Tunniste	Lokaalinimi
Aasia	AS	A0002_AS
Afrikka	AF	A0002_AF
Amerikka	AM	A0002_AM
Antarktis	AN	A0002_AN
Eurooppa	EU	A0002_EU
Oseania	OC	A0002_OC

39 http://www.iso.org/iso/country_codes.htm

5.2.3 Valtiot

Valtioita voidaan pitää poliittisen aluejaon ylimpänä tasona. Itsenäinen valtio ei ole mitenkään yksiselitteinen termi, vaan siihen liittyy usein myös tulkintoja eri poliittisten osapuolten kesken. Maailmassa on tällä hetkellä useita alueita kuten Taiwan, Palestiina, Kosovo ym., joiden statuksesta itsenäisinä valtioina kiistellään. Itsenäisen valtion ja autonomisen alueen raja onkin usein hyvin häilyvä ja riippuu pitkälti sopimuksista alueiden ja valtioiden välillä. Mitään absoluuttisen virallista listaa itsenäisistä valtioista ei siis voi olla, sillä se olisi samalla poliittinen kannanotto siitä, mitkä alueet hyväksytään itsenäisiksi valtioiksi ja mitkä ei.

Edellisessä luvussa mainittua YK:n ja ISO:n yhdessä laatimaa ja ylläpitämää alue- ja valtiokoodilistaa ei voida käyttää pohjana itsenäisten valtioiden aineistoksi. Listassa on mukana autonomisia alueita, joilla on jonkin tason itsehallinto, mutta jotka eivät selvästi ole itsenäisiä valtioita. Esimerkiksi Ahvenanmaa on luokiteltu omaksi maakseen tässä listassa. YK korostaakin, ettei listalla oteta kantaa valtion tai alueen kansainväliseen poliittiseen tai lailliseen statuksesta (YK 2007). ISO:n valtiolista on suora kopio YK:n listasta, sillä poikkeuksella, että ISO mainitsee myös Taiwanin. Taiwan ei ole YK:n jäsen eikä sen takia ole mukana YK:n valtio- ja aluelistassa (YK 2006). Noudattamalla YK:n listaa ISO haluaa painottaa puolueettomuuttaan[40], jotta sen laatimat maakoodit voitaisiin ottaa mahdollisimman laajasti käyttöön ympäri maailmaa.

YK:n ja ISO:n valtio- ja aluelistat eivät siis ota kantaa siihen, onko alue itsenäinen valtio vai ei, joten listaa itsenäisistä valtioista pitää etsiä muualta. YK:lla on tällä hetkellä 192 jäsenvaltiota (YK 2006), jotka kaikki voidaan kiistatta luetella itsenäisiksi valtioiksi. Jäsenistä kuitenkin puuttuu Vatikaani, joka ei ole hakenut jäsenyyttä YK:iin. Vatikaani tunnustetaan silti laajasti itsenäiseksi valtioksi. Myöskään Taiwan ei ole YK:n jäsen. Maa oli jäsen vuoteen 1971 asti, jolloin Kiina otti Taiwanin paikan jäsenenä YK:ssa. Tästä päätti YK:n yleiskokous 25.10.1971 päätöslauselmassa 2758 (XXVI).

ONKI-Paikassa noudatetaan ISO:n 194 valtion listaa ja käytetään sitä pohjana valtio-aineistolle. Vaikka YK:n ja ISO:n listoja ei voida suoraan käyttää valtioaineiston lähteinä, voidaan niitä kuitenkin käyttää valtioiden URI:en muodostamisessa. ISO:n ylläpitämä kaksikirjaiminen maakoodi on ajan suhteen riittävän staattinen, jotta sitä voisi käyttää valtioaineiston sisäisenä tunnisteena ja siten URI:n osana.

ISO:n kaksikirjaiminen maakoodi on käytössä muun muassa IANA:n[41] määrittelemissä Internetin maakohtaisissa ylätason ccTLD-verkkotunnuksissa[42]. Tämän takia ISO:n kaksikirjaiminen maakoodi on monelle jo entuudestaan tuttu, jonka takia valtion URI:sta tulee näin ihmiselle helppolukuisempi. Julkisen hallinnon tietohallinnon neuvottelukunnan suosituksen JHS 123 (2005) mukaan valtiotunnuksena on käytettävä ISO:n määrittelemää kaksikirjaimista maakoodia. Koska tässä työssä noudatetaan ensisijaisesti standardeja ja

40 ISO 3166 and the UN, http://www.iso.org/iso/country_codes/background_on_iso_3166/iso_3166_and_the_un.htm

41 Internet Assigned Numbers Authority, http://www.iana.org/

42 Country-Code Top-Level Domain, http://www.iana.org/cctld/

julkisia suosituksia päädyttiin käyttämään ISO:n kaksikirjaimista maakoodia osana valtioiden URI-tunnusta.

Valtioiden URI:n lokaalinimi muodostetaan siis seuraavalla tavalla:

<aineistokoodi>_<maakoodi>[_<aikamääre>],

jossa <aineistokoodi> on valtioaineiston koodi, <maakoodi> on ISO 3166-1-standardin (2006) kaksikirjaiminen maakoodi versaaleilla ja <aikamääre> on valinnainen päivämäärä, jolloin valtion nykyiset rajat on vahvistettu. Aikamääreen muoto esitettiin luvussa 4.4.2. Valtioaineiston aineistokoodi on numeerisen järjestyksen mukaan A0003. Taulukossa 2 on esitetty esimerkkinä Suomen valtioinstanssin tunnistetiedot ontologiapalvelussa.

Taulukko 2. *Esimerkki-instanssi valtioaineistosta*

Aineisto	Valtiot
Aineistotunniste	A0003
Esimerkkipaikka	Suomi (nykyiset rajat vahvistettu 15.7.1947)
Aineiston sisäinen paikkatunniste	FI_19470715
Paikan lokaalinimi	A0003_FI_19470715
Paikan URI	http://www.yso.fi/onto/suo/A0003_FI_19470715

5.3 Maanmittauslaitoksen Paikannimirekisteri

Paikannimirekisteri, eli PNR perustuu Maa- ja metsätalousministeriön alaisuudessa toimivan Maanmittauslaitoksen Maastotietokannassa ylläpidettävään Suomen peruskartan 1:20 000 paikannimistöön[43]. Rekisteri sisältää noin 800 000 peruskartassa esitettävää luonto- ja kulttuurinimeä.

Sijaintitietoina on paikan keskipisteen koordinaatit ja joen tapauksessa sen suun koordinaatit. Paikoista kerrotaan myös paikan tyyppi sekä missä kunnassa, seutukunnassa, maakunnassa, suuralueessa ja läänissä paikka sijaitsee. Näiden tietojen lisäksi löytyy paljon tietoa siitä, miten ja millä tavalla paikan nimi pitää sijoittaa kartalle, kuten tekstiin ja kirjasimiin liittyviä tietoja ja tekstin tarkka sijainti kartalla.

Paikat on jaettu 52 eri paikkatyyppiin[44], joille jokaiselle löytyy vastaava paikkaluokka SUO-ontologiassa. Paikkatyyppi määrittelee sen, minkä SUO-ontologian paikkaluokan instanssiksi paikka tulee ontologiapalvelussa. Paikannimirekisterissä paikan tyyppi on numeerinen koodi, joka ilmoitetaan jokaisen paikan tiedoissa. Ennen Paikannimirekisterin paikkojen automaattista RDF-konvertointia on laadittava lista siitä, mitä SUO-ontologian paikkaluokkaa jokainen paikkatyyppi vastaa.

43 Nimistörekisteri, http://www.maanmittauslaitos.fi/default.asp?id=923

44 Maanmittauslaitoksen Nimistörekisterin tekninen seloste, http://www.maanmittauslaitos.fi/PopUpDocuments/nimistorekisteri_tekninen_seloste.pdf

Paikannimirekisterin nimistä noin 720 000 on suomenkielisiä, 75 000 ruotsinkielisiä, 4 500 pohjoissaamenkielisiä, 3 800 inarinsaamenkielisiä ja 150 koltansaamenkielisiä. Rekisteri kertoo nimen virallisuuden sekä sen, onko nimen kieli enemmistöasemassa paikan sijaintikunnassa. Tästä tiedosta voidaan päätellä, mikä nimi asetetaan RDF:ssä *skos:prefLabel*-luokkaominaisuudella ensisijaiseksi nimeksi paikalle, siinä tapauksessa, että paikalle löytyy useita virallisia nimiä eri kielillä.

Saamenkielisten paikannimien takia on Paikannimirekisterin aineistossa käytetty ISO-8859-10-merkistöä, joka kattaa kaikkien pohjoismaissa käytettyjen kielten merkit. Merkistön määrittelee standardi ISO/IEC 8859-10 (1998). Saamen kielissä on aakkosia kuten Đ ja Ŋ, joita ei voi esittää yleisimmin Suomessa ja länsimaissa käytetyllä ISO-8859-1-merkistöllä. Sisäisesti ONKI-Paikassa käytetään UTF-8-merkistöä, joka kattaa suurimman osan maailman kielissä käytetyistä merkeistä ja myös kaikissa saamen kielissä käytetyt merkit. Tämän takia koko Paikannkimirekisterin aineisto on ennen RDF-muunnosta muunnettava UTF-8-merkistöön.

Paikannkimirekisterissä paikkojen koordinaatit ilmoitetaan kolmessa eri koordinaatistossa: KKJ/PKJ, KKJ/YKJ ja EUREF-FIN/UTM. Sisäisesti ONKI-Paikassa kaikkien paikkojen koordinaatit ilmoitetaan WGS84-koordinaatistossa, joka vastaa melko tarkasti EUREF-FIN maantieteellistä koordinaatistoa. EUREF-FIN/UTM ja EUREF-FIN maantieteellisen koordinaatiston välillä on siis tehtävä muunnos, joko aineiston RDF-muunnoksen tai RDF-aineiston indeksoinnin yhteydessä. Jotta aineistojen indeksointi sujuisi mahdollisimman kevyesti päädyttiin ontologiapalvelussa ratkaisuun, jossa RDF-varaston kaikkien paikkojen pitää sisältää koordinaatit WGS84-koordinaatistossa. Tällöin vastuu eri aineistojen omien koordinaattijärjestelmien muuntamisesta WGS84-koordinaatistoon jää aineiston RDF-muunnoksen tekijälle ja ONKI-Paikka voi toimia sisäisesti pelkästään WGS84-koordinaatistolla. Paikkojen koordinaattien käsittelystä kerrottiin tarkemmin luvussa 4.4.3.

Paikoilla on Paikannimirekisterissä myös sisäinen tunniste, jolla eri paikannimet yhdistetään johonkin tiettyyn paikkaan. Paikkatunniste on aineistossa muuttumaton, ellei paikalle tapahdu merkittäviä muutoksia kuten aluerajojen ja nimen muutos. Esimerkiksi kuntaliitoksissa ja kuntien rajojen muuttuessa syntyy uusia paikkoja ja niille uusia paikkatunnisteita Paikannimirekisteriin. Paikkatunniste on kahdeksan merkkiä pitkä numeerinen merkkijono, jonka takia se soveltuu sellaisenaan, ilman tarvetta muunnoksille, osaksi paikan lokaalinimeä. Esimerkiksi taulukossa 3 on Paikannimirekisterin aineistosta otettu, Helsingissä sijaitseva Kampin asuinalue. Paikannimirekisterin aineistokoodiksi on valittu A0008. Yhteenveto kaikista ONKI-Paikan aineistoista sekä niiden aineistokoodeista löytyy myöhemmin luvussa 5.8.

Paikannimirekisterissä suurimmat paikat aluehierarkiassa ovat kunnat. Rekisteri ei varsinaisesti sisällä tietoa esimerkiksi lääneistä, maakunnista ja seutukunnista. Nämä alueet voidaan kuitenkin päätellä rekisterin tiedoista, sillä jokaisen paikan kohdalla ilmoitetaan missä kunnassa, seutukunnassa, maakunnassa ja läänissä paikka sijaitsee. Näistä hallinnollisista alueista ilmoitetaan myös aluekoodi sekä alueen nimi suomeksi että ruotsiksi. Näiden tietojen perusteella voidaan siis luoda koko Suomen aluehallinnon hierarkia ja jokaisesta alueesta voidaan luoda instanssi SUO-ontologiaan.

Ongelma Paikannimirekisterin kanssa on kuitenkin se, ettei se ole virallinen lähde hallinnollisille alueille, vaan käyttää niitä paikkojen metatietona rekistereissään. Näissä tiedoissa saattaa olla virheitä, kuten olikin Paikannimirekisterin versiossa vuodelta 2006, jossa oli muun muassa muutamia virheitä maakuntien nimissä. Virheet havaittiin, kun Paikannimirekisterin aineisto muunnettiin tätä työtä varten RDF-muotoon. Suomessa virallista tietoa hallinnollisista alueluokituksista ylläpitää Tilastokeskus, jonka tietoja käytetään lähteenä myös ONKI-Paikassa. Paikannimirekisteristä tuodaan siis kaikki kuntaa pienemmät paikat ja liitetään ne kuntaan rekisterissä ilmoitetun kuntakoodin avulla.

Taulukko 3. *Esimerkki-instanssi Maanmittauslaitoksen Paikannimirekisteristä*

Aineisto	Maanmittauslaitoksen Paikannimirekisteri
Aineistotunniste	A0008
Paikka	Kamppi, Helsinki
Aineiston sisäinen paikkatunniste	10342635
Paikan lokaalinimi	A0008_10342635
Paikan URI	http://www.yso.fi/onto/suo/A0008_10342635

5.4 Suomen hallinnolliset alueet

Jotta koko hallinnollisten alueiden hierarkia olisi Suomen osalta täydellinen, tarvitaan virallisesta lähteestä tiedot lääneistä, maakunnista, seutukunnista ja kunnista. Tilastokeskus ylläpitää tätä virallista alueluokitusta yhdessä muiden tahojen kanssa, jotka myös antavat eri alueille numeeriset tunnukset. Näiden tunnusten avulla luodaan soveltuvin osin hallinnollisille alueille URI-tunnisteet ONKI-Paikassa.

Kaikista hallinnollisista alueista tarvitaan nimet ainakin suomeksi ja ruotsiksi sekä englanniksi, jos sellainen erikseen on olemassa. Alueista tarvitaan myös koordinaatit ja tieto siitä, mihin paikkaan alue on *suo:isPartOf*-suhteessa. Jokaiselle alueelle pitää myös määritellä aineiston sisäinen tunniste, jota käytetään paikan URI:n luomisessa.

5.4.1 Läänit

Suomen läänijako uudistui 1.9.1997, jolloin läänien määrä väheni kahdestatoista läänistä kuuteen. Läänijaon määrittelee lääninhallituslaki (22/1997)[45], jonka mukaan läänit ovat Etelä-Suomen lääni, Länsi-Suomen lääni, Itä-Suomen lääni, Oulun lääni, Lapin lääni ja Ahvenanmaan lääni. Tilastokeskus on antanut näille lääneille numerot yhdestä kuuteen edellä mainitussa järjestyksessä. Nämä numerot toimivat myös läänien hallinnollisten instanssien tunnisteina. Tunnistetta käytetään myös ONKI-Paikassa, jossa tunnisteen perään lisätään päivämäärä, jolloin instanssin nykyinen maantieteellinen alue on astunut voimaan. ONKI-Paikassa Suomen läänien aineistotunnisteeksi on valittu A0004. Esimerkissä alhaalla taulukossa 4 on Lapin läänin instanssin tunnistetiedot.

45 Lääninhallituslaki (10.1.1997/22), http://www.finlex.fi/fi/laki/ajantasa/1997/19970022

Taulukko 4. *Esimerkki-instanssi Suomen lääniaineistosta*

Aineisto	Suomen läänit
Aineistotunniste	A0004
Paikka	Lapin lääni
Aineiston sisäinen paikkatunniste	5_19970901
Paikan lokaalinimi	A0004_5_19970901
Paikan URI	http://www.yso.fi/onto/suo/A0004_5_19970901

5.4.2 Maakunnat

Suomessa läänit jaetaan maakuntiin, joiden nimet ja alueet on määritelty Valtioneuvoston päätöksessä maakunnista 26.2.1998/147[46]. Päätös astui voimaan 1.3.1998, ja on edelleen voimassa Suomen maakuntajaon perustana. Kuntaliitosten seurauksena maakuntien aluejakoihin on sittemmin tullut joitakin muutoksia, joiden takia maakunnista on olemassa ajallisesti myös useita eri maantieteellisiä instansseja. Tässä työssä otetaan mukaan kaikki vuonna 2008 voimassa olevat maakuntien instanssit. Esimerkiksi Pirkanmaan maakunnan alue muuttui 1.1.2007, kun siellä sijainnut Längelmäen kunta lakkautettiin ja jaettiin Jämsän ja Oriveden kesken[47]. Jämsä sijaitsee Keski-Suomen maakunnassa, jonka takia Längelmäen aluejako vaikutti Pirkanmaan maakunnan maantieteellisen alueen kattavuuteen. Taulukossa 5 on esitetty Pirkanmaan maakunnan instanssin tunnistetiedot.

Taulukko 5. *Esimerkki-instanssi Suomen maakunta-aineistosta*

Aineisto	Suomen maakunnat
Aineistotunniste	A0005
Paikka	Pirkanmaan maakunta
Aineiston sisäinen paikkatunniste	06_20070101
Paikan lokaalinimi	A0005_06_20070101
Paikan URI	http://www.yso.fi/onto/suo/A0005_06_20070101

Tilastokeskus on antanut maakunnille kaksinumeroiset tunnukset, joita käytetään ONKI-Paikassa maakunnan maantieteellisen instanssin paikkatunnisteen luomiseen. Kuten muissa hallinnollisten alueiden tunnisteissa, myös maakuntien tunnisteissa lisätään oletusarvoisesti perään päivämäärä, jolloin maakunnan maantieteellisen alueen kattavuus on astunut voimaan.

Yleensä maakunnista käytetään niiden kansallista nimeä myös ulkomaisissa kielissä. Joillakin maakunnilla on kuitenkin esimerkiksi englannin kielessä vakiintuneet nimet, jotka

46 http://www.finlex.fi/fi/laki/ajantasa/1998/19980147

47 Längelmäen kunnan lakkauttamisesta ja sen alueiden liittämisestä Jämsän kaupunkiin ja Oriveden kaupunkiin (523/2006), http://www.finlex.fi/fi/laki/alkup/2006/20060523

poikkeavat kansallisista nimistä. Sisäasianministeriössä on laadittu suositus maakuntien nimien vastineiksi englanniksi, saksaksi ja ranskaksi (Sisäasiainministeriö 2001). Esimerkiksi Karjala ja Pohjanmaa ovat maakuntia, joilla on englanniksi vakiintuneet nimet Karelia ja Ostrobothnia.

5.4.3 Seutukunnat

Seutukuntajako otettiin käyttöön vuonna 1994 aluekehittämislakien perusjaoksi[48]. Sisäasiainministeriö vahvistaa seutukuntajaon, mutta seutukunnat päättävät itse nimistään. Yhdessä Tilastokeskuksen kanssa Sisäasiainministeriö sopii seutukuntien tunnuksista, jotka ovat kolmenumeroisia lukuja. Myös seutukuntien tunnukset ovat hallinnollisen alueen eikä maantieteellisen alueen instanssin tunniste. Tämän takia myös seutukuntien paikkatunnisteen perään lisätään päivämäärä, jolloin seutukunnan maantieteellisen alueen kattavuus on astunut voimaan.

Myös seutukuntien tunnukset viittaavat hallinnolliseen alueeseen, joka on eri kuin maantieteellinen alue. Seutukunnissa tapahtuvien muutosten seurauksena seutukuntakoodit eivät välttämättä pysy samana, mutta myös seutukuntien paikkatunnisteiden perään lisätään päivämäärä, jolloin seutukunnan maantieteellinen kattavuus on astunut voimaan. Taulukon 6 esimerkissä on Härmänmaan seutukunta, jonka alueellinen kattavuus muuttui 1.1.2007, kun seutukunnassa mukana ollut Lapuan kunta siirtyi Seinäjoen seutukuntaan. Härmänmaan seutukuntakoodi 145 pysyi samana myös muutoksen jälkeen. Taulukossa 6 on esitetty Härmänmaan seutukunnan instanssin tunnistetiedot.

Taulukko 6. *Esimerkki-instanssi Suomen seutukunta-aineistosta*

Aineisto	Suomen seutukunnat
Aineistotunniste	A0006
Paikka	Härmänmaan seutukunta
Aineiston sisäinen paikkatunniste	145_20070101
Paikan lokaalinimi	A0006_145_20070101
Paikan URI	http://www.yso.fi/onto/suo/A0006_145_20070101

5.4.4 Kunnat

Kuntien numerotunnukset antaa Väestörekisterikeskus julkisen hallinnon suosituksen, JHS 110 (n.d.) mukaisesti. Kuntanumeroiden hyvä puoli on siinä, että hallinnollisesti uusien kuntien syntyessä, annetaan uusille kunnille myös uudet numerotunnukset. Numerotunnus ei kuitenkaan muutu silloin, kun jokin kunta yhdistyy toiseen kuntaan. Silloin kunta, johon liittyminen tapahtui, säilyttää kuntanumeronsa. Kunnat voidaan siis määritellä poliittisina tai hallinnollisina toimijoina, joille kuntanumero on annettu. Tämä hallinnollinen toimija ei siis ole sama kuin kunnan maantieteellinen alue, sillä toimijan hallinnoima maantieteellinen alue voi ajan myötä muuttua. Tämän takia kuntanumeroa ei

48 Laki alueiden kehittämisestä (1135/1993), http://www.finlex.fi/fi/laki/alkup/1993/19931135

voida yksinään käyttää kunnan URI:n luomiseen. Kuntanumero yhdistettynä kunnan rajamuutoksen päivämäärällä sen sijaan yksilöi kunnan ja sen kattaman maantieteellisen alueen tietyksi paikkainstanssiksi.

Kunta-aineistoon pitää siis myös lisätä päivämäärä, jolloin kunnan nykyiset aluerajat ovat astuneet voimaan. Taulukon 7 esimerkissä on vuonna 2007 perustettu Akaan kaupunki, joka syntyi kun Toijalan kaupunki kuntakoodilla 864 ja Viialan kunta kuntakoodilla 928 lakkautettiin ja perustettiin uusi Akaan kaupunki kuntakoodilla 020. Kumpaakaan kuntaa ei siis liitetty toiseen, vaan molemmat kunnat lopettivat toimintansa hallinnollisina toimijoina ja perustettiin uusi.

Taulukko 7. *Esimerkki-instanssi Suomen kunta-aineistosta*

Aineisto	Suomen kunnat
Aineistotunniste	A0007
Paikka	Akaa
Aineiston sisäinen paikkatunniste	020_20070101
Paikan lokaalinimi	A0007_020_20070101
Paikan URI	http://www.yso.fi/onto/suo/A0007_020_20070101

5.5 GEOnet Names Server

GEOnet Names Server eli GNS on kahden yhdysvaltalaisten viraston yhdessä ylläpitämä lista ulkomaisten maantieteellisten kohteiden nimistä. Virastot ovat National Geospatial-Intelligence Agency[49], eli NGA ja U.S. Board on Geographic Names[50], eli BGN. NGA on Yhdysvaltain puolustusministeriön alainen virasto, jonka tehtävänä on kerätä paikkatietoa ympäri maailmaa pääosin Yhdysvaltain hallituksen käyttöön. BGN:n tehtävä puolestaan on standardisoida ulkomaisten paikannimien kirjoitusasu Yhdysvaltain hallituksen julkaisuja varten. GNS ei sisällä Yhdysvaltojen rajojen sisällä olevien paikkojen nimiä. Tälle on oma rekisterinsä, Geographic Names Information System, eli GNIS joka esitetään luvussa 5.6.

GNS on vapaasti saatavilla NGA:n internetsivuilta yhtenä isona CSV-muodossa[51] olevana tekstitiedostona. Aineistoa saa vapaasti käyttää, eikä siihen liity mitään lisenssejä tai käyttörajoituksia. NGA kuitenkin suosittelee, että aineistoa käyttävä sovellus ilmoittaisi lähteensä seuraavalla tekstillä:

> *”Toponymic information is based on the Geographic Names Data Base, containing official standard names approved by the United States Board on Geographic Names and maintained by the National Geospatial-Intelligence Agency. More information is available at the Products and Services link at www.nga.mil. The National Geospatial-Intelligence Agency name, initials,*

49 National Geospatial-Intelligence Agency, http://www.nga.mil/

50 U.S. Board on Geographic Names, http://geonames.usgs.gov/

51 Comma Separated Values on pilkulla tai muilla erotinmerkeillä muotoiltu taulukkomuotoinen teksti

and seal are protected by 10 United States Code Section §445."

GNS sisältää noin 4 miljoonaa maantieteellistä kohdetta ja noin 5,5 miljoonaa nimeä näille kohteille. Paikat on luokiteltu eri paikkatyyppeihin kuten Paikannimirekisterissä. Eri paikkatyyppejä on kuitenkin huomattavasti enemmän kuin Paikannimirekisterissä, lähinnä sen takia että paikkatyyppien alaluokitus on huomattavasti hienojakoisempi. Kuten Paikannimirekisterissä tarvitaan RDF-muunnosta varten myös GNS:ssä lista kaikista paikkatyypeistä ja tieto siitä, mitä SUO-ontologian luokkaa ne vastaavat.

Paikkojen sijaintitiedoissa GNS käyttää maantieteellistä WGS84-koordinaatistoa, joten paikkojen koordinaatit voidaan ottaa ontologiapalvelussa käyttöön sellaisenaan. Aineisto käyttää myös UTF-8-merkistöä, joka helpottaa paikkatietojen muuntamista RDF:ksi. GNS käyttää maakoodeina kaksikirjaimisia aluekoodeja, jotka määrittelee standardi FIPS10-4[52] (1995). Nämä eivät täysin vastaa ISO 3166-1-maakoodeja. Jotta GNS:n paikat saataisiin sijoiteltua oikean valtion alle SUO-ontologian *suo:isPartOf*-hierarkiassa, tarvitaan lista FIPS10-4-koodeista ja niitä vastaavista kaksikirjaimisista ISO 3166-1-maakoodeista.

Paikkojen nimien kieli on merkitty aineistoon kolmikirjaimisella kielikoodilla[53], jonka määrittelee standardi ISO 639-3 (2007). Kielikoodi on suunniteltu kattamaan kaikki maailman tunnetut, sekä elävät että kuolleet ihmiskielet. ISO 639-3-koodit ovat laajennos kolmikirjaimisista ISO 639-2-kielikoodeista, jonka takia niitä voidaan käyttää silloin, kun jollekin kielelle ei löydy kielikoodia ISO 639-2-standardista.

Yksi ongelma GNS-aineiston paikannimien kielikoodimerkinnöissä on se, että isossa osassa nimistä puuttuu kielimerkintä. Tämä tuottaa selvän ongelman ontologiapalvelussa, jossa paikkojen nimien kielellä pitää voida tehdä hakurajauksia. Miljoonille nimille ei voida RDF-muunnosvaiheessa keksiä kielikoodeja, joten paikannimen kielimerkinnälle on keksittävä toinen ratkaisu. Sekä ISO 639-2 että ISO 639-3 määrittelevät kielikoodin *und* (engl. undetermined), määrittämättömälle kielelle. RDF/XML:ssä voitaisiin siis nimille määritellä attribuutti *xml:lang="und"*.

Attribuutin arvon tällä hetkellä määrittävä dokumentti RFC4646 (Phillips & Davis 2006) suosittelee kuitenkin, ettei *und*-kielikoodia käytettäisi, ellei sille ole painavia syitä sovelluksen toiminnan kannalta. Tällaisia painavia syitä ei ONKI-Paikassa ole, joten paikkojen nimimääritysten yhteyteen ei lisätä *xml:lang*-attribuuttia, jos paikan nimen kieli ei ole tiedossa. Jää siis indeksoinnin ja paikkojen hakutoiminnon tehtäväksi esittää käyttäjälle kielirajaushaut niin, että käyttäjä ymmärtää, ettei hakurajauksella voida taata kaikkien tietynkielisten nimien löytyminen.

Aineisto luokittelee paikat karkeasti yhdeksään pääluokkaan ja tarkemmin noin kuuteen ja puoleen sataan paikkatyyppiin. Paikkatyyppiluokittelu on siis varsin yksityiskohtainen. Esimerkiksi pyramidilla ja öljylähteellä on aineistossa omat paikkatyyppinsä. Jotta GNS:n aineistosta peräisin olevien paikkojen sijoittaminen SUO-ontologian instansseiksi olisi mahdollista, on ensin määriteltävä millä tavalla GNS:n paikkatyypit vastaavat SUO-ontologian paikkaluokkia. Koska GNS:n paikkatyypitys on huomattavasti hienojakoisempi

52 Federal Information Processing Standards Publication 10-4, http://www.itl.nist.gov/fipspubs/fip10-4.htm

53 ISO 639-3, http://www.sil.org/iso639-3/

kuin SUO-ontologian alkuperäinen tyypitys, on SUO-ontologian paikkaluokkahierarkiaa jouduttu laajentamaan GNS:n paikkatyypeillä. Näin aineiston RDF-muunnoksen yhteydessä kadotetaan mahdollisimman vähän informaatiota. Toinen vaihtoehto olisi ollut määritellä lista siitä, mitä SUO-ontologian paikkaluokkaa GNS:n paikkatyyppi parhaiten vastaa. Tässä vaihtoehdossa kuitenkin suuri osa paikkatyypeissä olevasta informaatiosta olisi kadonnut. Esimerkiksi GNS:n pyramidia määrittävä tyyppi olisi pitänyt korvata SUO-ontologian rakennusta tai muinaisjäännöstä määrittelevällä tyypillä.

GNS-aineiston sisäinen paikkatunniste UFI, eli Unique Feature Identifier on numero, joka suurella osalla paikoista on negatiivinen. Aineiston lähdesivulla eikä aineiston kenttien kuvauksissa tarjota mitään selitystä sille, miksi joidenkin paikkojen tunniste on negatiivinen. Asiasta lähetettiin kysely sähköpostitse, johon vastattiin, että paikkainstanssit, jotka ovat olleet olemassa vanhassa aineistossa ennen GNS-aineiston perustamista vuonna 1994, ovat saaneet vanhan aineiston sisäiseen tunnisteeseen etuliitteeksi miinusmerkin. Miinusmerkki ei tuota mitään ongelmia paikkainstanssin lokaalinimessä, joten sille ei ole tarvetta tehdä minkäänlaista muunnosta ja GNS-aineiston sisäistä paikkatunnistetta voidaan käyttää sellaisenaan osana paikkainstanssin lokaalinimeä. Taulukossa 8 on esimerkki GNS-aineistosta peräisin olevan paikkainstanssin tunnistetiedoista.

Taulukko 8. *Esimerkki-instanssi GNS-aineistosta*

Aineisto	GEOnet Names Server
Aineistotunniste	A0009
Paikka	Madrid, Espanja
Aineiston sisäinen paikkatunniste	-390625
Paikan lokaalinimi	A0009_-390625
Paikan URI	http://www.yso.fi/onto/suo/A0009_-390625

5.6 Geographic Names Information System

Edellä esitetty GEOnet Names Server ei sisältänyt lainkaan Yhdysvaltojen alueella olevia paikkoja. Näille paikoille on oma aineistonsa Geographic Names Information System, eli GNIS, jota ylläpitää yhdysvaltalainen virasto U.S. Board on Geographic Names[54] (BGN). Viraston tehtävänä on laatia liittovaltiotasolla säännöt maantieteellisten nimien käytölle ja kirjoitusasulle. Aineisto on pitkälti samantyyppinen kuin GNS muutamia poikkeuksia lukuun ottamatta.

GNIS käyttää omaa paikkatyyppiluokitusta, jonka takia tälle aineistolle on tehtävä oma lista siitä, mitä SUO-ontologian paikkaluokkaa GNIS-paikkatyyppi vastaa. Paikkatyyppejä on 65, joten tyypitys on huomattavasti karkeampi kuin GNS:ssä. Aineistossa määritellään paikalle virallisesti käytetty nimi, mutta nimen kieltä ei määritellä. Useilla alueilla, kuten Havaijilla ja Alaskassa, on useita alkuperäiskansojen kielistä peräisin olevia paikkojen nimiä. Toisaalta GNIS-aineoston yksi tehtävä on yhtenäistää paikkojen nimien kirjoitusasu

54 U.S. Board on Geographic Names, http://geonames.usgs.gov/

englannin kielellä, joten aineiston RDF-muunnoksessa voidaan tehdä oletus, että kaikkien paikkojen nimet ovat englanninkielisiä. Alkuperäiskielillä nimien kirjoitusasua on usein eri. Esimerkiksi Havaiji on englanniksi Hawaii, mutta havaijinkielellä Hawai´i.

Paikkojen koordinaatit ilmoitetaan NAD83-koordinaattijärjestelmässä, jota käytännössä, ainakin ONKI-Paikan tarpeisiin, voidaan pitää samana kuin WGS84. Näiden kahden koordinaattijärjestelmän välinen ero on maksimissaan 1,5 metriä[55], joka ei vaikuta ONKI-Paikan hakutulosten visualisointiin millään tavalla. Sisäisestihän ONKI-Paikassa koordinaatteja käytetään pelkästään paikkojen aluerajaushauissa sekä paikan esittämiseen kartalla. Näihin käyttötarpeisiin puolentoista metrin virhemarginaali ei vaikuta millään tavalla. GNIS-aineistossa paikoille ilmoitetaan myös korkeus metreinä, joka lisätään tässä aineistossa paikkojen koordinaattien määrittelyihin sekä koordinaattipisteiden URI-tunnisteisiin luvussa 4.4.3 esitetyn säännön mukaisesti.

Jokaiselle maantieteelliselle kohteelle on GNIS-aineistossa annettu yksilöllinen tunniste. Tunniste on korkeintaan kymmenen merkkiä pitkä, numeerinen merkkijono. Tätä tunnistetta käyttäen voidaan muodostaa GNIS-aineistosta peräisin olevien paikkojen URI-tunnisteet. Taulukossa 9 on esimerkki GNIS-aineistosta tuodun paikkainstanssin tunnistetiedoista.

Taulukko 9. *Esimerkki-instanssi GNIS-aineistosta*

Aineisto	Geographic Names Information System
Aineistotunniste	A0010
Paikka	Miami, Florida
Aineiston sisäinen paikkatunniste	295004
Paikan lokaalinimi	A0010_295004
Paikan URI	http://www.yso.fi/onto/suo/A0010_295004

Aineistosta käy ilmi paikkojen sijainti hallinnollisessa aluehierarkiassa. Yhdysvalloissa osavaltiot on jaettu piirikuntiin (engl. county). Jokaisen paikan kohdalla määritellään osavaltio ja piirikunta, jossa paikka sijaitsee. Myös osavaltiot ja piirikunnat on lueteltu paikkoina aineistossa, joten myös näille hallinnollisille alueille on helppo luoda URI:t. Ongelman tuottaa kuitenkin se, ettei paikkojen tiedoissa ilmoitettu osavaltio ja piirikunta ole määritelty GNIS-tunnisteella vaan erillisellä hallintoaluekoodilla. Aineistossa osavaltio, jossa paikka sijaitsee, ilmoitetaan kaksikirjaimisella FIPS 5-2[56] (1987) koodilla. Piirikunta ilmoitetaan kolminumeroisella FIPS 6-4[57] (1990) koodilla, joka on jokaiselle piirikunnalle yksilöllinen osavaltion sisällä.

Jotta paikat voitaisiin liittää piirikuntiin ja piirikunnat osavaltioihin *suo:isPartOf*-suhteella,

55 http://www.geod.nrcan.gc.ca/faq_e.php#23

56 Federal Information Processing Standards Publication 5-2, http://www.itl.nist.gov/fipspubs/fip5-2.htm

57 Federal Information Processing Standards Publication 6-4, http://www.itl.nist.gov/fipspubs/fip6-4.htm

pitäisi olla olemassa lista, joka määrittelee mitä FIPS-koodia hallinnollisten alueiden GNIS-paikkatunnisteet vastaavat. GNIS-aineistosta tätä tietoa ei voida päätellä, sillä kaikki hallinnolliset alueet luokitellaan samalla *Civil*-paikkatyypillä. Tämän takia ei voida tietää, onko kyseessä osavaltio vai piirikunta, eikä osavaltioiden ja piirikuntien GNIS-paikkatunnistetta voida siksi päätellä. Ratkaisun ongelmaan tuo BGN:n julkaisema Yhdysvaltojen hallinnollisten alueiden lista, joka luettelee erikseen osavaltiot ja piirikunnat GNIS-paikkatunnisteineen. Tämän listan avulla pystytään aineiston RDF-muunnoksen yhteydessä yhdistämään FIPS-koodit aineiston sisäisiin paikkatunnisteisiin.

Kuten GNS on myös GNIS-aineisto vapaasti käytettävissä. BGN kuitenkin suosittelee, että aineistoa käyttävissä sovelluksissa mainittaisiin aineiston alkuperäksi U.S. Geological Survey-virasto[58], jonka alaisuudessa BGN toimii.

5.7 Suomen ajallinen paikkaontologia

Suomen ajallinen paikkaontologia eli SAPO[59] (Kauppinen et al. 2008; Väätäinen 2008) on FinnONTO-projektin Semanttisen laskennan tutkimusryhmän kehittämä ontologia paikkatiedon ajallisten muutosten esittämiseen. Paikkaontologioiden ongelmana on se, että ne ovat yleensä vain hetkellinen kuvaus alati muuttuvassa paikkatiedon maailmassa. Etenkin hallinnollisten alueiden jaoissa, nimissä ja rajoissa tapahtuu jatkuvasti muutoksia, jotka pitää voida esittää paikkaontologiassa. Esimerkiksi kuntaliitokset tarkoittavat sitä, että jotkut kunnat lakkaavat olemasta samalla kun uusia kuntia syntyy.

SAPO on tällä hetkellä populoitu Suomen historiallisilla kunnilla viimeisten noin 150 vuoden ajalta. Ajallinen ulottuvuus paikkaontologiassa tulee hyödylliseksi etenkin aluehierarkioihin perustuvien paikkahakujen yhteydessä. Esimerkiksi nykyinen Lappeenranta peitti 1900-luvun vaihteen Viipurista yli kymmenen prosenttia. Sen ajan Viipurilla indeksoidut tiedot, kuten eri kylien kartat tai valokuvat ovat siis todennäköisesti osittain peräisin nykyisen Suomen alueelta. Tämä on tietoa, jota voidaan hyödyntää aineistojen hallinnassa ja ontologisiin suhteisiin perustuvien hakujen tehostamisessa (Sinkkilä 2008).

SAPO toimii yhteistyössä SUO-ontologian kanssa niin, että SAPO:ssa määritellyt paikat ovat SUO-ontologian paikkainstansseja. Koska SAPO:n aineisto on jo valmiiksi RDF-muodossa yhtenä XML-tiedostona, on sen tuominen ONKI-Paikan käyttöön suhteellisen yksinkertainen prosessi. Jokaisen paikan RDF-kuvaus on vain rekisteröitävä erikseen RDF-varstoon, jotta ONKI-Paikan indeksointijärjestelmä pystyy käsittelemään ne hakuja varten.

Yksi haaste SAPO:n aineiston kanssa on sen sisältämät päällekkäisyydet Suomen kunta-aineiston kanssa. Kaikki Suomen kunta-aineistossa olevat kunnat löytyvät myös SAPO:sta. Jotta päällekkäisyyksistä olisi mahdollista vältyttyä hakukäyttöliittymässä, on kehitettävä mekanismi, jolla samojen paikkojen välille voitaisiin luoda *owl:sameAs*-suhde. SUO-ontologia määrittelee kunta- ja kaupunkiluokille ominaisuuden *suo:kuntakoodi*, joka on Väestörekisterikeskuksen ylläpitämä numeerinen tunniste kunnille. Lisäämällä tämä sekä SAPO:n että Suomen kunta-aineiston kuntiin, voidaan aineiston tuomisen yhteydessä

58 U.S. Geological Survey , http://www.usgs.gov/

59 Suomen ajallinen paikkaontologia, http://www.seco.tkk.fi/ontologies/sapo/

määritellä ONKI-Paikan sisäisesti *owl:sameAs*-suhde kuntainstanssien välille.

5.8 Yhteenveto

Tämän työn tuloksena ONKI-Paikkaan lisätään 11 aineistoa paikkoineen taulukon 10 mukaisesti. Lopputuloksena on koko maailman kattava, useilla miljoonilla paikka-instansseilla populoitu ontologinen paikkatietopalvelu. Useimmilla aineistoilla muunnos RDF-muotoon on suhteellisen vaivatonta. Suurin työ liittyy paikkojen URI-tunnisteiden määrittämiseen ja paikkojen tyypityksen sovittamiseen SUO-ontologian luokkiin.

Aineistot A0001-A0007 on suurimmaksi osaksi koottu tätä työtä varten. Suoraan ONKI-Paikan tarpeisiin sopivia, vastaavia aineistoja ei ollut valmiina olemassa. Työ koostui enimmäkseen tietojen keräämisestä yhteiseen taulukkomuotoiseen aineistoon, jota on helppo ja yksinekertaista ylläpitää ja päivittää esimerkiksi taulukkolaskentaohjelmassa. Näissä aineistoissa suurin työ oli kerätä paikkojen viralliset nimet suomen, ruotsin ja englannin kielillä, sekä selvittää se, mitä aineiston sisäistä tunnistetta voisi käyttää paikkojen yksilöimiseen. Myös paikkojen koordinaattien selvittäminen oli varsin hankala työ johtuen eri aineistoissa käytetyistä erilaisista koordinaattijärjestelmistä sekä niiden välillä tehtävistä koordinaattimuunnoksista.

Aineistot A0008-A0010 olivat suhteellisen vaivattomia muuntaa RDF-muotoon. Suurin ongelma oli selvittää muiden valtioiden hallinnollisten alueiden hierarkiajako, joka vaihtelee suuresti valtiosta toiseen. GNS- ja GNIS-aineistoista aluehierarkiat olivat pääteltävissä.

Taulukko 10. *Yhteenveto paikkatiedon lähdeaineistoista*

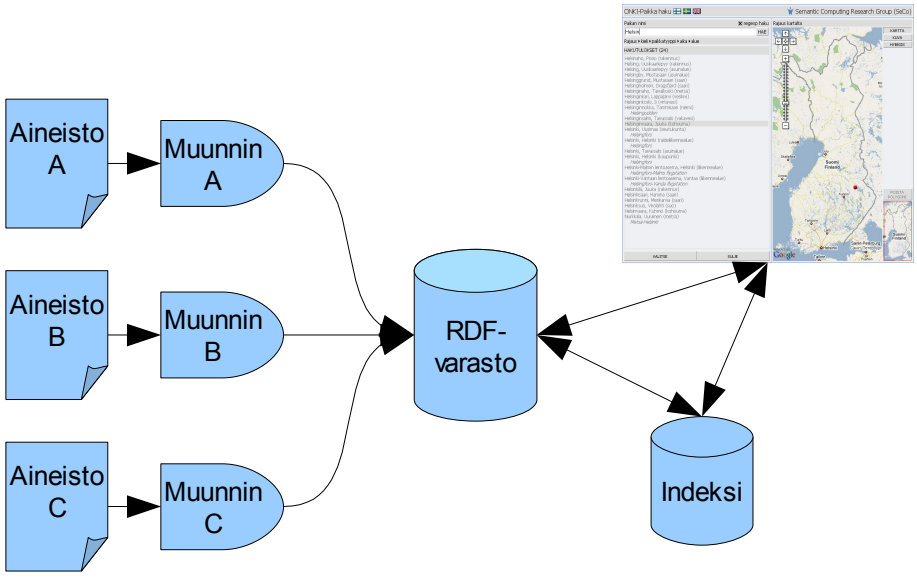
Aineistokoodi	Aineisto	Paikkainstanssien määrä vuoden 2008 alussa
A0001	maailma	1
A0002	maanosat	7
A0003	valtiot	194
A0004	Suomen läänit	6
A0005	Suomen maakunnat	20
A0006	Suomen seutukunnat	77
A0007	Suomen kunnat	415
A0008	Paikannimirekisteri	n. 800 000
A0009	GEOnet Names Server	n. 4 100 000
A0010	Geographic Names Information System	n. 2 000 000
A0011	Suomen ajallinen paikkaontologia	n. 650

6 ONKI-PAIKKA ONTOLOGIAPALVELU

6.1 Kokonaiskuva ja prosessit

Paikkatiedon ontologiapalvelun tehtävä on nimensä mukaan tarjota paikkatietoa ontologisessa muodossa. Useimmiten paikkatietoa ylläpidetään relaatiotietokannoissa tai XML-rakenteissa kuten luvussa 2.2 esitettiin. Jotta paikkatieto olisi hyödynnettävissä semanttisessa webissä, on se ensin muunnettava ontologiseen muotoon. Käytännössä tämä tarkoittaa sitä, että paikkatiedon resursseille on annettava URI:t, jonka jälkeen resursseihin voi liittää ominaisuuksia. Paikkatiedon resurssit ja niiden ominaisuudet määrittelee paikkatiedon ontologia, jota populoidaan luomalla instansseja paikkaontologian lähteistä.

Ensimmäinen askel ontologiapalvelun luomisessa on sopivan paikkatieto-ontologian valitseminen, jota käyttää pohjana paikkatiedon kuvaamiseen semanttisessa webissä. ONKI-Paikan lähtökohtana oli käyttää jo valmiiksi kehitettyä suomalaista paikkaontologiaa SUO:ta. Ontologian populoiminen paikkatiedolla on suhteellisen suoraviivainen prosessi. Ontologia itsessään määrittelee paikkatiedon sanaston semanttisessa webissä ja sen, millä tavalla paikkatieto on jaettu luokkiin ja niiden ominaisuuksiin. Haasteellisinta ontologian populoimisessa on eri paikkatietoresurssien URI:en luominen. Muut tiedot paikoista voidaan poimia suoraan aineistosta melkein sellaisenaan. Ainoana poikkeuksena on paikkojen koordinaattien vaihtelevat esitysmuodot.



Kuva 1. *Kokonaiskuva ONKI-Paikka-palvelusta, jossa paikkatietoaineistot A-C ensin muunnetaan omilla muuntimillaan RDF-muotoiseksi dataksi. Paikkojen RDF-kuvaukset tallennetaan RDF-varastoon, joka indeksoidaan nopeita hakuja varten.*

Kokonaisuudessaan aineiston tuominen ONKI-Paikan käyttöön tapahtuu neljässä eri vaiheessa, joissa jokaisessa lopputuloksena on jokin määritelty esitysmuoto paikkatiedosta (kuva 1):

1. Aineisto tallennetaan alkuperäismuodossaan ONKI-Paikan omaan aineistopankkiin. Pankki sisältää yhden kansion jokaista aineistoa kohden. Seuraavassa vaiheessa luotava muunnosluokka saa ajettaessa syötteenä alkuperäisaineiston kansion.
2. Aineistolle luodaan muunnoskäsittelijä Javassa, joka toteuttaa rajapinnan *fi.seco.paikka.convert.ConvertHandler*. Muunnosluokan nimeksi suositellaan *ConvertHandler<aineistokoodi>*. Muunnosluokalla aineistosta louhitaan paikkainstanssit, annetaan niille URI-tunnisteet ja luodaan paikoista RDF-kuvaukset, jotka sisältävät paikoista kaiken tiedon, joka on mahdollista tuoda aineistosta, ja jota voidaan kuvailla SUO-ontologian mukaisesti.
3. Aineisto muunnetaan RDF-muotoon ONKI-Paikan RDF-varastoon.
4. Aineiston RDF-varasto indeksoidaan hakuja varten.

Ensimmäisessä vaiheessa mainittu aineistopankki on yksinkertaisesti tietyn kaavan mukaisesti nimetty kansiorakenne, josta ONKI-Paikka löytää alkuperäisaineiston tarvittaessa. Kaava on muodoltaan

<juuri>/<aineistokoodi>/,

jossa <juuri> on aineistopankin juurikansio, joka on konfiguroitu ONKI-Paikkaan, ja <aineistokoodi> on aineistolle annettu tunnus, esimerkiksi A0008 Maanmittauslaitoksen Paikannimirekisterille. Käytännössä aineistopankkia tarvitaan vain uuden aineiston tuonnin yhteydessä, jolloin *ConvertHandler*-muunnosluokan on löydettävä muunnettava aineisto määrätyn kansiopolun alta. Aineiston kansion sisällä aineisto voi olla missä alkuperäisformaatissa tahansa. Tähän ei ONKI-Paikka ota kantaa. Ainoastaan aineiston muunnosluokan on osattava lukea paikka-aineiston alkuperäisformaattia.

6.2 Aineistojen RDF-muunnos

Paikkatietoaineistojen alkuperäisformaatti voi vaihdella suurestikin aineistosta toiseen. Tässä työssä käytetyt aineistot ovat kuitenkin kaikki tekstipohjaisia, CSV-muodossa olevia tiedostoja SAPO:n aineistoa lukuun ottamatta, joka on jo valmiiksi RDF:ää. SAPO:nkin aineistolle on kuitenkin tehtävä muunnos, sillä ONKI-Paikka olettaa RDF:n olevan tietyssä muodossa, jotta se voidaan lukea indeksointia varten. Siksi kaikkien paikkainstanssien RDF-kuvaukset tallennetaan tietyssä muodossa ONKI-Paikan omaan RDF-varastoon, josta paikkojen RDF-kuvaukset on nopeasti haettavissa. Jokaisen aineiston *ConvertHandler*-muunnosluokka täydentää RDF-varastoa aineistosta löytyneiden paikkainstanssien RDF-kuvauksilla.

6.2.1 RDF-varasto

Ontologisen paikkatietopalvelun ideana ei ole pelkästään tarjota helppoja ja tehokkaita hakutoimintoja paikkojen URI-tunnisteiden löytämiseksi. Vaikka tämä onkin ONKI-Paikan päätehtävä, on silti tärkeää myös saada paikkainstanssin koko kuvaus ontologisessa muodossa. ONKI-Paikka tarjoaa tätä varten rajapinnan, jonka kautta aineistoista peräisin olevien paikkainstanssien RDF-kuvaustiedostot voidaan hakea. Ideaalisinta olisi, jos

paikkainstanssien URI-tunnisteet toimisivat myös RDF-kuvaustiedostojen URL-osoitteina. Tämä on täysin mahdollista konfiguroida esimerkiksi *www.yso.fi*-osoitetta hallitsevassa HTTP-palvelimessa, jossa URL-uudelleenkirjoitussääntöjen avulla voidaan viitata ONKI-Paikan RDF-palvelun oletusosoitteeseen.

Oletusosoite ONKI-Paikan paikkainstanssien RDF-kuvaustiedostoille on:

<palvelun_URL>/rdf/<paikkainstanssin_lokaalinimi>,

jossa <palvelun_URL> on URL-osoite, jossa ONKI-Paikka-palvelu on saatavilla ja <paikkainstanssin_lokaalinimi> on URI-tunnisteen lokaalinimiosa paikkainstanssille, jonka RDF-kuvaus haetaan. Tällä hetkellä ONKI-Paikka-palvelun URL-osoite on http://demo.seco.tkk.fi/onkipaikka. Esimerkiksi taulussa 7 esitetyn Akaan kaupungin RDF-kuvaustiedoston URL-osoite olisi tässä esitetyn säännön mukaan http://demo.seco.tkk.fi/onkipaikka/rdf/A0007_020_20070101.

RDF-varaston populoimisen hoitaa jokaisen aineiston oma *ConvertHandler*-muunnos-luokka. Varaston tekniselle toteutukselle löytyy useita vaihtoehtoja. Lähinnä vaihtoehtoja löytyy siinä, miten paikkojen RDF-kuvaukset tallennetaan. Yksi vaihtoehto voisi olla sellainen, jossa jokaisen aineiston kaikkien paikkainstanssien RDF-kuvaukset tallennetaan yhteen RDF-tiedostoon. Ongelmana tämän lähestymistavan kanssa on, se että miljoonia paikkoja sisältävien aineistojen RDF-tiedostoista tulisi valtavia. Esimerkiksi GNS-aineiston alkuperäinen CVS-tiedosto, joka on saatavilla aineiston WWW-sivuilta, on kooltaan noin 900 megatavun kokoinen ja ZIP-pakattuna noin 210 megatavun kokoinen.

Muunnettuna yhdeksi isoksi RDF-tiedostoksi veisi aineisto vähintäänkin saman verran tilaa kuin alkuperäisaineisto. Näin isot tiedostot eivät itsessään ole ongelma. Epäkäytännöllisiksi isot tiedostot muuttuvat siinä vaiheessa, kun niistä pitää hakea pelkästään yhden tietyn paikkainstanssin RDF-kuvaus. Ensinnäkin hakeminen ei voisi tapahtua tekstitasolla, sillä haetun paikan URI voi sijaita useassa kohdassa RDF-tiedostoa riippuen siitä, mitä RDF-esitysmuotoa tiedostossa on käytetty. RDF-tiedosto pitäisi siis ensin tulkita RDF-tasolla, jotta haetun paikan RDF-kuvaus voitaisiin hakea.

Useiden gigatavujen kokoisten XML-dokumenttien käsittely muistissa olisi lähes mahdotonta ja yhden paikan RDF-kuvauksen hakeminen olisi näistä syistä johtuen aivan liian hidasta. Tämän takia yhden ison RDF-tiedoston vaihtoehto ei kelpaa ONKI-Paikan RDF-varaston tallennusmuodoksi. Toinen vaihtoehto olisi tallentaa jokaisen paikka-instanssin RDF-kuvaus erilliseen tiedostoon. Tässä vaihtoehdossa yhden aineiston kaikkien paikkainstanssien RDF-tiedostot voisivat sijaita yhden lähdeaineistokohtaisen kansion alla. Tässä vaihtoehdossa kuitenkin tulee vastaan tiedostojärjestelmien rajoitukset, jossa yhdessä kansiossa olevien tiedostojen määrä on rajoitettu. Tätä vaihtoehtoa testattiin Ubuntu Linux 7.04 -versiossa, jossa yhteen kansioon pystyi tallentamaan ainoastaan noin 65000 tiedostoa, jonka jälkeen tiedoston tallennuksen yritys tuotti virheilmoituksen.

Esimerkiksi GNS-aineiston noin 4 miljoonan paikkainstanssin kaikki RDF-tiedostot eivät millään mahtuisi yhteen kansioon näillä tiedostomäärien rajoituksilla. Voisi tietenkin jakaa tiedostot useaan kansioon, mutta silloin alakansioiden nimeämiselle pitäisi kehittää jokin johdonmukainen nimeämiskäytäntö, jonka mukaan olisi mahdollista esimerkiksi paikan

lokaalinimen perusteella löytää oikea kansiopolku, jonka alta paikkainstanssin RDF-tiedosto löytyy. Kansiopolkujen nimien hallinta toisi palveluun turhan kompleksisuustason, joka voisi olla vältettävissä jollain toisella RDF-varaston teknisellä toteutuksella.

Kolmas toteutusvaihtoehto perustuisi täysin tietokannan käyttöön. Ideana olisi se, että RDF-varaston tietokanta olisi yhteydessä aineistojen indeksointitietokantaan niin, että paikkainstanssien RDF-kuvaukset olisi haettavissa samojen tietokantahakujen kautta kuin mitä paikkahauissakin käytetään. RDF-kuvauksille luotaisiin tietokantaan taulu, josta oikean paikan RDF-kuvaus löytyisi esimerkiksi indeksitietokannan sisäisen paikka-tunnisteen avulla, joka mahdollistaisi RDF-kuvausten haun paikkainstansseille monimut-kaistenkin hakukyselyjen kautta. Voitaisiin esimerkiksi suoraan hakea RDF-kuvaukset paikoille, joilla on tietty nimi. Tämä olisi mahdotonta tehdä suoraan, jos paikkainstanssien RDF-tiedostot olisi tallennettu tiedostojärjestelmään. Tällöin pitäisi ensin tehdä tietokanta-haku, jolla palautetaan kaikkien löydettyjen paikkojen URI-tunnukset ja lokaalinimet. Vasta näiden avulla voitaisiin hakea paikkojen RDF-tiedostot tiedostojärjestelmästä.

Tietokantapohjaisella tallennuksella ei myöskään ole samoja rajoituksia kuin tiedosto-järjestelmillä. Yhden tietokantataulun tietueiden enimmäismäärä on miljardeissa, joten yhteen tietokantatauluun mahtuisi helposti ONKI-Paikan kaikkien aineistojen paikka-instanssien RDF-kuvaukset. Tietokantapohjainen tallennus RDF-kuvauksille on selvästi näistä kolmesta vaihtoehdoista paras, joten se on valittu ONKI-Paikan RDF-varaston tallennusalustaksi. RDF-varaston ja sen indeksoinnin tietokantaratkaisun kuvaus esitetään tarkemmin luvussa 6.3.

6.2.2 Päivitykset aineistoihin

Paikkainstanssien URI-tunnisteiden luomisesta, on paikkainstanssien määrittämisessä otettu lähtökohdaksi eri aineistoista peräisin olevien instanssien pitäminen omina instansseinaan. Tämä tarkoittaa sitä, että eri paikka-aineistoista peräisin olevat instanssit ovat helposti löydettävissä alkuperäisaineistosta aineiston sisäisen paikkatunnisteen avulla.

Tämä helpottaa huomattavasti myös paikkatiedon RDF-varaston ylläpitämistä silloin, kun alkuperäisaineistoihin tulee päivityksiä. Päivitetyt tiedot on helppo kohdistaa RDF-varaston, ja siitä luodun indeksin tietoihin alkuperäisaineiston tunnisteen ja aineiston sisäisen paikkatunnisteen avulla. Jos aineistoon on lisätty uusia paikkoja, on ne myös helppo tunnistaa uudesta tunnisteesta, joka ei löydy aiemmin luodusta RDF-varastosta.

Nopeinta aineiston päivittäminen on silloin kun pelkästään edellisestä versiosta muuttuneet tiedot on saatavilla. Jos saatavilla on pelkästään koko päivitetty aineisto, on se käytävä läpi paikka kerrallaan ja verrattava tiedot vanhoihin tietoihin, jotta muutokset löydettäisiin. Toinen vaihtoehto olisi poistaa RDF-varastosta koko aineisto ja ajaa se uudelleen sisään. Tässä vaihtoehdossa pitäisi kuitenkin ottaa huomioon se, että uudessa aineistossa ehkä on poistettu paikkoja, jotka olivat aiemmin olemassa aineiston vanhassa versiossa. Tämä voisi tuottaa ongelmia, jos poistetun paikan URI-tunnistetta on käytetty jossakin resurssien annotoinnissa.

Periaatteessa luotuja paikkainstansseja ei saisi RDF-varastosta poistaa, jos ne on kerran

sinne lisätty, koska ikinä ei voi tietää onko myöhemmin poistetun paikan URI-tunnistetta käytetty jonkin resurssin annotoinnissa. Aineistot voivat kuitenkin sisältää suoranaisia virheitä, kuten virheellisiä paikkojen nimiä tai tyyppejä. Tällöin paikkaa ei ole tarvetta poistaa, vaan korjataan ainoastaan paikan tiedot. Mutta jos paikan aineiston sisäisessä tunnisteessa on ollut virhe, niin silloin virheellinen tunniste on poistettava käytöstä. Jos virheelliseen tunnisteeseen pohjautuvaa paikan URI-tunnistetta on ehditty käyttää annotoinnissa, olisi kuitenkin hyvä saada tietoa siitä, että tunniste on todettu virheelliseksi ja ohjataan käyttäjää muuttamaan tunniste oikeaksi.

Ilmoitus vanhan URI-tunnisteen virheellisyydestä pitäisi tulla silloin, kun paikan URI:a käytettäisiin URL-osoitteena tai silloin, kun paikan URI-tunnisteen lokaalinimiosan perusteella haetaan paikkainstanssin RDF-kuvaus luvussa 6.2.1 esitetyn RDF-kuvauksen URL-osoitteen kautta.

6.3 Aineistojen indeksointi

Paikka-aineiston laajuuden ja koon takia aineistoa olisi ollut mahdotonta käsitellä kokonaisuudessaan muistissa esimerkiksi Jenan kautta. Tämän takia on päädytty ratkaisuun, jossa aineisto on tallennettu tekstitiedostoina ontologiapalvelun omaan RDF-varastoon. Hakuja ei kuitenkaan voida suorittaa aineistoon tiedosto- ja tekstitasolla eikä myöskään XML-tasolla RDF-formaatin eri mahdollisten lyhennettyjen syntaksien takia. Hakujen yhteydessä aineisto on siis tunnettava RDF-tasolla, jotta haettava tieto olisi löydettävissä.

Miljoonien paikkainstanssien ja niitä vastaavien RDF-tiedostojen reaaliaikainen selaaminen RDF-tasolla olisi täysin mahdotonta, sillä haut kestäisivät aivan liian pitkään. Myöskään koko aineiston lukeminen muistiin ja hakujen tekeminen muistissa ei testien perusteella juurikaan nopeuttanut hakuja. Perimmäisenä ongelmana hakujen yhteydessä on RDF-syntaksi ja subjekti-predikaatti-objekti-malli, jolla kaikki informaatio esitetään. Tämä malli on selvästi optimoitu tiedon mahdollisimman nopeaan lisäämiseen ja ilmaisuvoimaiseen esittämiseen, kun taas tiedon hakeminen ja löytäminen on hidasta (Abadi et al. 2007).

ONKI-Paikan yksi käyttäjärajapinta on paikkojen moninäkymähaku, jolla eri hakukriteerejä yhdistelemällä voidaan suorittaa hakuja RDF-varastoon. Aineiston valtavan laajuuden takia hakuja ei voida suorittaa suoraan paikkojen RDF-kuvauksiin, vaan hakuja varten on kehitetty RDF-kuvauksista tehty erillinen indeksitietokanta, jossa haut suoritetaan. Indeksi on toteutettu relaatiotietokannalla, jonne hakuja varten tarvittavat tiedot paikoista on tallennettu pelkästään nopeita hakuja varten. Indeksoitavat tiedot ovat ONKI-Paikan ensimmäisessä versiossa paikan nimi, paikan nimen kieli, paikkatyyppi, paikan koordinaatit, paikan olemassaolon alku- ja loppuaika sekä hierarkiapolku alueesta, jossa paikka sijaitsee *suo:isPartOf*-suhteella. Indeksiin on tallennettu myös tieto siitä, mistä aineistosta paikkainstanssi on peräisin. Näin voidaan hakuja kohdistaa pelkästään tiettyyn tai tiettyihin ONKI-Paikassa käytössä oleviin aineistoihin.

Indeksitietokannan ideana on tallentaa paikkatiedot hakuja varten mahdollisimman tiiviiseen muotoon. Pelkästään hauissa tarvittava tieto indeksoidaan ja tieto tallennetaan sellaisessa muodossa, että se mahdollistaa relaatiotietokannassa mahdollisimman nopeat haut. Paikoille luodaan oma tietokantataulu sekä relaatiotaulu, jolla paikkoja yhdistetään jonkin ontologisen suhteen avulla toisiinsa. Paikkatietoaineistosta pyritään myös löytämään paikkoihin liittyvät, usein toistuvat tiedot, joille luodaan omat tietokantataulut hakuja varten.

Yksi paikoilla usein toistuva tieto on paikan tyyppi, jolle luodaan oma taulunsa tietokantaan. Koska paikka voi olla vain yhtä tyyppiä, lisätään paikkojen tauluun viittaus tyyppien taulun tyyppitunnisteeseen. Paikkojen nimille on myös luotu oma taulunsa. Koska paikalla voi olla useita nimiä, on nimen kohdalla viittaus paikkojen taulun paikka-tunnisteeseen. Paikkojen nimiä ei voida tallentaa pelkästään merkkijonoina niin, että nimien taulussa yksi nimi esiintyisi pelkästään kerran. Paikan nimeen liittyy myös tieto paikan nimen virallisuudesta, joka RDF-tasolla esitetään *skos:prefLabel*- ja *skos:altLabel*-ominaisuuksien avulla (katso luku 5.1.2). Nimien taulussa tietue siis koostuu sekä paikan nimen merkkijonosta että sen virallisuudesta kertovasta tiedosta.

Jokainen paikka sisältää pakollisen tiedon siitä, minkä paikan kanssa sillä on *suo:isPartOf*-suhde. Tällä tavalla voidaan hakea kaikki paikat tietyn paikan alta. Voidaan esimerkiksi löytää kaikki järvet Espoon kaupungin alueella. Hakukäyttöliittymän aluerajaus ei kuiten-kaan saa rajoittua pelkästään yhteen hierarkiatasoon, vaan pitää olla mahdollista myös löytää esimerkiksi kaikki järvet Uudenmaan maakunnan alueella. Koska *suo:isPartOf* on transitiivinen ominaisuus, tarkoittaisi tällainen haku aluehierarkian rekursiivista läpi-käymistä Uudenmaan maakunnasta alaspäin. Maakunnathan koostuvat seutukunnista, jotka puolestaan koostuvat kunnista.

Rekursiivisia hakuja on kuitenkin hankala tehdä relaatiotietokannassa, joten tällaisia aluerajaushakuja varten luotiin paikkatunnisteista koostettu tieto paikan *suo:isPartOf*-hierarkiapolusta aina ylimpään tasoon eli maailmaan asti. Hierarkiapolusta muodostuu paikalle tieto aivan kuten paikan nimestäkin. Koska monella paikalla on sama hierarkia-polku, on myös tämä usein toistuva tieto, jolle luodaan indeksitietokannassa oma taulunsa. Paikkatietotauluun tallennetaan viittaus hierarkiapolkutauluun, johon voi tehdä suoraan hakuja. Jos polku sisältää haetun aluerajauspaikan tunnisteen, tarkoittaa se sitä, että paikka on joko suoraan tai välillisesti *suo:isPartOf*-suhteessa rajausalueen kanssa, jolloin haku palauttaa kaikki rajausalueen sisällä olevat paikat. Indeksi siis koostuu sekä RDF-aineistosta suoraan kerätystä paikkatiedosta että hakuja varten erikseen käsitellystä ja koostetusta tiedosta. Paikkojen hierarkiapolkujen luominen on ainoa semanttinen päättely, jota suoritetaan paikkatietoon indeksointivaiheessa.

Kaikki paikkatietoaineistoon tehtävät haut tapahtuvat indeksitietokannassa, jolloin indeksin on sisällettävä kaikki hakutuloksen näyttämiseen tarvittava informaatio. Hakukäyttö-liittymä esitetään tarkemmin luvussa 6.5, jossa moninäkymähaun eri osat esitellään. Hakutuloksen on minimissään sisällettävä löydettyjen paikkojen URI-tunnisteet, joiden avulla voidaan hakea paikkojen muut tiedot RDF-varastosta. Kuitenkin hakutuloksen mahdollisimman nopean esittämisen takia ei olisi järkevää aina joutua tuloksen esittämistä

varten hakemaan paikkojen muut tiedot RDF-tiedostoista. Tämän takia hakutuloksessa palautetaan aina kaikki paikkaan liittyvät indeksoidut tiedot. Indeksointitietokannan rakenne on esitetty yksityiskohtaisesti taulukossa 11.

Taulukko 11. *Indeksitietokannan taulut sekä taulujen kentät kuvauksineen*

Kenttä	Kuvaus
place Paikkainstanssin taulu, jossa paikkainstansseille yksilölliset tiedot säilytetään	
place_id	Indeksointitietokannan sisäinen tunniste paikalle. Viittaa *place_rdf*-taulun id-kenttään.
namespace	Paikan nimiavaruus. Jos nimiavaruus on tyhjä, käytetään ONKI-Paikan oletusnimiavaruutta, joka on sama kuin SUO-ontologian nimiavaruus, eli http://www.yso.fi/onto/suo/.
local_name	Paikkainstanssin lokaalinimi.
place_type_id	Viittaus paikkatyyppiin *place_type*-taulun id-kenttään.
coordinate1	Koordinaattipisteen leveyspiirikoordinaatti WGS84-desimaalina.
coordinate2	Koordinaattipisteen pituuspiirikoordinaatti WGS84-desimaalina.
coordinate3	Korkeus metreinä, jos tieto on saatavilla. Jos arvo ei ole saatavilla on kentän arvo 0.
time_begin	Paikan perustamisajankohta UNIX-aikaleimana.
time_end	Paikan lakkauttamisajankohta UNIX-aikaleimana.
place_ancestor_path_id	Viittaus *place_ancestor_path*-taulun id-kenttään.
source_id	Aineistokoodi lähdeaineistolle, josta paikkainstanssi on peräisin.
place_ancestor_path Paikkahierarkiapolut paikkainstansseille	
id	Indeksointitietokannan sisäinen tunniste paikkahierarkiapolulle.
path	Polku, joka muodostetaan paikkainstanssien sisäisistä *place*-taulun id-tunnisteista liittämällä tunnisteet yhteen eroteltuna kaksoispisteellä niin että polku myös alkaa ja loppuu kaksoispisteeseen. Polun alussa ei ole maailman paikkainstanssin sisäistä id-tunnistetta. Polku on siis muotoa ”:A:B:C:D:”.
count	Paikkainstanssien määrä, joilla on tämä paikkahierarkiapolku.
place_name Paikannimien tiedot sisältävä taulu	
id	Indeksointitietokannan sisäinen tunniste paikannimelle.

label	Paikan nimi.
language_code	Kielen kolmikirjaiminen ISO 639-3-kielikoodi.
is_pref_name	Tämä kenttä määrittelee sen, onko tämä nimi paikan virallinen ensisijainen nimi. Kentällä on joko arvo 0 tai 1.
place_id	Viittaus *place*-taulun id-kenttään. Paikka, jolla on nimi.
place_name_language Paikkojen nimien kielten tiedot sisältävä taulu	
id	Indeksointitietokannan sisäinen tunniste paikan nimen kielelle.
language_code	Kielen kolmikirjaiminen ISO 639-3-kielikoodi.
count	Paikannimien määrä, jotka ovat tätä kieltä.
place_rdf Paikkojen RDF-kuvaukset sisältävä taulu (RDF-varasto)	
id	Indeksointitietokannan sisäinen tunniste paikalle.
rdf	Paikkainstanssin RDF-kuvaus.
source_id	Aineistokoodi lähdeaineistolle, josta paikkainstanssi on peräisin.
place_relation Paikkojen välisten relaatioiden taulu	
id	Indeksointitietokannan sisäinen tunniste paikkarelaatiolle.
namespace	Relaation nimiavaruus. Jos nimiavaruus on tyhjä, käytetään oletusnimiavaruutta.
local_name	Relaation lokaalinimi.
place_id1	Viittaus *place_rdf*-taulun id-kenttään. Paikka, jolla on relaatio.
place_id2	Viittaus *place_rdf*-taulun id-kenttään. Paikka, johon relaatio kohdistuu.
place_type Paikkatyyppien tietojen taulu	
id	Indeksointitietokannan sisäinen tunniste paikkatyypille.
namespace	Paikkatyypin nimiavaruus. Jos nimiavaruus on tyhjä, käytetään oletusnimiavaruutta.
local_name	Paikkatyypin lokaalinimi.
count	Paikkainstanssien määrä, jotka ovat tätä tyyppiä.

Paikkatietojen indeksointi on ONKI-Paikassa toteutettu MySQL-tietokannan[60] avulla. Tietokantaohjelmistona voisi kuitenkin toimia mikä tahansa muu relaatiotietokanta. MySQL valittiin etenkin avoimuutensa ja ilmaisuutensa mutta myös laajan tukensa takia. Siksi MySQL-tietokantaan on helppo käyttää muun muassa Java-sovelluksen kautta. Paikkojen nimien tekstihaussa on hyödynnetty MySQL-tietokannan *Full-Text*-haku-ominaisuutta[61]. Ominaisuus nopeuttaa huomattavasti hakuja merkkijonoista silloin, kun haettu merkkijono on yli kolme merkkiä pitkä ja haku kohdistuu merkkijonojen alkuun. Tämä sopii varsin hyvin paikannimihakuihin, joissa yleensä haku nimenomaan kohdistuu paikan nimen alkuun.

Full-Text-haku toimii niin, että haku kohdistuu merkkijonon kaikkiin sanoihin. Sanat erottuvat toisistaan esimerkiksi välilyönnin tai viivan avulla. Muitakin merkkejä tulkitaan *Full-Text*-haussa sanavälimerkkeinä, mutta paikannimissä ainoat sanavälimerkit ovat juuri välilyönti ja viiva. Haku kohdistuu aina myös pelkästään sanan alkuun, sillä ”jokerimerkkinä” toimivan asteriskin voi sijoittaa vain haetun merkkijonon loppuun. Full-Text indeksiä käyttävä haku, jolla haetaan kaikkia ”Pyhä”-alkuisia paikannimiä näyttää MySQL:ssä seuraavanlaiselta:

```
select * from place_names where match(label) against('Pyhä*' in boolean mode);
```

Tietokannan *place_names*-taulun *label*-sarakkeesta on luotu sekä FULLTEXT- että BTREE-indeksit. *Full-Text*-haku käyttää hyödykseen FULLTEXT-indeksiä ja *like*-haku käyttää BTREE-indeksiä. Jotta saataisiin varmuutta siitä, onko *Full-Text*-haku varmasti ONKI-Paikan tarpeita ajatellen nopein vaihtoehto paikannimien hakuun, suoritettiin testi, jossa *Full-Text*-hakua verrattiin vastaaviin *like*- ja *regexp*-hakuihin. Vastaavan *like*-haun *where*-lauseke on yllä esitetylle ”Pyhä”-alkuisille paikannimille seuraavanlainen:

```
label like 'Pyhä%' or label like '% Pyhä%' or label like '%-Pyhä%'
```

Vastaava regexp-haku voidaan kirjoittaa kahdella eri tavalla. Toinen käyttää hyväkseen yhtä säännöllistä lauseketta ja toinen on yllä esitetyn *like*-haun vastine, jossa jokaisella eri sanavälimerkkivaihtoehdolla on oma lausekkeensa. Yllä esitetyn like-haun vastine regexp-haulla on:

```
label regexp '^Pyhä(.*)' or label regexp '(.*) Pyhä(.*)' or label regexp '(.*)-Pyhä(.*)'
```

Sama haku voidaan kirjoittaa tiiviimpään muotoon säännöllisten lausekkeiden sääntöjen mukaisestti:

```
lable regexp '(^|(.*) |(.*)-)Pyhä(.*)'
```

Näitä neljää yllä esitettyä hakulausekevaihtoehtoa verrataan keskenään eri merkkijonoille. Ennen jokaista hakua tyhjennetään MySQL:n kyselyvälimuisti, jotta haut käyttäisivät puhtaasti vain tietokannan indeksejä eikä välimuistia. Hakumerkkijonoista on valittu sekä pitkiä että lyhyitä sekä sellaisia, jotka täsmäävät moneen sekä sellaisia, jotka täsmäävät vain muutamaan paikannimeen. Testi on suoritettu tietokoneella, jolla ei ole samanaikaisesti muita käyttäjiä, eikä muita konetta kuormittavia prosesseja. Haku suoritettiin Paikan-

60 MySQL, http://www.mysql.com/

61 http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html

nimirekisteristä peräisin olevista noin 800 000 paikannimestä. Yksittäiset haut on suoritettu sata kertaa, joista on laskettu hakuaikojen keskiarvo. Taulukossa 12 on esitetty testihakujen tulokset.

Taulukko 12. *Eri SQL-kyselyvaihtoehtojen nopeuksien vertailu. Hakuajat on esitetty sekunneissa. Lukumäärä kertoo haulla löydettyjen paikannimien lukumäärän. Regexp 2 on säännöllinen lauseke yhdellä lausekkeella.*

Merkkijono	Lukumäärä	Full-Text	Like	Regexp 1	Regexp 2
Helsinki	9	0,00	0,96	2,85	5,34
Pyhä	409	0,02	0,95	2,52	5,04
Saa	4521	0,11	0,98	2,43	4,90
Hevosen	49	0,00	0,96	2,72	5.24
Ålö	8	0,00	0,95	2,51	5,09

Mielenkiintoinen havainto testistä on se, että yksi monimutkaisempi säännöllinen lauseke on noin kaksi kertaa hitaampi kuin kolmen yksinkertaisemman säännöllisen lausekkeen käyttö yhdessä kyselyssä. *Full-Text*-hakua vastaava *like*-haku oli varsin nopea *regexp*-hakuun verrattuna. Ylivoimaisesti nopein oli kuitenkin MySQL:n *Full-Text*-indeksiä hyödyntävä haku, joka suoritti vähän paikkoja löytäviä hakuja alle sekunnin sadasosan. Nopeutensa sekä paikannimihakuihin soveltuvien hakuominaisuuksiensa takia *Full-Text*-haku on valittu ONKI-Paikan oletushakumenetelmäksi. Koska *Full-Text*-haulla ei pysty hakemaan yhtä monipuolisilla hakurakenteilla kuin *regexp*-haulla, annetaan käyttäjälle mahdollisuus valita käyttääkö hän *Full-Text*- vai *regexp*-hakua.

6.4 Web Service -rajapinta

ONKI-Paikan ydin ei itsessään tarjoa minkäänlaista graafista käyttöliittymää esimerkiksi paikkojen hakemiseen. Kommunikointi palvelun ja asiakassovellusten välillä tapahtuu verkossa HTTP-protokollan kautta, kahden erilaisen Web Services -rajapinnan kautta. ONKI-Paikka tarjoaa asiakassovelluksilleen kaksi eri rajapintaa hakujen tekemiseen verkon yli. Toinen rajapinnoista käyttää hyväkseen selaimissa käytettävää Javascriptiä ja siinä toimivaa Ajax-teknologiaa. Toinen rajapinnoista toimii XML-tasolla viestien välittämiseen SOAP ja WSDL-protokollien avulla.

6.4.1 Rajapinta yleisellä tasolla

Riippumatta teknisestä toteutustavasta on paikkojen hakurajapinnalla aina käytössä tietyt metodit, jotka toimivat toteutustavasta riippumatta samalla tavalla. Näiden metodien avulla pitää olla mahdollista suorittaa monimutkaisiakin hakua kaikkien palveluun rekisteröityjen paikkojen seasta. Myös yksittäisille paikoille voidaan suorittaa kyselyjä, joiden avulla on mahdollista saada selville tietyn paikan semanttiset suhteen muihin paikkoihin.

ONKI-Paikan Web Service -paikkahakupalvelua kutsutaan tässä dokumentissa nimellä

PlaceFinder. Tämä on myös palvelun oletusnimi, jonka voi muuttaa paikkahakupalvelimen asetuksista. Teknisesti koko palvelu koostuu yhdestä Java-luokasta, jonka metodit tarjoavat kaiken sen, mitä tarvitaan asiakassovelluksen paikkahakukäyttöliittymän rakentamiseen. Alla on taulukkomuodossa lueteltu palvelun metodit, niiden argumentit, sekä metodien palauttamien arvojen rakenne.

Jokaisen Web Service -rajapinnan metodin ensimmäinen argumentti on kielikoodi, sille kielelle, jolla asiakassovelluksen käyttöliittymää käytetään. Esimerkiksi *search*-metodin palautusarvossa oletusarvoisesti hakutulos järjestellään paikan ensimmäisen virallisen nimen mukaan. Jos käyttöliittymää käytetään esimerkiksi ruotsin kielellä, voidaan hakutulos järjestetään paikkojen ruotsinkielisten nimien mukaan. Hakutuloksessa annetaan myös paikan tyypin nimi, joka palautetaan sillä kielellä, mikä on tässä argumentissa määritelty. Kieli määritellään aina kolmikirjaimisella ISO 639-3-kielikoodilla.

Taulukko 13. *Web Service -rajapinnan metodien kuvaukset*

getLanguages(String *uiLanguage*)	
Kuvaus	Metodilla haetaan lista kaikista paikan nimien kielistä lisätietoineen.
Argumentit	*uiLanguage*: Kieli, jolla asiakassovelluksen käyttöliittymää käytetään.
Palautusarvo	Metodi palauttaa *FinderLanguage*-olioista koostuvan listan, järjestetty kielen nimen mukaan sillä kielellä, joka on annettu *uiLanguage*-argumentissa.
getPlace(String *uiLanguage,* String *placeURI*)	
Kuvaus	Metodilla haetaan tietyn paikan olio paikan URI-tunnisteen perusteella.
Argumentit	*uiLanguage:* Kieli, jolla asiakassovelluksen käyttöliittymää käytetään. *placeURI*: Haetun paikan URI-tunniste.
Palautusarvo	Metodi palauttaa *FinderPlace*-olion.
getRelatedPlaces(String *uiLanguage,* String *placeURI*, String *relationURI*)	
Kuvaus	Metodilla haetaan tiettyyn paikkaan jollakin semanttisella suhteella liittyviä muita paikkoja. ONKI-Paikassa yleisin semanttinen suhde on *http://www.yso.fi/onto/suo/isPartOf*, jolla kuvataan etenkin hallinnollisten alueiden hierarkkisia suhteita toisiinsa. Esimerkiksi Suomessa jokin tietty kunta on osa jotakin seutukuntaa. Metodi ei suorita paikkojen haussa päättelyä, joten transitiiviset suhteet kuten SUO-ontologian *isPartOf* palauttavat pelkästään suorat naapurit suhdeverkossa.
Argumentit	*uiLanguage:* Kieli, jolla asiakassovelluksen käyttöliittymää käytetään. *placeURI*: URI-tunniste paikalle, jonka relaatiot haetaan. *relationURI*: Semanttisen relaation URI-tunniste.
Palautusarvo	Metodi palauttaa *FinderPlace*-olioista koostuvan listan, järjestetty paikan nimen mukaan sillä kielellä, joka on annettu *uiLanguage*-argumentissa.
getTypes(String *uiLanguage*)	

Kuvaus	Metodilla haetaan lista kaikista SUO-ontologian paikkatyypeistä lisätietoineen.
Argumentit	*uiLanguage:* Kieli, jolla asiakassovelluksen käyttöliittymää käytetään.
Palautusarvo	Metodi palauttaa *FinderType*-olioita koostuvan listan, joka on järjestetty paikkatyypin nimen mukaan sillä kielellä, joka on annettu *uiLanguage*-argumentissa.
search(String *uiLanguage,* HashMap<String,String> *searchParameters*)	
Kuvaus	Tämä metodi on koko paikkahakutoiminnon ydin. Metodille syötetään hakuehdot, joiden perusteella ONKI-Paikasta palautetaan hakuehtoihin täsmäävät paikat, sekä tilastoa haun tuloksesta, kuten löydettyjen paikkojen määrä ja määrät paikkatyypeittäin.
Argumentit	*uiLanguage:* Kieli, jolla asiakassovelluksen käyttöliittymää käytetään. *searchParameters*: Lista hakuparametreista arvoineen, joilla paikkahakua rajataan. Kaikki hakuparametrit ovat valinnaisia, mutta liian väljästi rajattu haku saattaa palauttaa enemmän paikkoja, kuin on sallittua. Palvelun kuormituksen hillitsemiseksi palautettavien paikkojen määrä on rajattu 2000 paikkaan jokaista hakutulosta kohden. Metodin hakuparametrit ovat seuraavat: ● *nameFragment*: Paikan nimen osa. ● *isRegexp*: Jos tämä parametri on totuusarvoltaan tosi, tarkoittaa se sitä, että parametri *nameFragment* tulkitaan haussa säännölliseksi lausekkeeksi. Tällöin paikan nimi haetaan *nameFragment*-parametrin sisältävän säännöllisen lausekkeen mukaan. ● *coordinatePolygon*: Lista koordinaattipisteitä, jolla rajataan haku koskemaan tiettyä aluetta monikulmion sisällä. Parametrin arvona on lista WGS84 desimaalimuodossa olevia leveyspiiri-pituuspiiri-pareja, jotka muodostavat maanpinnalla monikulmion. ● *nameLanguage*: Lista haettujen paikannimien kielten kolmikirjaimisista ISO 639-3-kielikoodeista. ● *placeType*: Lista haettujen paikkojen SUO-ontologian paikkatyypeistä. Parametrin arvona on lista paikkatyyppien URI-tunnisteista. ● *source*: Lista aineistojen koodeista, joista haku tehdään ● *timeBegin*: Tämä parametri määrää haun koskemaan ainoastaan paikkoja, jotka ovat olleet olemassa määrätyn ajanhetken jälkeen. Jos paikalle ei ole määritelty perustamis- tai lakkauttamisaikaa oletetaan haussa paikan olleen aina olemassa. ● *timeEnd*: Määrää haun koskemaan paikkoja, jotka ovat olleet olemassa ennen määrättyä ajanhetkeä. ● *partOfPlace*: Hallinnollinen alue, jonka sisältä paikkoja haetaan. Parametrin arvona on hallinnollisen alueen URI-tunniste.

Palautusarvo	Metodi palauttaa *FinderResult*-olion, jossa paikka-olioiden lista on järjestetty paikan nimen mukaan sillä kielellä, joka on annettu *uiLanguage*-argumentissa.

Paikkahaun yhteydessä käytettävät *FinderLanguage*-, *FinderPlace*-, *FinderResult*- ja *FinderType*-oliot ovat staattisen datan esittämiseen tarkoitettuja, monimutkaisempia tietovarastoja, joita ei voida esittää suoraan esimerkiksi Javan *HashMap*-olion avulla. ONKI-Paikan sisäinen *Place*-olio ei myöskään suoraan sovellu Web Service -palvelun vastauksen tietotyypiksi, sillä se ei toteuta JavaBean-mallia[62] (Hamilton 1997). Jotta olioiden ominaisuuksien hakeminen olisi helppoa Web Service -rajapinnan kautta käytetään olioissa JavaBean-mallia olioiden ominaisuuksien asettaja- ja hakijametodien nimeämisessä.

Taulukko 14. *Web Service -rajapinnassa käytettyjen olioiden ominaisuuksien kuvaukset*

Ominaisuuden nimi	Ominaisuuden kuvaus
FinderLanguage	
label	Kielen nimi hakusovelluksen käyttöliittymän kielellä.
code	Kielen ISO 639-3-kielikoodi.
count	Tällä kielellä olevien paikkojen nimien lukumäärä ONKI-Paikan tietokannassa.
FinderPlace	
label	Paikan virallinen nimi hakusovelluksen käyttöliittymän kielellä. Jos paikalla ei ole nimeä annetulla kielellä, saa tämä ominaisuus arvokseen paikan nimen sillä kielellä, joka on virallinen enemmistön kieli paikan sijaintialueella. Ominaisuus on muodoltaan lista, jossa ensimmäinen elementti, avaimella 0, on nimi ja toinen elementti, avaimella 1, on nimen kieli ISO 639-3-kielikoodilla.
altLabels	Lista paikan muista virallisista nimistä, esimerkiksi muilla kielillä. Listan elementit ovat samaa muotoa kuin *label*-ominaisuus.
URI	Paikan URI-tunniste.
typeURI	Paikan tyypin URI-tunniste.
typeLabel	Paikan tyypin nimi sillä kielellä, joka on annettu argumenttina *search*-metodiin.
coordinate	Lista koordinaateista WGS84-desimaalimuodossa. Listan avain 0 on leveyspiirikoordinaatti ja avain 1 on pituuspiirikoordinaatti. Jos paikan koordinaattipisteelle on määritelty korkeus annetaan se

62 Java SE Desktop Technologies - Java Beans, http://java.sun.com/beans

	listassa avaimella 2. Korkeus ilmoitetaan aina metreinä.
coordinateURI	Koordinaattipisteen URI-tunniste.
timeBegin	Paikan perustamisajankohta, jos tieto saatavilla.
timeEnd	Paikan lakkauttamisajankohta, jos tieto saatavilla.
partOfURI	URI-tunniste alueelle, jonka sisällä paikka sijaitsee.
partOfLabel	Nimi alueelle, jonka sisällä paikka sijaitsee, sillä kielellä, joka on annettu toisena argumenttina *search*-metodiin.
FinderResult	
defaultNamespace	Oletusnimiavaruus kaikille URI-tunnisteille. ONKI-Paikassa tämä on sama kuin SUO-ontologian nimiavaruus, eli *http://www.yso.fi/onto/suo/*.
placeList	*FinderPlace*-olioista koostuva lista.
placeCount	Haulla löydettyjen paikkojen lukumäärä.
maxPlaceCount	Enimmäismäärä paikkoja, jotka palautetaan yhdellä haulla haun suorittavalle sovellukselle. Jos löydettyjen paikkojen lukumäärä ylittää tämän arvon tulee *placeList* olemaan tyhjä.
langStatistics	Lista niiden paikkojen nimien kielistä, jotka löytyivät haulla. Listan avaimena on kielen kolmikirjaiminen ISO 639-3-kielikoodi.
typeStatistics	Lista niiden paikkojen SUO-ontologian tyypeistä, jotka löytyivät haulla. Listan avaimena on tyypin lyhennetty URI-tunniste, jossa SUO-ontologian nimiavaruus on lyhennetty *suo*.
sqlQuery	Tietokantaan tehdyn haun SQL-kysely. Tämä on lähinnä virhetilanteiden selvittämistä varten oleva ominaisuus.
FinderType	
label	Paikkatyypin nimi hakusovelluksen käyttöliittymän kielellä.
URI	Paikkatyypin URI SUO-ontologiassa.
count	Tätä tyyppiä olevien paikkojen lukumäärä ONKI-Paikan tietokannassa.

6.4.2 Javascript / Ajax -rajapinta

Kuten monet verkossa toimivat sovellukset nykyään, toimii myös ONKI-Paikan oma hakukäyttöliittymä selaimessa. Selaimesta on yhä useammin tulossa hyvin monipuolistenkin sovellusten ajoympäristö. Tässä kehityksessä tärkeä osa on ollut Javascriptin yhä parempi tuki eri selaimissa. Etenkin viime vuosina Javascriptiin perustuvan Ajax-tekniikan yleistymisen ansiosta selaimet pystyvät päivittämään sisältöä sivulle lataamatta koko dokumenttia uudelleen selaimen ikkunaan. Tämä tekee selaimessa toimivan

sovelluksen käytöstä miellyttävämpää ja nopeamman tuntuista kun koko näkymää ei tarvitse päivittää jokaisella hiiren painalluksella.

Myös ONKI-Paikka tarjoaa yhtenä vaihtoehtona Ajax-tekniikkaan perustuvaa rajapintaa. Rajapinta on toteutettu DWR:n avulla. DWR, eli Direct Web Remoting[63] on Ajax[64]-toteutus, jolla palvelimen Java-luokkia voidaan tuoda selaimessa Javascriptin käyttöön. Palvelimen Java-luokista ja luokkien metodeista viedään selaimelle Javascript-versiot, joita voi selaimessa käyttää esimerkiksi graafisen paikkahaun käyttöliittymän rakentamiseen. Kommunikaatio hakusovelluksen ja palvelimen välillä hoituu automaattisesti DWR:n peruskirjaston kautta, jolloin asiakassovelluksen toteuttaja voi toimia täysin Javascriptin oliotasolla, välittämättä siitä, miten palvelimella olevasta tiedosta luodaan oliot Javascriptin tasolla.

Perinteisesti termillä ”Web Service” tarkoitetaan SOAP / WSDL -tyylisiä XML-rajapintoja, joiden avulla sovellusten on mahdollista keskustella keskenään verkon yli. W3C kuitenkin määrittelee termin ”Web Service” yksinkertaisesti ohjelmallisena rajapintana, jonka kautta sovellukset pystyvät kommunikoimaan keskenään verkon yli[65]. DWR sopii siis tähän kuvaukseen täydellisesti, sillä se tarjoaa nimenomaan rajapinnan, jolla selaimessa Javascriptillä toteutettu sovellus pystyy suoraan kommunikoimaan palvelimella Javalla toteutettu sovelluksen kanssa.

DWR-rajapinta tuo asiakassovelluksen Javascriptin käyttöön *PlaceFinder*-nimisen olion, jonka sisältämät metodit tarjoavat kaiken tarvittavan ONKI-Paikkaa hyödyntävän paikkahakukäyttöliittymän rakentamiseen. Kaikki *PlaceFinder*-olion metodit on lueteltu aiemmin taulukossa 13, jossa määritellään ONKI-Paikan Web Service -rajapinnan rakenne. Koko palvelun ytimenä on *search*-niminen metodi, jolla varsinaiset paikkahaut suoritetaan. Selaimessa toimivalle, Ajax-tekniikkaa käyttävälle sovellukselle tärkeitä ovat myös hakukäyttöliittymän rakentamiseen tarvittavat metodit *getLanguages* ja *getTypes*. Näiden metodien avulla on mahdollista rakentaa kieli- ja paikkatyyppivalikot käyttöliittymään, joista saadaan oikeat arvot *search*-metodin haunrajausparametreihin.

6.4.3 SOAP / WSDL -rajapinta

Etenkin DWR:n mahdollistaman rajapinnan avulla on varsin helppoa ottaa ONKI-Paikan palvelut käyttöön muissakin web-palveluissa *mash-up*-tyylisesti Web 2.0-hengen mukaisesti (Viljanen et al. 2008). Paikkatietoa voidaan helposti sisällyttää toisiin, selaimissa toimiviin ja Javascriptillä toteutettuihin palveluihin. DWR ei kuitenkaan sovellu rajapinnaksi sellaisiin asiakassovelluksiin, joissa ei ole käytössä Javascript tai tuki Ajax-toiminnoille. Sovellukset eivät välttämättä aina ole rakennettu toimimaan selaimessa, jolloin näille on tarjottava toisenlainen rajapinta ONKI-Paikan palvelimen kanssa kommunikointiin. Näille sovelluksille yleensä luontevin vaihtoehto on XML-pohjainen SOAP / WSDL -rajapinta. Siinä tieto palvelusta ja sen metodeista kuvataan viesteissä WSDL:n, eli Web Service Description Language -kuvauskielen (Christensen et al. 2001)

63 Direct Web Remoting, http://getahead.org/dwr

64 Asynchronous JavaScript and XML (AJAX), http://developer.mozilla.org/en/docs/AJAX

65 Web Services Activity, http://www.w3.org/2002/ws/

avulla. Itse metodikutsut ja niiden tuottamat tulokset välitetään asiakas- ja palvelinsovellusten välillä SOAP, eli Simple Object Access Protocol -viestien (Gudgin et al. 2007) avulla.

Rajapinta on ONKI-Paikassa toteutettu Apachen Axis[66]-nimisen SOAP-toteutuksen avulla. Axis on tekniikka, jolla pystyy automaattisesti, tai pienten apuluokkien avulla muuntamaan Java-luokkia suoraan Web service -palveluiksi. Axis hoitaa automaattisesti WSDL-dokumentin luomisen luokan tiedoista. WSDL kuvaa tarkasti palvelun metodit, niiden parametrit sekä parametrien että palautusarvojen tyypit. Näiden tietojen avulla voidaan rakentaa sovelluksia, jotka hyödyntävät palvelun julkaisemia metodeja. Palvelu on rakenteeltaan täysin samanlainen kuin yllä DWR-toteutuksessa. Itse asiassa Axis käyttää täysin samoja JavaBean-rajapintaa toteuttavia luokkia kuin DWR palvelun rajapinnan toteutuksessa. Tämän ansiosta luokan metodit ja argumentit ovat täysin samat molemmissa Web Service -rajapintojen vaihtoehdoissa.

6.5 Hakukäyttöliittymä annotoinnin apuvälineenä

Web Service -rajapintojen lisäksi tämän työn tuloksena on kehitetty paikkahauille graafinen käyttöliittymä, joka on eräänlainen esimerkki siitä, miten Javascript / Ajax -rajapintaa voidaan hyödyntää selainpohjaisessa sovelluksessa. Hakukäyttöliittymä hyödyntää kaikkia niitä ominaisuuksia, joita hakurajapinta tarjoaa asiakassovelluksen käyttöön. Kuten jo aiemmin Web Service -rajapinnan kuvauksessa todettiin, on *search*-metodi koko rajapinnan ydin, jolla haut suoritetaan ONKI-Paikka tietokantaan. Pelkällä tekstipohjaisella hakutuloslistalla on varmasti mahdollista löytää haettu paikka, mutta varsinaisen tehon paikkahakuihin tuo visuaalinen tulosten selailu muun muassa kartan avulla, jossa on heti nähtävissä, missä jokin löydetty paikka sijaitsee.

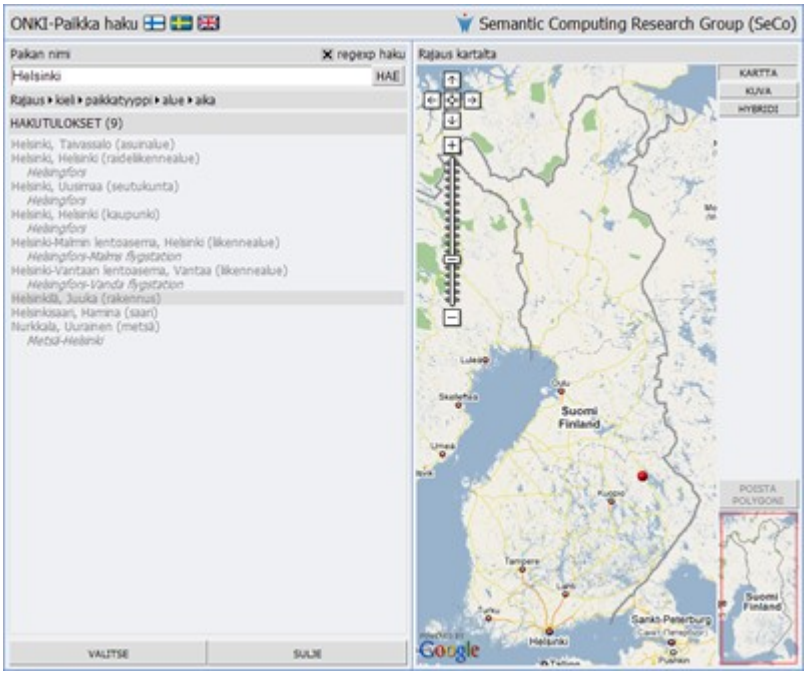
Hakukäyttöliittymän pääasiallinen käyttötarkoitus on olla resurssien annotoijan apuväline paikkainstanssien URI-tunnisteiden noutamiseen weblomakkeeseen. Lomakkeessa voisi esimerkiksi olla tekstinsyöttökenttä paikan URI-tunnisteelle, sekä painike syöttökentän vieressä, josta hakukäyttöliittymä avautuu tunnisteen noutamista varten. Kuvassa 2 on esitetty, millaiselta paikan URI:n syöttökenttä voisi näyttää annotaatiojärjestelmän ylläpitolomakkeessa.

Kuva 2. *Esimerkkikenttä kuvitteellisesta annotaatiojärjestelmän lomakkeesta, jossa syöttökenttään on valittu jonkin paikan URI. Uutta arvoa kentälle voidaan hakea kuvassa 3 esitetyllä hakukäyttöliittymällä, joka avautuu klikkaamalla syöttökentän vieressä olevaa painiketta.*

Eri hakurajausvaihtoehtoja yhdistelemällä on juuri hakemansa paikan löytäminen tehty mahdollisimman helpoksi. Hakukäyttöliittymä tukee kaikkia niitä hakujen rajauksia, jotka Web Service -rajapinta tarjoaa. Mikään hakujen rajausvaihtoehdoista ei ole pakollinen, aivan kuten rajapinnassa on määritelty. Palvelimen kuormituksen hillitsemiseksi ei kuitenkaan palauteta yli 2000 paikkainstanssia löytävien hakujen tuloksia, vaan kehotetaan

66 Apache Axis, http://ws.apache.org/axis/

käyttäjää rajaamaan hakuaan enemmän. Kuvassa 3 on esitelty hakukäyttöliittymän ikkuna, jossa paikkahakuja suoritetaan monipuolisilla haun rajausehdoilla.



Kuva 3. *ONKI-Paikan oletushakukäyttöliittymä, jossa on haettu nimellä Helsinki ja valittu hiiren osoittimella hakutuloslistasta Itä-Suomessa sijaitseva Helsinkilä*

Jokaisella haun rajausvaihtoehdolla on oma osansa hakukäyttöliittymässä. Varmaankin eniten käytetty haun rajaus on paikan nimellä tehtävä rajaus. Tämän takia nimellä haku on tehty mahdollisimman helpoksi pitämällä nimihaun tekstikenttä aina esillä (kuva 4). Nimellä haku on tehty automaattitäytöllä (engl. autocompletion), joka tarkoittaa sitä, että haut suoritetaan samanaikaisesti sitä mukaan, kun tekstikenttään lisätään kirjaimia. Näin haku rajautuu itsestään koko ajan sitä mukaan, mitä pidemmän merkkijonon kirjoittaa. Koska nimihaku toimii MySQL:n *Full-Text*-indeksiä hyödyntäen, löytää haku kaikki ne paikannimet, jotka sisältävät sanan, joka alkaa syötetyllä merkkijonolla. Esimerkiksi hakemalla merkkijonolla ”Simo” löydetään muun muassa ”Simonmäki” mutta myös ”Iso Simolampi” ja ”Yli-Simola”, koska kaikki nämä sisältävät *Full-Text*-indeksin mukaan erillisen sanan, joka alkaa merkkijonolla ”Simo”.

Kuva 4. *Nimihakukenttä, jossa valittu haku säännöllistä lauseketta käyttäen. Tekstikentän alla näkyy hakukäyttöliittymän kieli-, paikkatyyppi-, alue- ja aikahakurajausvalintojen painikkeet.*

Käyttöliittymässä voi myös valita haetaanko hitaammalla, mutta huomattavasti monipuolisemmalla, säännöllistä lauseketta käyttävällä paikannimihaulla. Jos on valittu *regexp*-haku, ei hakua suoriteta automaattitäytöllä, sillä haku voidaan suorittaa vasta kun säännöllinen lauseke on kirjoitettu valmiiksi. Tämän takia, *regexp*-hakuja tehtäessä, ilmestyy syöttö-

kentän viereen hakupainike, jota painamalla haku suoritetaan. Haku käynnistyy myös painamalla tekstikentässä näppäimistön rivinvaihtopainiketta (engl. enter).

kieli	haussa	kaikkiaan
suomi	2804	715799
ruotsi	1	75052
pohjoissaame	2725	4829
inarinsaame	3622	3969
koltansaame	149	149
englanti	0	253

Kuva 5. *Haun rajaus paikannimen kielen mukaan. Kuvassa on valittu kaikki Suomessa puhuttavat saamen kielet.*

Jokaisen haun rajauksen yhteydessä päivittyy löydettyjen paikkojen lista. Paikan nimen mukaan tehtävien rajausten lisäksi on nimihaun tekstikentän alla linkkejä muiden rajausvaihtoehtojen säätöikkunoiden avaamiseen. Ensimmäisenä on rajaus, joka määrittelee sen, millä kielellä nimihakua suoritetaan (kuva 5). Oletuksena haku kohdistuu kaikenkielisiin paikannimiin. Rajaamalla haku koskemaan pelkästään suomenkielisiä paikannimiä, löydetään esimerkiksi merkkijonolla ”Helsing” Taivalkoskella sijaitseva Helsinginaho muttei Mustasaaressa sijaitsevaa Helsingby-nimistä asuinaluetta. Nimen kielen rajaavassa ikkunassa kerrotaan myös tilastoa siitä, kuinka monella paikalla haussa sekä tietokannassa on nimi tietyllä kielellä.

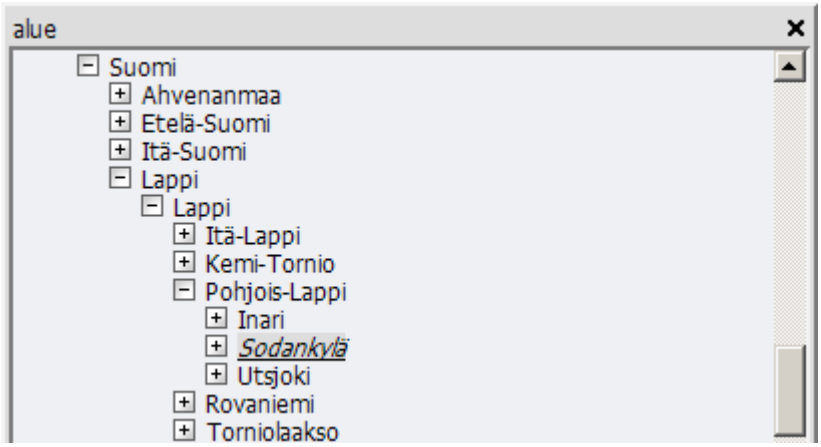
paikkatyyppi	haussa	kaikkiaan
allas	0	12
alue	0	1
asuinalue	2	25062
erämaa-alue	0	3
hautausmaa	0	17
kaatopaikka	0	5
kansallispuisto	0	28

Kuva 6. *Paikan tyyppirajaus, jossa haussa löydetty kaksi asuinaluetta*

Seuraava rajausvaihtoehto on paikan tyypin mukainen haun rajaus (kuva 6). Rajaus toimii aivan samalla tavalla kuin paikan nimen kielen rajaus. Oletuksena kaikki paikkatyypit on valittuna. Myös tässä ikkunassa kerrotaan, kuten kielirajausikkunassa, tilastoa löydettyjen sekä kaikkien paikkojen lukumäärästä tyyppiä kohden. Tyypit ovat suoraan SUO-ontologian mukaisia paikkatyyppejä. Paikkojen tyypit ovat samat, kuin mitä paikoille on RDF-muunnoksen yhteydessä annettu. Tyypin mukaan hakeminen on usein kätevä rajaus esimerkiksi silloin, kun haetaan vaikkapa ainoastaan valtioita tai muita hallinnollisia alueita kuten kuntia.

Seuraavana rajausvaihtoehtona on alue, jonka sisältä paikkoja haetaan (kuva 7). Paikat ovat aina mereologisessa hierarkiasuhteessa johonkin toiseen paikkaan, joka yleensä on jonkinlainen hallinnollinen alue. RDF-tasolla suhde ilmoitetaan paikan *suo:isPartOf*-ominaisuudella. Rajaus ottaa huomioon *suo:isPartOf*-suhteen transitiivisuuden, jonka takia

esimerkiksi Suomella rajattu haku ei palauta ainoastaan läänejä vaan myös järviä ja asuinalueita. Alueita voi rajaukseen valita useita. Oletuksena hauissa ei ole aluerajauksia, jolloin haetaan kaikkia paikkoja.



Kuva 7. *Aluerajausikkuna, jossa valittu aluehierarkiasta Sodankylä, jonka alueelta haetaan paikkoja*

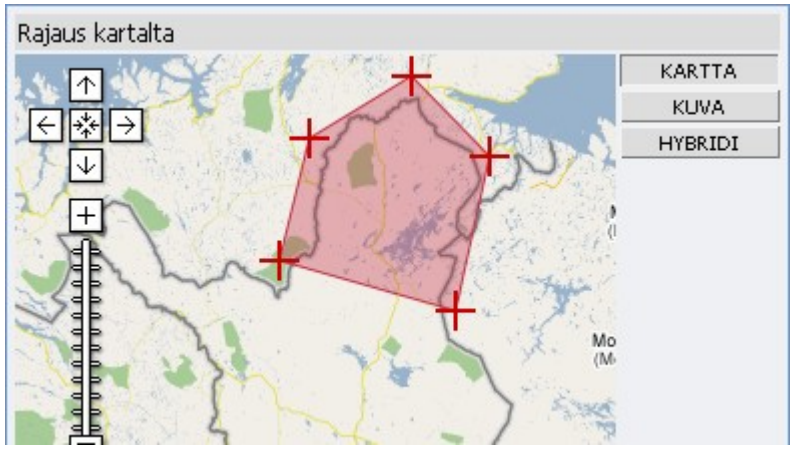
Hakua voidaan rajata myös paikan ajallisen olemassaolon mukaan. Tämä on seuraava rajausvaihtoehto hauille ONKI-Paikan hakukäyttöliittymässä (kuva 8). Paikan olemassaolon alku- ja loppuaika on ONKI-Paikan perusaineistojen osalta peräisin pelkästään SAPO:n aineistosta. Yksikään muista perusaineistoista ei määrittele paikkojen ajallista olemassaoloa. Jos paikalle ei ole määritelty perustamis- tai lakkauttamisaikaa, oletetaan hakujen yhteydessä paikan olleen aina olemassa. Aikarajauksessa määritellään perustamis- ja/tai lakkauttamisvuosi. Haussa palautetaan kaikki ne paikat, jotka ovat olleet olemassa määritellyn aikajakson aikana.

Kuva 8. *Ajan mukaan tehtävän haunrajauksen ikkuna*

Hakukäyttöliittymässä on vielä kuudes rajausmahdollisuus, jossa käytetään hyväksi käyttöliittymän karttanäkymää. Kartta on toteutettu Google Maps -karttapalvelun avulla, joka tarjoaa monipuolisen Javascript / Ajax -rajapinnan kartan esittämiseen ja kartan muokkaamiseen. Kartan kohdistusta on mahdollista siirtää, karttaa voi suurentaa ja pienentää ja kartalle voi lisätä yksinkertaista grafiikkaa, kuten kuvia, viivoja ja monikulmioita. Kartta auttaa huomattavasti esimerkiksi hahmottamaan löydettyjen paikkojen sijaintia. Viemällä hiiren osoittimen löydettyjen paikkojen listassa jonkin löydetyn paikan ylle, näkyy kartalla pisteenä paikan sijainti.

Karttaa voidaan käyttää myös haun rajauksessa määrittelemällä koordinaattipisteistä muodostuva monikulmio, jonka peittämältä alueelta paikkoja haetaan (Kauppinen et al. 2006). Monikulmio määritellään klikkaamalla hiirellä kartalle, jolloin jokaisella klikkauksella kartalle jää risti merkiksi monikulmion janojen päätepisteistä (kuva 9). Jos monikulmion kulmapisteitä haluaa siirtää, voidaan niihin tarttua kiinni hiiren osoittimella ja raahata toiseen kohtaan, jolloin monikulmion muoto muuttuu. Monikulmion saa poistettua painik-

keesta kartan oikealla puolella.



Kuva 9. *Kartalle Pohjois-Lappiin piirretty monikulmio rajausehtona paikkahauille*

Näillä kuudella haun rajausvaihtoehdolla on paikan löytäminen varsin helppoa. Yleensä on aina tiedossa paikan nimi, tyyppi tai jonkinlainen arvio paikan sijainnista. Esimerkiksi jos haetaan Itä-Suomessa olevaa Haapajärvi-nimistä vesistökohdetta voitaisiin klikata kartalle monikulmio, joka suunnilleen määrittäisi Itä-Suomen alueen ja kirjoittaa paikannimi-kenttään Haapajärvi. Jos haku ei tuottanut yhtään tulosta, voidaan monikulmion kulma-pisteitä liikuttaa niin, että monikulmion peittämä alue laajenee. Näin voidaan Itä-Suomea laajentaa, kunnes jokin Haapajärvi löytyy.

7 TULOSTEN ARVIOINTIA

Tämän työn alkuperäisiin tavoitteisiin liittyi semanttisen paikkatietopalvelun suunnittelu ja toteutus. Suunnittelun perustana oli SUO-ontologia, jota populoidaan paikkainstansseilla eri paikkatietolähteistä. Vaikein tehtävä oli määritellä paikkainstansseille globaalit ja uniikit URI-tunnisteet, jotka ovat perustana koko semanttisen verkon instanssien yksilöimisessä. Vaikeaa oli myös määritellä, milloin paikka on erillinen instanssi toisesta paikasta. Jos esimerkiksi kahdessa eri paikkatietoaineistossa on viittaukset Helsingin kaupunkiin, niin voidaanko olettaa että molemmat viittaavat kaikilta ominaisuuksiltaan täysin samaan paikkainstanssi? Tuloksena oli, ettei näin voida olettaa, sillä jokaisella aineistolla on yleensä oma näkökulmansa paikkatietoon, jolloin myös aineistossa esiintyvät paikat saattavat sisältää tiedoiltaan eriäviä arvoja.

Olisi myös lähes mahdotonta koneellisesti päättää, viittaako eri aineistoissa esiintyvät paikat samoihin paikkainstansseihin. Kuten työssä selvitettiin, on paikkojen disambiguoiminen nimen, sijainnin tai paikkatyypin mukaan lähes mahdotonta tehdä täydellä varmuudella. Näiden syiden takia otettiin tässä työssä aineistolähtöinen lähestymistapa paikkainstanssien tunnistamiseen. Jokaisen aineiston sisältämät paikat ovat siis uniikkeja paikkainstansseja, omine URI-tunnisteineen. Tämä helpottaa huomattavasti semanttisen paikkatietopalvelimen ylläpitoa ja kehitystä, sillä jokaista aineistoa voidaan käsitellä täysin toisista aineistoista riippumattomina lähteinä. Paikkainstanssien URI-tunnisteiden luomiseen käytetään, jos mahdollista, hyväksi paikkatietoaineiston sisäisiä tunnisteita, jolloin aineiston päivittäminen paikkatietopalveluun on helpompaa.

Koko palvelun perustana on hakemiston luominen paikkojen virallisille tunnisteille, eli semanttisen verkon URI-osoitteille, joilla globaalisti voidaan viitata koko verkon laajuudessa tiettyyn paikkainstanssiin. Tämä on etenkin tärkeää eri resurssien annotoinnin yhteydessä, jolloin hyödytään siitä, että eri resursseissa käytetään samaa tunnistetta tietylle paikalle. Tällä tavalla eri resurssit saadaan verkottumaan toisiinsa yhteisten paikkaviittausten avulla.

7.1 Tavoitteiden saavuttaminen

Alkuperäisten tavoitteiden saavuttaminen on työssä onnistunut varsin hyvin. Eri paikkatietoaineistojen paikkainstanssien URI-tunnisteiden luomiselle on laadittu tarkat säännöt, joita noudattamalla uusia paikkatietoaineistoja voidaan myöhemmin lisätä palveluun. Myös paikkatiedon muiden instanssien, kuten koordinaattipisteiden URI-tunnisteiden muoto on tarkasti tässä työssä määritelty. Näiden määritysten avulla on varsin suoraviivaista ylläpitää palvelun sisältämiä paikkainstansseja.

ONKI-Paikka-palvelun tehtävä ei ole olla yleinen tietovarasto erilaiselle paikkatiedolle, vaan paikkainstanssien diambiguoimiseen räätälöity työkalu. Tämän takia hakupalveluun tallennetaan paikka-aineistojen indeksoinnin yhteydessä pelkästään tätä tehtävää varten tarvittavia tietoja paikoista. Tällä pyritään myös selkeään ja käytöltään mahdollisimman yksinkertaiseen palveluun, jonka käyttöönottokynnys olisi mahdollisimman alhainen. Selkeän käyttöliittymän ja tehokkaan moninäkymä- tai yhdistelmähaun ansiosta palvelua

on helppo käyttää esimerkiksi selaimen kautta. ONKI-Paikan oletushakukäyttöliittymä tarjoaa käyttäjälleen visuaalisen tavan löytää hakemiaan paikkoja. Apuna on Googlen kehittämä, selaimessa toimiva karttapalvelu, jonka ansiosta hauissa löydettyjen paikkojen sijainti on helposti nähtävissä. Hakukäyttöliittymän suhteen työn tavoitteet on saavutettu hyvin. Haut ovat suhteellisen nopeita ja hakutuloksen visuaalisuuden ja informatiivisuuden ansiosta etsitty paikka myös löytyy varsin nopeasti.

7.2 Jatkokehitys

Tässä ratkaisussa tiedon kulku on hakukäyttöliittymän kautta yksisuuntainen. Käyttäjät eivät siis pääse tallentamaan omia paikkojaan palveluun, niin että muut käyttäjät voisivat hyödyntää niitä omissa annotoinneissaan. Tämä olisi yksi idea jatkokehitykselle, jossa palveluun luotaisiin aineisto käyttäjien omille paikoille. Voidaan hyvin kuvitella, että olisi tarvetta tietyille yhteisöille luoda omia paikkatietoaineistoja pelkästään heidän tarpeita tyydyttämään.

Esimerkiksi jokin arkeologinen seura tai arkeologiaa harjoittava ja kaivauksia suorittava taho voisi tarvita kaivauspaikoista oman luettelonsa. Kaivauspaikat toki sijaitsevat jossain, jo ennestään nimetyllä alueella, mutta tämä alue ei tarpeeksi tarkasti yksilöi aluetta, jossa varsinainen kaivaus tehdään. Tällöin olisi kätevä jos arkeologit voisivat ylläpitää omaa aineistoa, joka kattaa pelkästään kaivauksia. Näin myös muut arkeologiset seurat ja yhteisöt voisivat aineistoissaan viitata muiden kaivauksiin yhteisten URI-tunnisteiden avulla.

Se, että käyttäjille annetaan mahdollisuus paikkojen lisäämiseen tuo mukanaan myös tarpeen käyttäjien ja käyttöoikeuksien hallinnalle. Tämä toisi koko palveluun yhden ylimääräisen kompleksisuustason, jolle ei ONKI-Paikan alkuvaiheissa ole nähty riittävää tarvetta. Käytännössä omien paikkojen hallintaan pitäisi rakentaa oma sovellus, jonne kirjaudutaan sisään. Sovelluksessa käyttäjä voi muokata niiden aineistojen paikka-instansseja, joihin hänelle on annettu siihen tehtävään oikeudet. Myös hakukäyttö-liittymässä tulisi ottaa huomioon käyttäjien omat paikkatietoaineistot. Pitää määritellä se, olisivatko aineistot kaikkien käyttäjien vai pelkästään tiettyjen käyttäjien käytettävissä. Tämä tarkoitta sitä, että myös hakukäyttöliittymässä vaadittaisiin sisäänkirjautuminen, jos halutaan tehdä hakuja myös ei-julkisiin paikkatietoaineistoihin.

Toinen vaihtoehto olisi perustaa omia instansseja ONKI-Paikka-plavelimesta, jossa ylläpidettäisiin omia, ei-julkisia paikkatietoaineistoja. Tällöin on huolehdittava siitä, että paikkainstanssien URI-tunnisteiden nimeämisessä ei tapahdu konflikteja ja päällek-käisyyksiä muiden ONKI-Paikka-palvelinten aineistojen kanssa. Tässä työssä on määritelty pelkästään tämän työn tuloksena syntyneen ONKI-Paikka-palvelimen instanssin paikkojen URI-tunnisteiden nimeämiskäytäntö. Mikään ei estä sitä, että jokin arkeologinen yhteisö perustaisi oman paikkatietoaineistonsa, jonka sisältämien paikkojen instanssien nimi-avaruus voisi olla esimerkiksi http://www.arkeologit.fi/onto/. Myös siinä nimiavaruudessa olevat paikkainstanssit olisivat SUO-ontologian paikkainstansseja. Tämä vain ei olisi suoraan havaittavissa instanssien URI-tunnisteista.

Tässä toteutuksessa ainoa semanttinen suhde paikkojen välillä, joka tallennetaan hakuja

varten, on *suo:isPartOf*. Se antaa mahdollisuuden hakea paikkoja tietyn hallinnollisen alueen sisältä. Spatiaalisia suhteita on kuitenkin useita, joiden avulla olisi mahdollista suorittaa entistä tehokkaampia ja täsmällisempiä hakuja. Käyttökelpoinen olisi esimerkiksi ontologinen suhde, joka määrittelisi hallinnollisten alueiden naapuruussuhteet. Voisi olla kätevää hakea esimerkiksi jonkin tietyn kunnan kaikki naapurikunnat. Tällä hetkellä SUO-ontologia ei määrittele tällaista suhdetta, joka voisi olla nimeltään esimerkiksi *isNeighborOf*. SUO-ontologiassa paikkojen naapurit määritellään paikkojen muodon ja sijainnin määrittelevien geometristen primitiivien, kuten monikulmioiden ja janojen suhteiden kautta. Monikulmioiden välille voidaan määritellä *suo:touches*-suhde, joka tarkoittaa sitä, että kaksi aluetta ovat toistensa naapureita. Tässä toteutuksessa tällainen haku ei ole mahdollista siitä syystä, että lähdeaineistoissa tällaista naapuruussuhdetta eikä myöskään alueiden geometrisiä primitiivejä ole määritelty. Pelkistä paikkojen keskipisteen määrittelevistä koordinaattipisteistä ei voida päätellä, onko jokin kunta toisen kunnan naapuri.

ONKI-Paikka palvelin tarjoaa kuitenkin mahdollisuuden lisätä rajattomasti suhteita eri paikkainstanssien välille. Riippuu täysin lähdeaineistosta, tarjoaako se tietoa paikkojen välisistä suhteista. Jää siis lähdeaineiston RDF-muunnoksen tekijän tehtäväksi huolehtia paikkainstanssien välisten semanttisten suhteiden lisäämisestä paikkojen RDF-kuvauksiin. Web Service -rajapinnan *getRelatedPlaces*-metodin avulla on mahdollista hakea minkä tahansa suhteen mukaan.

Käytännössä jokaisen uuden suhteen lisääminen palveluun tarkoittaisi uuden haun rajauksen näkymän lisäämistä hakukäyttöliittymään. Jos palveluun lisätään *isNeighborOf*-suhde, tarvitaan myös käyttöliittymässä näkymä, jossa valitaan paikka tai paikat, jonka naapureita haetaan. Tämä olisi kuitenkin käyttöliittymän rakentajalle täysin mahdollista Web Service -rajapinnan metodeja käyttämällä.

Erilaisten semanttisten suhteiden myötä kasvaa myös tarve päättelyille. Tällä hetkellä päättelyä tehdään automaattisesti ainoastaan *suo:isPartOf*-suhteen mukaan tehtävissä hauissa. Jos esimerkiksi haetaan paikkoja Suomen sisältä ei voida palauttaa pelkästään läänejä, jotka ovat paikkahierarkiassa suoraan Suomen alla, vaan kaikkia paikkoja Suomen rajojen sisältä. Jotta tämä olisi mahdollista on tehtävä päättelyä, jossa huomioidaan *suo:isPartOf*-suhteen transitiivisuus. Käytännössä tämä tapahtuu RDF-aineiston indeksoinnin yhteydessä, jossa tietokantaan luodaan valmiiksi tieto paikkojen koko *suo:isPartOf*-hierarkiasta. Näin päättelyä ei tarvitse suorittaa hakua tehtäessä.

Muita päättelyjä, joita ei tässä toteutuksessa tueta, on eri suhteiden negaatiot. Voitaisiin esimerkiksi hakea paikkoja, jotka eivät ole *suo:isPartOf*-suhteessa toisiinsa tai paikkoja, jotka eivät ole toistensa välittömiä naapureita. Myös tämänlaiset haut vaatisivat omat lisäyksensä hakukäyttöliittymään. Monilla eri hakurajausten yhdistelmillä on mahdollista saada selville yksityiskohtaisempaa tietoa paikoista. Esimerkiksi naapurikuntien haku saattaa olla hyvinkin tarpeellinen toiminto tietyissä sovelluksissa, mutta paikkainstanssien disambiguoimiseen tämänlaisesta hausta on tuskin hyötyä. Vaikka kunta on Suomessa hallinnollisten alueiden alin taso, ei kuntien disambiguoimisessa ole yleensä ongelmia. Jos tietää haetun kunnan naapurikunnan nimen, on helppo ensin etsiä tiedetty kunta kartalta, ja

tämän jälkeen rajata hakua koskemaan pelkästään *suo:kunta*-tyyppisiä paikkoja ja rajata kartalta monikulmiolla alue tunnetun kunnan ympäriltä. Näin löydetään nopeasti jonkin kunnan naapurikunnat tai sen lähellä olevat kunnat.

Iso ratkaisematon kysymys on myös eri aineistoista peräisin olevien täysin samojen paikkainstanssien välisten *owl:sameAs*-suhteiden automaattinen luominen. Kyseessä olisi disambiguointitehtävä, jossa aineistoista saatujen paikkojen tietojen perusteella pitäisi voida koneellisesti päätellä viittaavatko kaksi paikkaa samaan maantieteelliseen kohteeseen. Kahdessa paikkatietoaineistossa on harvemmin täysin kiistattomasti viittaus samaan maantieteelliseen kohteeseen. Esimerkiksi Maanmittauslaitoksen Paikannimirekisterissä Helsinki on luokiteltu kaupunkimaiseksi kunnaksi osana Suomen hallinnollisten alueiden hierarkiaa. GEOnet Names Server -aineistossa Helsinki on luokiteltu poliittisen toimijan, eli valtion pääkaupungiksi. Nopeasti pääteltynä on aivan selvää että molemmat viittaavat Helsinki nimiseen kaupunkiin, joka siis myös on Suomen pääkaupunki. Toisaalta näissä kahdessa eri aineistoissa luokittelu antaa oman semanttisen vivahteensa paikkainstanssille. Määriteltäväksi jää se, mikä on se varsinainen maantieteellinen kohde, johon molemmat aineistot viittaavat paikan nimellä Helsinki.

Paikannimirekisterin Helsinki on kunta, jonka ominaisuuksina on kunnan rajat. Kuten luvussa 5.4.4 esitettiin, on kunnan instanssin määrittämisessä otettava huomioon kunnan maantieteellinen kattavuus eli kunnan rajat. Jos rajat muuttuvat, syntyy samalla uusi kuntainstanssi ja vanha lakkaa olemasta. Jos taas Helsinkiin viitataan valtion pääkaupunkina, ei tähän instanssiin selvästikään liity kuntaan liittyviä tietoja kuten kunnan rajoja. Pääkaupunki Helsinki pysyy samana pääkaupunkina ja samana paikkainstanssina vaikka kunta, jossa pääkaupunki sijaitsee muuttaa rajojaan. Tästä voidaan päätellä, että olisi tarve jonkinlaiselle asutetun paikan alkeistyypille nimeltä Helsinki, joka olisin jonkin suhteen kautta yhteydessä kaikkiin eri Helsinki-instansseihin. Suhde ei kuitenkaan voisi olla symmetrinen *owl:sameAs*-suhde, sillä silloin päättelyn kautta kaikki alkeistyyppiin liitoksissa olevat paikkainstanssit olisivat myös keskenään *owl:sameAs*-suhteessa, joka ei tässä tapauksessa ole totta.

Kuten tästä pohdiskelusta voi todeta ei paikkainstanssin määrittäminen aina ole helppo tehtävä. Paikkatiedon määrittämisessä on lukuisia eri näkökulmia, jotka vaikuttavat siihen, mitä tietyllä yksittäisellä paikalla tarkoitetaan.

LÄHDELUETTELO

Abadi, D., Marcus, A., Madden, S., Hollenbach, K. (2007) *Scalable Semantic Web Data Management Using Vertical Partitioning*. VLDB 2007 - 33rd International Conference on Very Large Data Bases, Wien, Itävalta, 23.-27.9.2007.

Antoniou, G., van Harmelen, F. (2004) *A Semantic Web Primer*. Cambridge, Massachusetts, The MIT Press. 238 s.

Berners-Lee, T. (1998) *Cool URIs don't Change* (online). Päivitetty 1998 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/Provider/Style/URI>

Berners-Lee, T., Fielding, R., Masinter, L. (2005) *Uniform Resource Identifiers (URI): Generic Syntax* (online). Päivitetty tammikuussa 2005 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://tools.ietf.org/html/rfc3986>

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F. (2006) *Extensible Markup Language (XML) 1.0 (Fourth Edition)* (online). Päivitetty 29.9.2006 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/xml/>

Brickley, D., Guha R.V. (2004) *RDF Vocabulary Description Language 1.0: RDF Schema* (online). Päivitetty 10.2.2004 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/rdf-schema/>

Christensen, E., Curbera, F., Meredith, G., Weerawarana, S. (2001) *Web Services Description Language (WSDL) 1.1* (online). Päivitetty 15.3.2001 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/wsdl>

Dean, M., Schreiber, G. (2004) *OWL Web Ontology Language Reference* (online). Päivitetty 10.2.2004 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/owl-ref/>

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T. (1999) *Hypertext Transfer Protocol - HTTP/1.1* (online). Päivitetty kesäkuussa 1999 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://tools.ietf.org/html/rfc2616>

FIPS 5-2 (1987) *Codes for the Identification of the States, the District of Columbia and the Outlying Areas of the United States, and Associated Areas*. National Institute of Standards and Technology, Information Technology Laboratory.

FIPS 6-4 (1990) *Counties and Equivalent Entities of the United States, Its Possessions, and Associated Areas*. National Institute of Standards and Technology, Information Technology Laboratory.

FIPS 10-4 (1995) *Countries, Dependencies, Areas of Special Sovereignty, and Their Principal Administrative Divisions*. National Institute of Standards and Technology, Information Technology Laboratory.

Gudgin, M., Hadley, M., Mendelsohn, M., Moreau, J., Nielsen H.F., Karmarkar, A., Lafon, Y. (2007) *SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)* (online). Päivitetty 27.4.2007 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/soap12-part1/>

Hamilton, G. (1997) *Sun Microsystems JavaBeans™*, Version 1.01-A. Mountain View, CA, Sun Microsystems. 114 s.

Henriksson, R., Kauppinen, T., Hyvönen, E. (2008) *Core Geographical Concepts: Case Finnish Geo-Ontology*. Location and the Web (LocWeb) 2008 workshop, 17th International World Wide Web Conference WWW 2008, ACM International Conference Proceeding Series; Vol. 300, Pages 57-60, Peking, Kiina, 21.-25.4.2008

Holi, M., Hyvönen, E. (2006) Modeling Uncertainty in Semantic Web Taxonomies. Teoksessa: Zhongmin Ma (toim.). *Soft Computing in Ontologies and Semantic Web.* Springer-Verlag.

Hyvönen, E. (2001) *Semantic Web - kohti uutta merkitysten Internetiä*. Esitelmä Semantic Web Kick-Off in Finland -tilaisuudessa, 2.11.2001. Helsinki, Porthania, Helsingin yliopisto.

Hyvönen, E., Viljanen, K., Mäkelä, E., Kauppinen, T., Ruotsalo, T., Valkeapää, O., Seppälä, K., Suominen, O., Alm, O., Lindroos, R., Känsälä, T., Henriksson, R., Frosterus, M., Tuominen, J., Sinkkilä, R., Kurki, J. (2007) *Elements of a National Semantic Web Infrastructure - Case Study Finland on the Semantic Web* (Invited paper). Proceedings of the First International Semantic Computing Conference (IEEE ICSC 2007), Irvine, Kalifornia, Syyskuussa 2007. IEEE Press.

Ishida, R. (2006) *Language Tags in HTML and XML* (online). Päivitetty 9.11.2006 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/International/articles/language-tags/>

ISO 19136 (2007) *Geographic Information - Geography Markup Language (GML)*. International Organization for Standardization. 349 s.

ISO 3166-1 (2006) *Codes for the Representation of Names of Countries and their Subdivisions - Part 1: Country Codes*. International Organization for Standardization. 69 s.

ISO 639-1 (2002) *Codes for the Representation of Names of Languages - Part 1: Alpha-2 Code*. International Organization for Standardization. 37 s.

ISO 639-2 (1998) *Codes for the Representation of Names of Languages -- Part 2: Alpha-3 Code*. International Organization for Standardization. 66 s.

ISO 639-3 (2007) *Codes for the Representation of Names of Languages -- Part 3: Alpha-3 Code for Comprehensive Coverage of Languages*. International Organization for Standardization. 12 s.

ISO/IEC 8859-10 (1998) *Information Technology - 8-bit Single-byte Coded Graphic Character Sets - Part 10: Latin Alphabet No. 6*. International Organization for Standardization / International Electrotechnical Commission. 10 s.

JHS 110 (n.d.) *Kuntien numerotunnus*. Julkisen hallinnon tietohallinnon neuvottelukunta. 2 s.

JHS 153 (2006) *ETRS89-järjestelmän mukaiset koordinaatit Suomessa*. Julkisen hallinnon tietohallinnon neuvottelukunta. 22 s.

JHS 154 (2006) *ERTS89 -järjestelmään liittyvät karttaprojektiot, tasokoordinaatistot ja karttalehtijako*. Julkisen hallinnon tietohallinnon neuvottelukunta. 33 s.

JHS 162 (2007) *Paikkatietojen mallintaminen tiedonsiirtoa varten*. Julkisen hallinnon tietohallinnon neuvottelukunta. 9 s.

Kauppinen, T., Henriksson, R., Väätäinen, J., Deichstetter, C., Hyvönen, E. (2006) *Ontology-based Modeling and Visualization of Cultural Spatio-temporal Knowledge*. Developments in Artificial Intelligence and the Semantic Web - Proceedings of the 12th Finnish AI Conference STeP, 26.-27.10.2006.

Kauppinen, T., Väätäinen, J., Hyvönen, E. (2008) *Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal*. S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 ESWC 2008, LNCS 5021, Teneriffa, Espanja, 1.-5.6.2008. s. 110-123

Kavouras, M., Kokla, M., Tomai, E. (2005) Comparing Categories Among Geographic Ontologies. *Computers & Geosciences* 31, 2. s. 145-154.

KOTUS, Kotimaisten kielten tutkimuskeskus (2006) *Suomen paikannimet vieraskielisissä teksteissä* (online). Päivitetty 5.12.2006 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.kotus.fi/index.phtml?s=599>

KOTUS, Kotimaisten kielten tutkimuskeskus (2007) *Saamen kielet* (online). Päivitetty 22.12.2007 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.kotus.fi/index.phtml?s=207>

Manola, F., Miller, E. (2004) *RDF Primer* (online) Päivitetty 10.2.2004 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/rdf-primer/>

Miles, A., Bechhofer, S. (2008) *SKOS Simple Knowledge Organization System Reference* (online). Päivitetty 25.1.2008 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/skos-reference/>

Moats, R. (1997) *URN Syntax* (online). Päivitetty toukokuussa 1997 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://tools.ietf.org/html/rfc2141>

NIMA TR8350.2 (2000) *Department of Defense World Geodetic System 1984, Its Definition and Relationships With Local Geodetic Systems*, Third Edition. National Imagery and Mapping Agency. 175 s.

OGC 05-047r3 (2006) *GML in JPEG 2000 for Geographic Imagery (GMLJP2) Encoding Specification*. Open Geospatial Consortium Inc. 170 s.

Phillips, A., Davis, M. (2006) *Tags for Identifying Languages* (online). Päivitetty syyskuussa 2006 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.rfc-editor.org/rfc/rfc4646.txt>

Sanastokseskus TSK ry (2005). *Geoinformatiikan sanasto*. Gummerus Kirjapaino Oy, Saarijärvi. 54 s.

Sauermann, L., Cyganiak, R., Ayers, D., Völkel, M. (2008) *Cool URIs for the Semantic Web* (online). Päivitetty 31.3.2008 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.w3.org/TR/cooluris/>

Saur, K.G. (2007) *International Standard Bibliographic Description (ISBD)*. München, International Federation of Library Associations and Institutions. 320 s.

Sinkkilä, R. (2008) *Käsitteen kontekstiperustainen valinta semanttisessa webissä*. Pro gradu -tutkielma. Helsingin yliopisto, Tietojenkäsittelytieteen laitos. Helsinki. 68 s.

Sisäasiainministeriö (2001) *Aluekehitysalan sanasto suomi-ruotsi-englanti-saksa-ranska* (online). Päivitetty 2001 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.intermin.fi/intermin/images.nsf/files/642951652C173123C2256CD90048A507/$file/Aluekehityssanasto.pdf>

Tilastokeskus (2007) *Valtiot ja maat maanosittain 2007* (online). Päivitetty 21.9.2007 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.stat.fi/tk/tt/luokitukset/lk/valtio_21_index.html>

Viljanen, K., Tuominen, J., Hyvönen, E. (2008) *Publishing and Using Ontologies as Mash-Up Services*. Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008), Teneriffa, Espanja, 1.-5.6.2008.

Väätäinen, J. (2008) *Ajallisesti muuttuvan paikkatiedon hallinta*. Insinöörityö. EVTEK-ammattikorkeakoulu, Mediatekniikan koulutusohjelma. Espoo. 69 s.

YK, Yhdistyneet kansakunnat (2006) *List of Member States* (online). Päivitetty 3.10.2006 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.un.org/members/list.shtml>

YK, Yhdistyneet kansakunnat (2007) *Countries or Areas, Codes and Abbreviations* (online). Päivitetty 28.8.2007 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://unstats.un.org/unsd/methods/m49/m49alpha.htm>

YK, Yhdistyneet kansakunnat (2008) *Composition of Macro Geographical (Continental) Regions, Geographical Sub-regions, and Selected Economic and other Groupings* (online). Päivitetty 31.1.2008 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://unstats.un.org/unsd/methods/m49/m49regin.htm>

YLE, Yleisradio (2008) *Kiviniemi: Läänit lakkautetaan 2010* (online). Päivitetty 25.3.2008 [viitattu 30.5.2008]. Saatavilla WWW-muodossa: <URL:http://www.yle.fi/uutiset/24h/id86191.html>