FinnONTO-hanke loi ontologisen perustan kansalliselle webin tietoinfrastruktuurille

Eero Hyvönen

Aalto-yliopiston ja Helsingin yliopiston vetämässä FinnONTO-hankkeiden sarjassa (2003–12) kehitettiin ja otettiin käyttöön ensimmäinen kansallisen tason prototyyppi semanttisen webin tietoinfrastruktuurista. Siihen kuuluu eri alojen ontologioita, ONKI-ontologiakirjastopalvelut näiden hyödyntämiseksi sovelluksissa sekä malli prosessista, jolla monialainen ontologityö voitaisiin Suomessa jatkossa järjestää. Vuodesta 2008 koekäytössä olleesta ONKI-palvelusta vakinaistettiin vuoden 2014 alussa Kansalliskirjaston toimesta ontologiapalvelu Finto, jota rahoittavat opetus- ja kulttuuriministeriö sekä valtiovarainministeriö ja jonka piirissä ontologiatyö on pilottivaiheen jälkeen jatkumassa.

World Wide Webin (WWW) sisään on nopeasti rakentumassa uudenlainen semanttinen tiedon verkko, Web of Data, josta WWW:n ”isä” Tim Berners-Lee käyttää myös nimityksiä *Linked Data* (Heath, Bizer, 2011) ja *Giant Global Graph* (GGG). GGG on verkkomuotoista (meta)tietoa maailmasta, esimerkiksi että Galapagos-saarilla tavattava sinijalkasuula on lintulaji tai että Akseli Gallen-Kallela maalasi Aino-triptyykin 1. version Pariisissa vuonna 1889. Tiedon webin perustan muodostavat ontologiat, jotka määrittelevät tiedon kuvailussa käytettävät sanastot tietokoneiden ”ymmärtämällä” tavalla. Google käyttää omasta GGG-verkostaan nimeä *Knowledge Graph* ja Microsoft buddhalaiseen valaistumiseen viittaavaa sanaa *Satori*.

GGG rakentuu pala palata toisiinsa yhdistettävistä datajoukoista, semanttisen verkon palasista, joita eri tahot tuottavat. Esimerkiksi kulttuurialalla datajoukkoja ovat mm. eri museoiden, kirjastojen ja arkistojen kokoelmatiedot (Hyvönen, 2012). WWW:n mittakaavassa keskeinen datajoukko on erikielisistä Wikipedioista algoritmisesti louhittu DBpedia-verkko[1], johon on n. kymmenen vuoden kuluessa linkitetty muita tietojoukkoja Linked Open Data -pilveksi[2], kuten yli 8 miljoonaa paikkaa sisältävä GeoNames[3]. Linkitetyssä datapilvessä lasketaan tätä kirjoitettaessa olevan 2 122 tietojoukkoa, yhteensä 62 miljardia tietojen välistä yhteyttä[4], kuten että Sibelius on säveltäjä. Datajoukkojen linkitys tarkoittaa sitä, että esimerkiksi DBpedian verkossa oleva Helsingin käsite (tunniste) on yhdistetty samuutta kuvaavalla semanttisella linkillä GeoNames-verkossa olevaan Helsingin käsitteeseen, jolloin tietokoneet voivat yhdistää ja rikastaa molemmista lähteistä saatavaa tietoa. Käsitteiden tunnistamiseen käytetään verkko-osoitteita (HTTP URI/IRI -tunnisteita), joiden kautta, esimerkiksi selaimen avulla, on mahdollista saada lisätietoa käsitteistä.

GGG:n dataverkko on esitetty ”semanttisesti” eli tietokoneen ”ymmärtämällä” tavalla, jossa tieto esitetään käsitteisiin liitettävien ominaisuuksien (*property*) ja näiden arvojen avulla (Antoniu, van Harmelen, 2012). Esimerkiksi maalauksen aihe-ominaisuuden (*subject*) arvona voi olla viittaus jonkin datajoukon, esimerkiksi DBpedian tai GeoNamesin Helsinki-käsitteeseen. Jos ominaisuuksien arvot eri tietojoukoissa valitaan yhteisesti sovituista ontologiasta (käsitteistöistä), tiedot yhdistyvät automaattisesti toisiinsa niiden kautta, ja datan linkittäminen helpottuu ratkaisevasti. Tietotekniikassa ”ontologialla” (Staab, Studer, 2009) tarkoitetaan tällaisia yhtei-

1 http://dbpedia.org/
2 http://linkeddata.org/
3 http://geonames.org/
4 http://stats.lod2.eu/

sesti määriteltyjä tietokoneen tulkittavissa olevia käsitteistöjä, esimerkiksi tesaurusta tai paikannimirekisteriä. Filosofiassa ontologia on tutkimushaara, jossa pohditaan olemassaolon perimmäisiä kysymyksiä.

Ontologiat esitetään ja niitä käytetään WWW:n maailmanlaajuista infrastruktuurityötä koordinoivan W3C-järjestön standardien avulla, joista tärkeimpiä ovat Resource Description Framework RDF -verkkotietomalli, ontologioiden ja sanastojen esittämiseen tarkoitetut RDF Schema[5], Simple Knowledge Organization System SKOS[6] ja Web Ontology Language OWL[7] sekä SPARQL-kyselykieli[8]. Nämä standardit määrittelevät yleisiä, yhteentoimivuuden ja loogisen päättelyn kannalta keskeisiä semanttisia periaatteita. Esimerkiksi RDF Schemassa määritellyllä subClassOf-suhteella voidaan ilmaista käsitettä yleisempi käsite, esimerkiksi että leijonat ovat kissaeläimiä. Jos tiedetään Leon olevan leijona, voidaan päätellä sen olevan myös kissaeläin.

Ontologiat kansallisena haasteena

Suomessa semanttisen webin ontologioita alettiin kehittää laajamittaisemmin MuseoSuomi-järjestelmän[9] yhteydessä vuoden 2002 alussa (Hyvönen ym., 2005). Työssä linkitettiin eri museoiden kokoelmatietoja ja tässä tarvittiin yhteisiä käsitteistöjä mm. esinetyypeille, materiaaleille, henkilöille ja paikoille. Käytettävissä ei ollut kotimaisia ontologioita ja kansainvälisiä järjestelmiä ei voitu sellaisinaan hyödyntää suomen kielestä ja kansallisista sisällöistä ja käytännöistä johtuen. Tästä kokemuksesta syntyi visio siitä, että maamme tarvitsee avoimeen dataan perustuvan semanttisen webin ontologisen sisältöinfrastruktuurin, jonka varaan verkkopalvelut voitaisiin rakentaa kustannustehokkaasti.

Vision toteuttamiseksi käynnistyi vuonna 2003 FinnONTO-projektisarja[10], joka oli tiettävästi ensimmäinen kansallisen tason semanttisen webin infrastruktuurihanke maailmassa. Sen yhtenä ydintavoitteena oli eri alojen ydinsanastojen ontologisoiminen ja linkittäminen eri alat ylittäväksi harmoniseksi kokonaisuudeksi, joka olisi kaikkien käytettävissä avoimena datana (*Open Data*) erityisten ontologiakirjastopalveluiden kautta. Tätä pääosin Tekesin rahoittamaa hanketta toteuttamaan koottiin aluksi 14 yrityksen ja julkisen organisaation rahoittajakonsortio, joka kasvoi suurimmillaan 39 organisaatioon. FinnONTO-hanketta johti Aalto-yliopiston ja Helsingin yliopiston Semanttisen laskennan tutkimusryhmä SeCo[11], jonka suojissa myös pääosa tutkimustyöstä on tehty.

W3C:n standardit eivät ota kantaa ontologioiden tai metadatan varsinaiseen sisältöön, vaan ovat luonteeltaan sovellusriippumattomia ja perustuvat logiikkaan. Esimerkiksi subClassOf-suhde mahdollistaa leijonan ja kissaeläimen välisen yleisen yläluokkasuhteen ilmaisemisen, mutta varsinaiseen eläinlajien taksonomian muodostamiseen ei oteta kantaa. Sisältö- ja ontologiatyö jää kunkin alan asiantuntijatahojen ja sovellusten kehittäjien tehtäväksi. FinnONTO-hanke tarttui tähän haasteeseen tavoitteenaan kansallisen, monialaisen ontologiainfrastruktuurin kehittäminen maahamme ja teknologian pilotointi hyötysovelluksissa.

Ontologiainfrastruktuurin osat

Ajatusta kansallisesta semanttisen webin ontologiainfrastruktuurista (Hyvönen ym., 2008) voi verrat tie-, sähkö- ja puhelinverkkojen muodostamiin perinteisiin infrastruktuureihin. Semanttisten käsitteiden verkosto on luonnollisesti abstrakti ja näkymätön, mutta mahdollistaa hieman vastaavaan tapaan yhteyksiä kuin vaikkapa tie- tai puhelinverkko. Visiona on, että jos FinnONTO-infrastruktuurin pelisäännöillä tuotettu sisältö julkaistaan verkossa, voidaan se kytkeä ja hyödyntää automaattisesti muiden toimijoiden semanttisissa verkoissa hieman vastaavaan tapaan, kuin uusi maantienpätkä tieverkoston osana. Etuna on, että uuden sisäl-

5 http://www.w3.org/standards/techs/rdf#w3c_all
6 http://www.w3.org/standards/techs/skos#w3c_all
7 http://www.w3.org/standards/techs/owl#w3c_all
8 http://www.w3.org/standards/techs/sparql#w3c_all
9 http://www.museosuomi.fi/
10 http://www.seco.tkk.fi/projects/finnonto/
11 http://www.seco.tkk.fi/

lön arvo rikastuu ”ilmaiseksi” infrastruktuurin avulla verkon muusta sisällöstä (vastaavasti kuin uuden tieosuuden arvo syntyy sen kytkeytymisestä muihin teihin), ja toisaalta muiden verkon julkaisijoiden sisältöjen arvo rikastuu uuden tiedonsirpaleen avulla (vastaavasti kuin uusi tie parantaa muiden teiden keskinäistä saavutettavuutta). Sekä itse infrastruktuuri että siinä jo oleva tieto voidaan hyödyntää toisissa sovelluksissa, mikä säästää merkittävästi järjestelmien kehityskustannuksia.

Ontologiainfrastruktuurissa voidaan erottaa kolme komponenttia:

- Ontologiat. Joukko toisiinsa linkitettyjä, W3C:n standardeihin ja kansalliseen käsitteistöön perustuvia ontologioita, joilla on linkkejä myös kansainvälisiin sanastoihin ja tietosisältöihin.
- Ontologiapalvelut. Ontologiakirjastopalveluiden ONKI12 kautta ajantasaiset ontologiat voidaan ottaa kustannustehokkaasti käyttöön eri organisaatioissa verkkopalveluina (Tuominen ym., 2009; Viljanen ym., 2009; Suominen ym., 2012).
- Ontologiajärjestelmän ylläpitoprosessi. Systemaattinen ja koordinoitu kansallinen organisaatio ja mekanismi, joka kantaa vastuun eri aloilla tarvittavien sanastojen yhteisöllisestä kehittämisestä.

Tarvittavia ontologioita voidaan ryhmitellä seuraavasti:

Yleiskäsiteontologiat. Nämä ontologiat vastaavat käsitteistöltään karkeasti nykyisiä asiasanastoja ja tesauruksia (ilman vapaan indeksoinnin termejä), kuten YSA[13], MASA tai Getty-säätiön Art and Architecture Thesauruksessa[14], jossa on yli 51 000 käsitettä ja 269 000 termiä.

Toimijaontologiat. Toimijaontologiat ovat henkilö ja organisaatiorekistereitä, ja muistuttavat kirjastoissa käytettyjä niin sanottuja auktoriteettitietokantoja. Toimijaontologian avulla samannimiset toimijat voidaan yksilöidä ja erottaa toisistaan eri tunnisteilla ja lisätietojen avulla, esimerkiksi henkilökaimat syntymävuoden ja paikan avulla. Esimerkiksi Getty-säätiön Union List of Artist Names (ULAN) -sanastossa[15] on noin 120 000 toimijaa (kuten Akseli Gallen-Kallela), joilla on noin 293 000 erilaista nimeä, ja toimijoita on luokiteltu kansallisuuden ja ammatin mukaan, joiden määrä lasketaan sadoissa.

Paikkaontologiat. Paikkaontologiat vastaavat kansallisten maanmittauslaitosten ylläpitämiä paikannimirekistereitä. Niiden avulla voidaan yksilöidä paikat, sijoittaa ne koordinaatistoon ja tallentaa paikkoihin liittyviä lisätietoja, kuten paikkatyyppi (kylä, järvi, asema jne.).

Aikaontologiat. Monella alalla keskeistä on ajan esittäminen. Lineaarisen kalenteriajan ohella voidaan viitata myös päivän ja vuodenkierron aikoihin (kevät, aamu) sekä nimettyihin aika- ja tyylikausiin (esim. rauta-aika, valistuksen aika, art deco).

Tapahtumaontologiat. Tapahtumat, kuten Porvoon valtiopäivät tai Napoleonin kruunajaiset, liittävät toisiinsa henkilöitä, paikkoja, aikoja ja toisia tapahtumia, ja mahdollistavat kulttuuristen sekä provenienssitietoon liittyvien ilmiöiden kuvaamisen semanttisesti yhteentoimivalla tavalla. Tästä johtuen mm. museoalalla standardoidun CIDOC-CRM-järjestelmän[16] perustaksi on valittu tiedon esittäminen tapahtumina, eikä esimerkiksi Dublin Coren[17] dokumenttiperustaista tietomallia.

Nimistöontologiat. Monilla tieteenaloilla on käytössä laajoja nimistöjä, joita voidaan liittää yleisontologioihin. Tällaisia ovat esimerkiksi biologian eliöiden lajilistat, geologian mineraalit, lääketieteen taudit ja lääkeaineet, kielitieteen kielet yms. Esimerkiksi FinnONTO:n YSO-ontologian käsite ”linnut” lavenee kattavaksi, yli 11 000 maailman lintulajin taksonomiaksi AVIO-ontologian kautta.

12 http://www.onki.fi/
13 http://vesa.lib.helsinki.fi/ysa/
14 http://www.getty.edu/research/tools/vocabularies/aat/
15 http://www.getty.edu/research/tools/vocabularies/ulan/
16 http://www.cidoc-crm.org/
17 http://dublicore.org/

Ontologiat yhdistävät tietoa ja parantavat tiedon hakua

FinnONTO:ssa luotu kansallinen ONKI-ontologiapalvelupilotti on ollut verkossa vuodesta 2008 ja sillä on ollut noin 14 000 ihmiskäyttäjää kuukaudessa. Palvelun rajapintoja on rekisteröitynyt käyttämään yli 400 eri tahoa. Ontologiasysteemin ytimessä ovat yleiskäsiteontologiat, joita muut ontologiat eri tavoin tarkentavat. Myös ONKI:sta tuotettu Finto-palvelu keskittyy näihin ontologioihin. Esittelen seuraavassa lyhyesti FinnONTO:n tuloksia tällä ontologiatyön osa-alueella.

Kehitystyön lähtökohdaksi otettiin maassamme jo käytössä olevat asiasanastot (Hyvönen, 2005). Niiden käytöllä voidaan parantaa ja yhdenmukaistaa tiedon indeksointia, jolloin tiedon haussa päästään parempaa tarkkuuteen (*precision*) ja saantiin (*recall*). Erikielisiin asiasanoihin perustuvaan tietojen indeksointiin, linkittämiseen ja hakuun liittyy kuitenkin vakavia haasteita. Oletetaan, että etsit ”helsinkiläisiä ravintoloita” tietojärjestelmästä. Perinteinen tiedonhakujärjestelmä vertailee hakutermejä ”Helsinki” ja ”ravintola” tietokantaan tiedon tallennusvaiheessa indeksoituihin kuvailutermeihin ja palauttaa hakutuloksena vastaavat tietueet. Mikäli hakusanaa vastaavaa merkkijonoa ei löydy tietueesta sopivasta paikasta saadaan tyhjä tulos. Jos esimerkiksi tietokantaan on talletettu ”Kalliossa” oleva ”pizzeria”, ei se löydy hakusanoilla ”Helsinki” ja ”ravintola”. Tässä ei auta sanojen taivutusmuotojen tai synonyymien huomioiminenkaan. Myös kielimuurit ovat perinteisen merkkijonohaun haasteena: ”ravintolalla” haettaessa ei ”restaurant” löydy.

Ontologiaperustainen hakupalvelu sen sijaan ymmärtää tarjota tuloksena ”Kalliossa” sijaitsevaa ”pizzeriaa”, koska haun apuna käytetään maailman käsitteitä kuvaavia ontologioita. Ne kertovat, että ”Kallio” on osa Helsinkiä ja että ”Kallio” tässä yhteydessä ei tarkoita luonnonmuodostelmaa tai presidentti Kyösti Kalliota. Samoin ”pizzeria” voidaan tunnistaa eräänlaiseksi ”ravintolan” alatyypiksi. Myös kielimuurit luhistuvat, sillä ontologinen indeksointi ja haku perustuvat kieliriippumattomiin tunnisteisiin eikä kielellisiin ilmauksiin. Ontologisten viittausten kautta tieto yhdistyy luontevasti toisiin tietoihin: esimerkissämme järjestelmä voi tarjota lisätietoa pizzakulttuurista Wikipedian kautta, mikäli hakupalvelussa käytetään samoja käsitteitä kuin DBpediassa. Voidaan myös luoda linkki lähellä olevan teatterin Italia-aiheista revyytä käsittelevään artikkeliin, koska pizza on italialainen ruokalaji, tai tarjota matkapuhelimeen navigointipalvelu Kallioon.

Linkitetty ontologiapilvi KOKO

FinnONTO:n ontologisointityö alkoi maamme käytetyimmästä asiasanastosta, Kansalliskirjaston Yleisestä Suomalaisesta Asiasanastosta YSA, joka sisälsi FinnONTO-hankkeen alussa vuonna 2003 yli 20 000 yleiskäsitettä eri aloilta. Sanaston laajentaminen oli edelleen tarpeen, mutta kaikkien eri alojen asiantuntemuksen saaminen sanastotyöhön oli haasteellista, vaikkakin ylläpitoa varten oli olemassa monialainen työryhmä. Samaan aikaan maassamme oli käytössä ja kehitettiin lukuisia YSA:n kanssa erityisesti yleisempien käsitteiden osalta päällekkäisiä erityissanastoja, mikä tuntui monasti tarpeettomalta moninkertaiselta työltä. Lisähaasteena oli, että samannimisillä käsitteillä saattoi olla eri sanastoissa erilainen merkitys, mikä johtaa sekaannuksiin tietoja yhdistettäessä.

FinnONTO tarttui näihin sisällöllisiin ja organisatorisiin haasteisiin esittämällä kansallisesti koordinoitua mallia, jossa sanastotyö voitaisiin yhteistuumin jakaa eri alojen sanastojen kehittäjäryhmien kesken ja yhdistää tulokset lopulta yhdeksi laajaksi, eri alat kattavaksi ontologiaksi. Vision mukaan tämä edistäisi merkittävästi eri alojen tietosisältöjen semanttista yhteentoimivuutta ja säästäisi samalla kehityskustannuksia päällekkäisen työn vähentyessä.

Työn tuloksena syntyi prototyyppi KOKO-ontologiasta, joka on eri alojen ontologioiden muodostama *Linked Data* -ontologiapilvi. Sen ytimenä on YSA:sta kehitetty Yleinen Suomalainen Ontologia YSO, joka muodostaa KOKO:n yläontologian (*upper ontology*) sisältäen merkitysalaltaan laajimmat käsitehierarkian käsitteet. Erikoisalojen tarkempi käsitteistö ”ripustuu” sit-

Kuva 1. Yhteisöllinen kokonaisontologia KOKO koostuu yläontologiasta YSO ja sitä tarkentavista alaontologioista. Ontologioiden keskinäiset leikkaukset kuvassa ovat vain viitteellisiä ja tarkoitettu yleiskuvan havainnollistukseksi.

ten YSO:n eri haaroihin hierarkioita syventäen. Kuva 1 havainnollistaa mallia (Seppälä, Hyvönen, 2014).

Lähes jokainen KOKO-pilven ontologia perustuu maassamme jo aiemmin käytettyyn vastaavaan asiasanastoon. Alojen jako taulukon 1 mukaisesti ei ole optimaalinen, vaan esimerkiksi monia pienempiä toisiinsa liittyviä ontologioita kannattaa jatkossa yhdistää laajemmiksi kokonaisuuksiksi. FinnONTO-hankkeen puitteissa tähän ei kuitenkaan haluttu ottaa kantaa eikä hankkeessa myöskään ryhdytty osaontologioiden päällekkäisyyksien poistamiseen: nämä tehtävät jätettiin sanastojen kehittäjätahojen asiaksi.

KOKO:n osaontologiat on ontologisoitu yhtenäisellä menetelmällä lukuun ottamatta kansainvälistä lääketieteen MeSH-sanastoa, jonka rakenne on alkuperäinen. Joissain tapauksissa ontologisoinnin yhteydessä alkuperäistä sanastoa laajennettiin ja YSA/YSO:n osalta valmistui sanaston ensimmäinen englanninnos, mikä mahdollistaa sen linkittämistä kansainvälisiin ontologioihin.

Haasteena asiasanaston ontologisoinnissa on sen semanttisen rakenteen täsmentäminen niin, että siinä toteutuvat koneellisessa päättelyssä, esimerkiksi kyselyn laajentamisessa, tarvittavat ominaisuudet. Keskeisimmät ontologiset muutokset ja tarkistukset, joihin työssä keskityttiin, olivat:

Luokkahierarkian täydentäminen. Ontologisen luokkahierarkian muodostaminen ja täydentäminen niin, että kaikilla käsitteillä (kaikkein ylin pois lukien) on ainakin yksi yläkäsite. Tämä parantaa käsitteiden linkitettävyyttä ja mahdollistaa muun muassa päättelyä.

Laajempi termi/suppeampi termi -suhteiden tarkentaminen. Asiasanastoissa käytettyjen laajempi termi/suppeampi termi (LT/ST) -suhteiden tarkentaminen yläluokkasuhteiksi (rdfs:subClassOf), osa–kokonaisuus-suhteiksi tai assosiatiivisiksi suhteiksi. Työssä keskityttiin subClassOf-suhteiden kehittämiseen.

KOKO-ONTOLOGIAAN KUULUVAT ONTOLOGIAT

Yleinen suomalainen ontologia	n. 26 000	Yleinen suomalainen asiasanasto
Julkishallinnon ontologia (JUHO)	n. 6 350	Valtioneuvoston asiasanasto
Kaunokirjallisuuden ontologia (KAUNO)	n. 5 100	Kaunokki-asiasanasto
Kielitieteen ontologia (KTO)	n. 950	Uralistiikan tutkimuksen bibliografian asiasanaluettelo
Kirjallisuudentutkimuksen ontologia (KITO)	n. 850	Kirjallisuudentutkimuksen asiasanasto
Kulttuurien tutkimuksen ontologia (KULO)	n. 1 500	Kulttuurien tutkimuksen asiasanasto
Liiketoimintaontologia (LIITO)	n. 3 400	Yrityssuomi.fi-portaalin käsitteistö
Merenkulkualan ontologia (MERO)	n. 1 400	Merenkulkualan asiasanasto
Musiikin ontologia (MUSO)	n. 1 000	Musiikin asiasanasto
Puolustushallinnon ontologia (PUHO)	n. 2 000	Puolustushallinnon asiasanasto
Taideteollisuusalan ontologia (TAO) / Museoalan ontologia (MAO)	n. 8 100	Muotoilun ja viestinnän asiasanasto / Museoalan asiasanasto
Terveyden ja hyvinvoinnin ontologia (TERO)	n. 6 500	TESA, Stameta-asiasanasto, n. 2500 MeSH-käsitettä (Medical Subject Headings)
Valokuvausalan ontologia (VALO)	n. 2 000	Valokuvan asiasanasto
Viikin tiedekirjaston ontologia (AFO)	n. 6 000	Agriforest-asiasanasto

Taulukko 1. KOKO-ontologiapilven osaontologioita. KOKO-ontologiapilveen kuului vuoden 2013 lopulla taulukossa luetellut 15 ontologiaa, yhteensä yli 71 000 käsitettä.

Yksilö–luokka-suhteiden transitiivisuuden tarkistaminen. Keskeinen ontologioissa käytetty periaate luokkahierarkian muodostamisessa on, että kaikkien luokkien yksilöiden tulee olla samanaikaisesti *kaikkien* yläluokkiensa yksilöitä. Tämä ominaisuus on hyödyllinen koneellisessa päättelyssä, esimerkiksi ominaisuuksien periytymisessä ja kyselyn laajentamisessa.

Monimerkityksisten asiasanojen merkitysten erottelu. Asiasanastojen monimerkityksisten käsitteiden merkitysten jakaminen eri käsitteiksi tarpeen mukaan niin, että ne voidaan sijoittaa luokkahierarkiaan.

Tarkastellaan esimerkkinä transitiivisuuden haasteesta asiasanoja taskupeilit ja peilit, jotka on määritelty asiasanastoissa seuraavien laajempien termien avulla (LT-suhde):

taskupeilit LT *peilit*
peilit LT *huonekalut*

Tämän mukaan taskupeilit on peilejä ja toisaalta peilit ovat huonekaluja, mikä kuulostaa järkevältä, mutta transitiivisuus molempien suhteiden ylitse ei toimi, sillä taskupeilit eivät ole huonekaluja. Jos LT-suhteet muutetaan mekaanisesti luokkasuhteiksi (subclassOf) seurauksena on, että haettaessa huonekaluja hakutulokseen voi tulla virheellisesti mukaan taskupeilejä. Transitiivisuuden tarkistaminen on erityisen haastavaa KOKO-ontologiapilvessä, kun luokkahierarkiat kulkevat eri ontologioiden kautta. Taskupeilit löytyy Museoalan asiasanastosta MASA, muttei YSA:sta.

Esimerkiksi asiasanojen monimerkityksellisyyden haasteista sopii YSA:ssa oleva asiasana ”lapset”, joka voi tarkoittaa mm. ikäluokkaa tai perhesuhdetta. Haasteena on, että aikuisetkin ovat omien vanhempiensa lapsia. Siksi YSO-ontologiassa lapset-käsitteestä on erotettu merkitykset ”lapset (ikään liittyvä rooli)” ja ”lapset (perheenjäsenet)”. Mikäli erottelua ei tehtäisi, saattaisi tietojärjestelmä esimerkiksi suositella aikuisille tarpeettomasti lastenkirjoja luettavaksi, koska nämä ovat vanhempiensa lapsia.

FinnONTO-hankkeessa neuvoteltiin lupa julkaista kaikki KOKO-ontologiat avoimena datana. Tulosten hyödyntäminen myös kaupallisiin tarkoituksiin on vapaata, ainoastaan aineiston kehittäjätaho on mainittava (MIT License, CC 3.0). Edellytykset työn jatkamiselle ja tulosten hyödyntämiselle ovat tältä osalta paremmat, kuin hankkeen alkaessa.

Ontologiatyön arviointia

Ontologiatyö, jo pelkästään insinöörien filosofiasta uusiokäyttöön ottama termi ontologia, on herättänyt myös vastustusta. Kritiikkiä on synnyttänyt ontologisoinnin yhteydessä syntyvä yhä tarkempi jaottelu merkityksiin, mikä johtaa yhä isompaan ja mutkikkaampaan käsitehierarkiaan. Tämä lisää paitsi ylläpitotyötä myös työtä indeksoinnissa ja saattaa rajata termeihin liittyviä vivahteita liikaa. Jos esimerkiksi kuvaillaan Albert Edelfeltin teos ”Pariisin Luxembourgin puistossa”, pitääkö siinä leikkivät lapset kuvata erikseen sekä ikäryhmänä, perhesuhteena että sosiaalisena ryhmänä, ja katetaanko näilläkään lapsiin liittyvä merkitysten ala?

Ontologiatyössä mennäänkin helposti liian hienojakoisiin ratkaisuihin, jos järjestelmän käyttötarkoitus pääsee unohtumaan. FinnONTO:ssa tähän vaaraan varauduttiin asettamalla tavoitteeksi minimaalinen ja mahdollisimman yleiskäyttöinen sanastojen ontologisointi, joka kuitenkin ratkaisisi eräitä keskeisimpiä asiasanastoihin ja asiasanastotyöhön liittyviä haasteita tietotekniikan sovellusten kannalta. Ontologian laajenemisen ongelma liittyy myös metatietomallien kehittämiseen, mikä kannattaa ottaa jatkossa huomioon. Ontologioitahan käytetään metatiedon esittämiseen eikä yksinään. Esimerkiksi FinnONTO-hankkeen piirissä kehitetty julkishallinnon JHS suositus 183 ”Julkisen hallinnon palvelujen tietomalli ja ryhmittely verkkopalveluissa” päätyi luomaan yhden laajan palveluontologian sijasta yleisemmän metatietomallin, jonka metatietokentille voidaan valita arvoja eri ontologioista.

Ehkä keskeisin periaatteellinen kysymys ontologiatyössä on, missä määrin maailmaa voidaan ja on tarkoituksenmukaista kuvata loogisena struktuurina. Maailma ja siihen liittyvä tieto on esimerkiksi monin tavoin epävarmaa (*uncertainty*), puutteellista (*incomplete*) ja sumeaa (*fuzzy*). Eikö maailman loogisessa kuvaamises-

sa jo epäonnistuttu kerran 1980-luvulla tekoälytutkimuksen asiantuntijajärjestelmien yhteydessä? Semanttisen webin lähtökohdat ovat kuitenkin logiikan soveltamisen suhteen hyvin erilaiset verrattuna tekoälyn asiantuntijajärjestelmiin. Linked Data -ajattelussa korostetaan, että ontologioiden ei tarvitse olla virheettömiä ja loogisesti eheitä voidakseen olla silti hyödyllisiä, päinvastoin kuin vaikkapa sytostaattihoitoa annostelevan asiantuntijajärjestelmän. Eihän WWW ylipäätään ole virheetön, mutta silti hyödyllinen. On kuitenkin selvää, että täsmälliseen tietoon perustuva logiikka asettaa rajoitteita luonteeltaan monimuotoisen ja epätäsmällisen maailman kuvaamisella ja että yhä automaattisemmin menetelmin ja yhteisöllisemmin tuotetun yhdistetyn datan laatu tulee olemaan keskeisiä haasteita semanttisessa webissä.

Haasteista huolimatta webin kehitys on kuitenkin astunut uudelle semanttiselle tasolle eikä paluuta entiseen ole. Web of Data luo WWW:n sisään uuden, W3C:n standardeihin perustuvan sisältökerroksen ja infrastruktuurin, joka on monin tavoin hyödyllinen. FinnONTO:n ontologioita ja ONKI-palvelua esimerkiksi on käytetty hyväksi monissa sovelluksissa, kuten Terveyden ja hyvinvoinnin laitoksen Tervesuomi-palvelussa (Suominen ym., 2011), Yleisradion verkkosivuilla, kymmeniä kulttuurialan kokoelmia linkittävässä Kulttuurisampo-palvelussa (Mäkelä ym., 2012), työ- ja elinkeinoministeriön YritysSuomi-palvelussa, monissa maamme taide- ja kulttuurihistoriallisissa museoissa sekä yleisten kirjastojen Kirjasampo-portaalissa (Mäkelä ym., 2011), jolla on nykyisin jo 65 000 käyttäjää kuukaudessa.

Lähteitä

G. Antoniou, F. van Harmelen: *A Semantic Web Primer* (3. p.). The MIT Press, 2012.

J. Aitchison, A. Gilchrist, D. Bawden: *Thesaurus Construction and Use: A Practical Manual*. Europa Publications, London, 2000.

Tom Heath, Christian Bizer: *Linked Data: Evolving the Web into a Global Data Space*. Claypool & Morgan, CA, USA, 2011.

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

Eero Hyvönen: Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita? *Tietolinja*, 2005.

Eero Hyvönen: *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Claypool & Morgan, CA, USA, 2012.

Eero Hyvönen, Kim Viljanen, Jouni Tuominen, Katri Seppälä: Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach. *Proceedings of the European Semantic Web Conference ESWC 2008*, Springer-Verlag, 2008.

Eetu Mäkelä, Eero Hyvönen , Tuukka Ruotsalo: How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web Journal*, 2012.

Eetu Mäkelä, Kaisa Hypén, Eero Hyvönen: BookSampo – Lessons Learned in Creating a Semantic Portal for Fiction Literature. *Proceedings of ISWC-2011*, Bonn, Germany, Springer-Verlag, 2011.

Katri Seppäla, Eero Hyvönen: *Asiasanaston muuttaminen ontologiaksi. Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista*. Suunnitelmia, selvityksiä, oppaita. Kansalliskirjasto, 2014 (ilmestyvä). http://www.doria.fi/handle/10024/67050

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

Steffen Staab, Rudi Studer (toim.): *Handbook on Ontologies*. Springer-Verlag, 2009

Jouni Tuominen, Matias Frosterus, Kim Viljanen, Eero Hyvönen: ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.

Kim Viljanen, Jouni Tuominen, Eero Hyvönen: Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.

Kirjoittaja on Aalto-yliopiston professori Semanttisen laskennan tutkimuskeskuksessa (SeCo).