Avoin data ja semanttinen verkko — yhdessä kohti älykkäämpää internetiä

Hannu Lohtander

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Pro gradu -tutkielma
Ohjaaja: Erkki Mäkinen
Huhtikuu 2013

Tampereen yliopisto
Informaatiotieteiden yksikkö
Tietojenkäsittelyoppi
Hannu Lohtander: Avoin data ja semanttinen verkko – yhdessä kohti älykkäämpää internetiä
Pro gradu -tutkielma, 55 sivua, 14 liitesivua
Huhtikuu 2013

Digitaalisen vallankumouksen tuoma datan määrän räjähdysmäinen kasvu on tuonut esiin toisaalta haasteita mutta myös mahdollisuuksia datan hyödyntämiseksi. Samaan aikaan käynnissä oleva avoimen ideologian esiinmarssi ja datan hyödyntämiseen tähtäävien teknisten menetelmien kehitys on muuttamassa suhtautumistamme dataan. Datasta on tulossa seuraava internetin resurssi. Internetin standardointiin tähtäävän W3-organisaation tavoitteena on tukea tätä kehitystä, ja se tuottaa tätä varten datan laadun parantamiseksi tarkoitettuja määrittelyitä. Datan kuvaamiseen tehdyt ja semanttisen datan ja semanttisen verkon mahdollistavat määrittelyt ovat näistä keskeisimmät. Avoimen datan ideologia on saanut julkiset instituutiot avaamaan dataa, ja tässä yhteydessä datan laadulle asetetaan vaatimuksia. Arvioidessani julkisen avoimen datan laatua tähän tarkoitukseen esitellyllä viiden tähden asteikolla tulen siihen tulokseen, ettei tämän datan laatu vastaa semanttisen verkon vaatimuksia.

Avainsanat ja -sanonnat: avoin data, semanttinen data, semanttinen verkko

Sisällys

1. Johdanto

Tietoyhteiskunnan toiminnot ja niiden tukemiseen tarkoitetut lukuisat verkkosovellukset tuottavat suuria määriä tallennettavaa dataa. Useimmiten tämä data on suljettua ja tallennettuna sovellusten yksityisiin tietovarastoihin. Osa datasta voi olla näkyvissä tietovarantoa hyödyntävien sovellusten käyttäjille, mutta silloin se ei usein ole alkuperäisessä muodossa raakadatana. Viime vuosina joitakin tällaisia tietovarantoja on avattu avoimesti saatavaksi, ja tällöin suljetusta datasta on tullut avointa dataa.

Raakadata on usein tietoalkioita tietokantariveinä, oliokantojen olioina tai Excel-taulukoita. Datan merkitys avautuu usein vasta, kun se otetaan käyttöön jossakin kontekstissa. Merkityksen luominen dataan, kun se ei ilmene sen käyttöyhteydestä, tapahtuu kiinnittämällä siihen metatietoa, datan merkityksen kuvaavaa informaatiota. Tällainen merkitysten liittäminen dataan toteutetaan metatiedolla, ja se tekee mahdolliseksi datan käsittelyn ohjelmallisesti niin, että dataa työstävät sovellukset kykenevät ottamaan merkityksen huomioon ja tuottamaan parempia tuloksia. Metatiedon semanttinen rikkaus riippuu siitä, miten laajasti dataa on kuvailtu ja kuinka syvä on käytetty metatietokuvaus.

Linkittämällä kuvauksia sisältävää dataa, tuottamalla siihen merkityksiä, jotka aukeavat seuraamalla luotuja polkuja toisiin datajoukkioihin ja niistä löydettyihin kuvauksiin, saadaan aikaan linkitettyä dataa. Avoin linkitetty data on avointa dataa, joka liitetään osaksi laajempaa datan verkkoa, jossa liikutaan solmujen välillä seuraamalla kuvauksiin tuotettuja reittejä. Linkitetty data sisältää siis aina myös kuvausinformaation. Tällaisen datan luomaa verkkoa kutsutaan semanttiseksi verkoksi (eng. Semantic Web). Hakukoneet ovat ilmeinen sovellusalue, jossa tällaisella semanttisella kuvauksella on keskeinen merkitys sovelluksen toiminnalle, joskin informaatiota etsivä agenttisovellus on semanttisen verkon tärkein käyttäjä.

Suljetun datan avaaminen avoimeksi raakadataksi tai dataksi, jossa on semanttinen tieto mukana, tai peräti avoimeksi linkitetyksi dataksi, vaatii tietyt vaiheet, joissa data kuvataan, tehdään saavutettavaksi ja yhdistetään toisiin datajoukkioihin. Tällaiset kuvaukset ovat hyödyllisimmillään, kun ne tehdään yhteisten sopimusten mukaan. Näihin kuvauksiin ja sopimuksiin liittyen tämän tutkielman tarkoitus on selvittää, millä tavalla piiloon jäävä data tuodaan julkiseksi ja internetin kautta vapaasti saatavaksi, missä vaiheessa semanttinen verkko on, mitä merkitysten antaminen datalle tarkoittaa, miten datalle annetaan merkityksiä, miten data linkittyy toisiin datajoukkioihin, ja millaisia työkaluja on olemassa tai kehitteillä.

Tutkin avointa dataa tarjoavien julkisten instituutioiden julkaisemia datoja tarkoituksenani selvittää, millä tavalla ne liittyvät semanttiseen verkkoon. Tutkielman toisena tarkoituksena on selvittää, millä tavoin avoimena datana julkaistua raakadataa voidaan jalostaa osaksi semanttista verkkoa. Kolmanneksi toivon tutkielman myös jäsentävän eri käsitteet niin, että lukija kykenee

seuraamaan käynnissä olevaa aktiivista yhteiskunnallista keskustelua ymmärtäen paremmin siinä käytetyt termit, teorian ja niiden osoittamien asioiden nykytilan.

Luvussa 2 tutkin digitalisoitumisen ja tietoyhteiskunnan muutoksia ja mitä näistä on seurannut käsitteen ”avoin data” suhteen. Esittelen avoimen datan käsitteen ja siihen sisältyvät merkitykset ja ideologisen ulottuvuuden. Esittelen myös olemassa olevia avoimen datan verkkosivustoja ja niihin tuotettavaa dataa.

Luvussa 3 käyn läpi merkitysten antamisen raakadatalle niin, että siitä tulee semanttista dataa. Erityisesti W3-organisaation puitteissa työstetään standardeja datan kuvaamiseen, ja esittelen tämän työn tuloksia. Tutkin siis sitä, millä tavalla data linkittyy keskenään.

Luvussa 4 tarkastelen semanttisen verkon käsitettä. Semanttinen verkko muodostuu semanttisen verkon solmuista, joista agenttisovellukset etsivät semanttista dataa. Esittelen tämän verkon taustalla olevan teknologian.

Luvussa 5 otan tarkempaan tarkasteluun avoimen datan sivustoja ja tutkin niiden kelpoisuutta semanttisen verkon laatukriteereillä mitattuna. Esittelen olemassa olevan laatumittarin ja sovitan sen esimerkiksi valittuihin avoimen datan sisältöihin.

Lopuksi luvussa 6 teen yhteenvedon avoimen datan analyysin tuloksista ja pohdin hieman sitä, miltä tutkimuksen valossa näyttää avoimen datan ja semanttisen verkon lähitulevaisuus – ja onko niillä edessä yhteinen vai erikseen kuljettava taival.

2. Avoin data

Kaikki on dataa, energiaa tai materiaa, ja sen alkuperä on alkuräjähdyksessä. Koska tämä tutkielma ei ole kirjoitettu teoreettisen fysiikan tai filosofian alueelta, sen aihepiiri rajataan yritysten, yhteisöjen ja yhteiskunnan tuottamaan dataan. Erityisesti mielenkiinto kohdistuu internetissä julkaistuun dataan ja sellaiseen dataan, joka sinne olisi julkaistavissa. Datan rakenne ja merkityksen tuottaminen siihen on tässä tutkielmassa keskeistä.

Avoin data voidaan nähdä tarkoittavan kahta asiaa: yhtäältä se liitetään yhteiskunnalliseen keskusteluun kansalaisille kuuluvasta julkisesta datasta ja datan kaupallisesta merkityksestä, ja toisaalta sillä tarkoitetaan tietyn teknisen määrittelyn mukaista dataa. Poliitikot ja kansalaiset näkevät keskustelusta sen ideologisen puolen, kun taas sovelluskehittäjät näkevät keskustelun taustalla olevat tekniset haasteet.

Tässä luvussa tarkasteluun otetaan data. Data on analogista tai digitaalista. Sen kantamaan tietoon liittyy aina vallankäytön ja ideologian ulottuvuus, ja filosofian puolelta siihen liittyy tiedon, ymmärryksen ja viisauden kaltaisia käsitteitä. Tässä luvussa tarkastelen näitä ja selvitän, mistä data ja avoimuus – avoin data – tulevat.

2.1. Digitaalinen vallankumous

Digitaalinen vallankumous on siirtymä, jossa analoginen teknologia on vaihtunut digitaaliseen teknologiaan. Erityisesti transistorin ja mikroprosessorin keksiminen (transistori vuonna 1947 ja mikroprosessori vuonna 1971) ovat mahdollistaneet tämän siirtymän. Analogisesta digitaaliseen on siirtynyt mm. musiikin tallentaminen, television signaalin lähettäminen, puhelimet ja kamerat. Nämä välittävät ja tallentavat tietoa digitaalisessa muodossa. Internet on valtava digitaalisen informaation tallennuspaikka, jonka informaation tuottamisen vauhtia kuvaa Neil Spencerin koostama visualisointi joka minuutti tapahtuvien internetpalveluiden käytön määrästä (Spencer, 2012).

Useimmiten dataa tallentuu erilaisten sovellusten käytön yhteydessä. Sovellukset puolestaan voivat olla hyvin erilaisia: ne voivat olla verkkosovelluksia, joiden nimenomaisena tarkoituksena on kerätä dataa, tai ne voivat olla kauppapaikkoja, joiden käytön seurauksena syntyy dataa. Välineet voivat olla myös digitaalikameroita tai automaattisia antureita, jotka mittaavat säätä tai tuotantoprosessia. Dataa syntyy yksityisten ja julkisten toimintojen piiristä: Suomessa esimerkiksi Tilastokeskus kerää tilastollista dataa ja sisäministeriön alaiset toimijat tuottavat viranomaistiedotteita hätäkeskustapahtumista. Yritykset synnyttävät dataa toimintansa ohella, tai niiden toiminta perustuu omaan, itse kerättyyn tai sille luovutettuun dataan. Data voi olla myös yrityksen myymä tuote.

Googlen toimitusjohtaja Eric Schmidt arvioi vuonna 2011 (Underwood, 2011), että internetissä olevan datan määrä on 5 miljoonaa teratavua. Samassa yhteydessä hän esitti väitteen, jonka mukaan Googlen hakukone on indeksoinut vain 0,004 prosenttia tästä datamäärästä. Datan määrän kasvu on ollut eksponentiaalista, ja tämä on mahdollista, koska sovellukset tuottavat siitä suurimman osan. Tällaisessa maailmassa datan hyödyntäminen on kasvava haaste, ja tämän haasteen ajankohtaisuutta kuvaa uusi käsite ”Big Data”, jonka käyttö on yleistynyt vuoden 2012 aikana. Termin taustalla on tämän valtavan digitaalisen datamassan analysointi tarkoituksena löytää sieltä hyödyllistä informaatiota.

2.2. Suljettu data ja sen avaaminen

On olemassa syitä, joiden vuoksi dataa ei voida antaa tai ei ole annettu avoimeen käyttöön. Data voi olla yksittäisen sovelluksen käyttämää dataa, jonka avaamiselle ei ole nähty mitään syytä – eikä sellaista välttämättä olekaan. Usein data on suojattua vahingonteon estämiseksi, lain velvoitteesta tai koska sen sisältämällä tiedolla on kaupallista arvoa. Erityisesti yrityksellä voi olla dataa, jota se ei voi tai halua avata kilpailijoiden nähtäväksi ja käytettäväksi. Myös julkishallinto voi joutua kilpailulainsäädännön vuoksi pitämään datan suljettuna. Usein data on suojattua sen saavutettavuuden osalta kuitenkin vain tottumuksesta tai varmuuden vuoksi – datan pitämiseksi suljettuna ei ole välttämättä perusteita. Data voi olla myös suljettua dataa, mutta ei yksityistä dataa. Esimerkiksi julkishallinto tuottaa paljon dataa, joka saattaa olla suljettua, mutta ei ole yksityistä. Suljettu data on siksi mielenkiintoista, että se sisältää paljon potentiaalisesti avointa dataa, ja tästä esimerkkinä osa aiemmin suljetusta datasta on julkaistu viime vuosien aikana avoimena datana.

Avoin data on dataa, joka on saatavilla internetistä ja jonka lisenssi ei rajoita sen käyttöä. Tämä on se muotoilu, jonka voi katsoa olevan tämänhetkinen avoimen datan määritelmä. Hyvää yksiselitteistä vakiintunutta määritelmää käsitteelle ei ole, mutta Wikipedia (”Open Data”, 2013) ja Open Definition (”Open Definition”, 2013) -sivustot sekä muut vastaavat tahot määrittävät sen saman suuntaisesti. Saatavuuden ja rajoitusten aste voi vaihdella. Saatavuus on toteutettu yksinkertaisimmillaan siten, että raakadata annetaan avoimesti ladattavaksi verkko-osoitteesta ilman minkäänlaisia esteitä, kuten rekisteröitymistä palveluun tai maksua. Saatavuuteen liittyy löydettävyyden haasteet. Data voi olla avointa, mutta internetin koon ja hakukoneiden vajavuuksista johtuen dataa ei välttämättä löydetä. Tällaista avointa mutta piiloon jäävää dataa on olemassa suuria määriä. Yhtenä pienenä keinoa tähän haasteeseen vastauksia etsittäessä on dataa kuvailevan informaation lisääminen datan yhteyteen sen löytämisen ja hyödyntämisen helpottamiseksi. Lisäksi voidaan perustaa portaalipalveluita, joihin data koostetaan keskitetysti löydettävyyden parantamiseksi.

Jos unohdetaan hetkeksi avoimen datan nykymerkitys, voidaan todeta, että termiä on käytetty aiemminkin. Se on liitetty keskusteluun aiheissa, joissa avoimuus on ollut uhattuna, kuten

tieteellisen tutkimuksen piirissä. Ei kaikki data ole suinkaan ollut suojattua, mutta varsin paljon on sellaista dataa, jonka voi katsoa kuuluvaksi vapaasti jokaisen saatavaksi. Sen kysymyksen selvittäminen, kuka tietoa omistaa ja millaiset eettiset tai kaupalliset kysymykset aiheeseen liittyvät, on toisen debatin aihe. Mielenkiintoista on, että julkishallinto ja osaltaan myös yrityksen ovat ymmärtäneet, että dataa ja sen sisältämää tietoa kannattaa avata. Datan avaamisen toivotaan luovan myös uusia kaupallisia mahdollisuuksia ja siten mahdollistavan uusia innovaatioita internetin muuttamassa taloudessa.

2.3. Ideologinen ulottuvuus

Julkisesti rahoitettujen instituutioiden keräämä data käsitetään yhteiseksi omaisuudeksi, ja poliitikot ovat toimineen tämän vallitsevan ideologian mukaisesti. Valtiot ovat viime vuosina Suomen tapaan ilmoittaneet avaavansa julkista dataa verkkoon. Esimerkiksi Yhdysvallat avasi oman avoimen datan portaalin vuonna 2009, Norja vuonna 2010, Espanja vuonna 2011 ja Intia vuonna 2012. Suomessa vastaavaa portaalia suunnitellaan Tietoyhteiskunnan kehittämiskeskus ry:ssä, jonka jäsenistö muodostuu keskeisistä yhteiskunnallisista ja liike-elämän toimijoista. Valtioita tai suuria ylikansallisia julkisia instituutioita, joilla on avoimen datan julkaisuun tarkoitettuja sivustoja, on ainakin yli 20 – todennäköisesti niitä on olemassa enemmän, ja joka tapauksessa niitä on tulossa lisää. Myös pienemmät toimijat tuottavat avoimen datan portaaleita, ne vain jäävät suurten toimijoiden pimentoon.

Avoin data liittyy yleisempään avoimen, vapaan tai ilmaisen ideologian ympärillä käytävään yhteiskunnalliseen keskusteluun. Tätä keskustelua on käyty jo pitkään ennen avoimen datan keksimistä: Steven Levy (2001) kirjasi vuonna 1984 julkaistuun hakkerietiikkaan sen kolmanneksi periaatteeksi: ”informaatio tahtoo olla ilmaista” (ajatus tulee kuitenkin joidenkin lähteiden, mm. Wagnerin (2003), mukaan alun perin Steve Brandilta).

Lisensoinnilla tarkoitetaan niitä ehtoja, joilla jonkun omistama keksintö, idea tai tuotos voidaan ottaa omaan käyttöön. Tässä tutkielmassa käydään läpi joukko määrittelyitä, jotka ovat lisenssien alaisia tuotoksia. Mikään tässä tutkielmassa myöhemmin esitellyistä määrittelyistä tai standardeista ei kuitenkaan ota kantaa sen tiedon lisensointiin, jota määrittelyiden mukaan kuvataan ja käsitellään. Määrittelyt itsessään ovat vapaasti käytettävissä W3C:n patenttikäytännön mukaisesti (Daniel, 2002). Niiden perusteella tehdyt tuotokset, tässä tapauksessa metatietokuvailut, voidaan julkaista millä hyvänsä lisenssillä – kuvaus ei siis peri määrittelyn lisenssiä. Metatieto voidaan julkaista avoimella tai suljetulla lisenssillä. Määrittelyitä tekevissä ryhmissä on akateemisten ja yhteiskunnallisten toimijoiden edustajia, mutta myös yritysten edustajia, joiden kaikkien etu on toimia yhteisesti hyväksytyin säännöin. Osa määrittelyistä saatetaan tehdään yrityksissä näiden tunnistamiin tarpeisiin, ja ne viimeistellään yhteisissä työryhmissä – näin on esimerkiksi tehty luvussa 4 esiteltävä SPARQL-kyselykielen laajennus. Yritys voi siis antaa tuotoksensa vapaasti

yleiseen käyttöön, ja jos se osallistuu W3:n työryhmiin, sen on sitouduttava tähän työryhmän tuotosten osalta.

Avoin, vapaa ja ilmainen eivät ole toisensa synonyymejä, mutta lähestyvät samaa kysymystä: kenelle tieto kuuluu? Tämän tutkielman tarkoitus ei ole olla eettinen tai moraalinen pohdiskelu aiheesta. Käytännön toiminnallaan valtiot ja yritykset osoittavat, että niissä ainakin jonkin verran uskotaan ilmaisuudesta ja avoimesta olevan sellaista etua, että ne osallistuvat avoimen ideologian piirissä tapahtuvaan toimintaan. Avoimen ideologiaan liittyy muitakin keskeisiä suuntauksia, mm:

- Avoin saatavuus (Open access)
- Avoin sisältö (Open content)
- Avoin lähdekoodi (Open source)
- Avoin tietämys (Open knowledge)
- Avoin kulttuuri (Free culture).

Avoin saatavuus liittyy tieteellisen tutkimuksen vapaaseen käyttöön. Avoin sisältö puolestaan on vapaasti käytettäviä kuvia, tekstejä ja muuta luovaa sisältöä. Avoin lähdekoodi tarkoittaa sovelluskoodia, josta jokainen saa sovelluksen kopion käyttöönsä tai voi käyttää sitä osana omaa sovellustaan. Avoin tietämys sisällyttää itseensä muita avoimen käsitteitä, erityisesti avoimen datan. Avoin kulttuuri on alun perin kirjan nimi, josta on syntynyt samaa nimeä käyttävä liike, joka toimii erityisesti tekijänoikeuskysymysten alueella. Termi liitetään myös sellaisiin yhteisöllisiin palveluihin, kuten avoin tietosanakirja Wikipedia. Avoin data on tästä kaikesta avoimen ideologiasta pieni osa, ja muut avoimen käsitteet usein sisältävät sen. Avoin data on kuitenkin monesti hyvin perustavalla tavalla se ensimmäinen avoin, josta muut johtavat omat tuotoksensa.

Digitaalisuus muuttaa perinteisiä liiketoimintamalleja. Perinteinen länsimainen markkinavetoinen talous perustuu niukkoihin resursseihin, joita joku tuottaa ja joku tarvitsee ja joita vaihdetaan vapailla markkinoilla. Tämän mallin lisäksi ilmainen on tullut yhä suositummaksi liiketoimintamalliksi (Andresson, 2009), kun kaikista resursseista ei ole enää niukkuutta digitaalisen maailman teknologioiden kehittyessä. Paitsi että digitaalisuus on mahdollistanut suurien datamäärien keräämisen ja tallentamisen, myös jakamisesta, vastaanottamisesta ja jatkokäsittelystä tulee koko ajan huokeampaa – käytännössä ilmaista. Tällainen kehitys on mahdollistanut mm. Googlen Youtube-palvelun, joka ei ole muuta kuin suunnaton määrä digitaalista videota, jota tuottavat digitaalisilla videokameroilla ja editointisovelluksilla tavalliset ihmiset ja media-alan yritykset ympäri maailman.

Avoimen lähdekoodin sovellusten alueella on jo käytännössäkin osoitettu, että avoimelle ideologialle voidaan rakentaa tuottavaa liiketoimintaa. Yritykset ovat olleet edelläkävijöitä ja tarttuneet tähän mahdollisuuteen ja tuottaneet avoimen lähdekoodin lisensseillä julkaistuja

ohjelmistoja. Vielä useammat yritykset käyttävät avoimen lähdekoodin sovelluksia osana liiketoimintamalliaan. Valtioiden rooli avoimen lähdekoodin ja avoimen datan osalta ei voisi erota toisistaan enempää: valtiot eivät juurikaan tuota avoimen lähdekoodin ohjelmistoja, vaikka käyttävät sellaisia, mutta sen sijaan avointa dataa julkaistaan erityisesti valtioiden, yhteiskunnallisten ja yhteisöllisten toimijoiden toimesta. Yritykset pyrkivät hyötymään tästä datasta, mutta eivät juuri itse tuota avointa dataa, kuten tuottavat sovelluksia. On mielenkiintoista nähdä, syntyykö tulevaisuudessa yritysmaailmaan liiketoimintaa, joka perustuu itse tuotettuun avoimeen dataan samalla tavoin kuin avoimen lähdekoodin ympärille on syntynyt.

Avoimen ideologia ei tarkoita epäkaupallista toimintaa. Usein ideologian taustalla on arvomaailma, joka pyrkii muihin kuin kaupallisiin tuloksiin. Nämä tavoitteet ovat kuitenkin usein sovitettavissa yhteen kannattavan liiketoiminnan kanssa. Tällaisen rinnakkaiselämän mahdollisuus lupaa hyvää avoimen aatteelle ja sen tulevaisuudelle. Avoimen ideologia on toimiva työkalu, ja on oletettavaa, että se tuottaa tulevaisuudessa yhä enemmän avointa dataa.

2.4. Tiedon portaat

Digitaalisen vallankumouksen luoma ja avoimen aatteen julkiseksi tuoma data on vasta ensimmäinen askel sen hyödyntämiseksi. Ackoff (1989) on käyttänyt tiedon portaat -käsitettä esittääkseen, miten ihminen mieltää dataa. Ackoffin portaat ovat data, informaatio, tieto, ymmärrys ja viisaus. Raakadata on dataa vailla minkäänlaista merkitystä. Ottamalla tuo data ja jäsentämällä sitä, kuvailemalla sitä ja löytämällä siitä sääntöjä ja merkityksiä, voidaan nousta tiedon portaita. Tässä luvussa rinnastetaan nuo portaat ja semanttinen data. Pohjustan tällä tavalla semanttisen datan olemusta, joskin sen tarkempaan määrittelyyn pureudun luvussa 3.

2.4.1. Data

Raakadataa (eng. data) on kaikkialla ympärillämme. Kaikki aistein havaittava on dataa, vailla niitä merkityksiä, joita luodaan sen mukaan, miten kukin prosessoi saamansa informaation. Tietojenkäsittelyn sovellusalueella dataa tallennetaan yleensä jonkin sovelluksen taustajärjestelmiin, jollaisia ovat mm. tietokannat, hakemistot ja tiedostot. Tietojenkäsittelyssä dataa ovat bitit, ykköset ja nollat, joita yhdistelemällä saadaan aikaan monimutkaisempia rakenteita. Data tässä merkityksessä on Merriam Websterin sanakirjan datan kolmannen määritelmän (”informaatiota numeerisessa muodossa, jota voidaan digitaalisesti välittää ja muokata”) mukaista.

Tämän tutkimuksen yhteydessä on mielekästä käsitellä dataa tietokantojen riveinä tai olioina, taulukoina tai jonkin muun järjestyneen rakenteen omaavina dokumentteina, vaikkakin data on lopulta näidenkin sovellusten taustalla vain lukuja yksi ja nolla esittäviä muutoksia jossakin tallennusjärjestelmässä. Sovellukset kadottavat käyttäjältään datan ja näyttävät yleensä siitä johdettua informaatiota.

2.4.2. Informaatio

Informaatiota (eng. information) tuottavia sovelluksia on useita; Ackoff itse nostaa esiin tietokantasovellukset esimerkeiksi tällaisesta sovelluksesta. Suuri osa esimerkiksi verkkosovelluksissa esitettävästä tiedosta on suoraan lähes sellaisenaan tietokannoista noudettua informaatiota. Tuotteen tiedot verkkokaupan katalogissa ovat informaatiota tarkasteltavasta tuotteesta – ne kostuvat teksteistä, sanoista ja numeroista, joilla oletetaan olevan jotakin merkitystä lukijalle senhetkisessä kontekstissa.

Shadboltin (2006) mukaan ”The Semantic Web is a Web of action-able information -- information derived from data through a semantic theory for interpreting the symbols.” Shadboltin määritelmästä on löydettävissä se tärkeä seikka, että informaation perusteella voidaan tehdä jotain, ja toisaalta se, että informaatiota löydetään datasta tulkinnan avulla.

2.4.3. Tieto

Tieto (eng. knowledge) on moniselitteinen käsite, jonka tutkimusta on usealla tieteenalalla ja jota voidaan tutkia eri yhteyksissä useilla eri tavoilla. Sen tarkempi määrittely ei tässä yhteydessä ole mielekästä. Ackoff tarkoittaa tiedolla sellaisia asioita kuin ulkoa opeteltujen kertolaskutaulukoiden osaamista.

Semanttisen verkon tieto on Terzin ja muiden (2003) mukaan kohdealueen sanasto (eng. vocabulary) ja tähän sanastoon liittyvät säännöt. Tällaisia sanastoja käsittelen myöhemmin. Verkkokauppainformaatiosta kuluttaja voi oppia tuotteen hinnan ja tietää tämän jälkeen, mitä pankkitilille tapahtuu, jos hän tuotteen ostaa – hän osaa siis sanaston ja ymmärtää siihen liittyvät säännöt tämän toimialan suhteen.

2.4.4. Ymmärrys ja viisaus

Ymmärrys (eng. understanding) ja viisaus (eng. wisdom) ovat Ackoffin portaiden ylimmät askeleet. Ymmärtääkö kuluttaja hinnan suhteessa kaikkeen siihen hintainformaatioon, jota hän on saanut elämänsä aikana, ja osaako hän tehdä vieläkin laajemman ymmärryksen perusteella viisaan ostopäätöksen? Tämä on jo laajempi kokonaisuus.

Tekoäly (eng. Artificial Intelligence, AI) on se osa-alue, joka tutkii tietokoneella luotavaa älyä. Termi on otettu käyttöön jo vuonna 1956 (McCorduck, 2004). Semanttisen verkon osalta tässä tutkielmassa ei tutkita tarkemmin tekoälyä hyödyntäviä sovelluksia vaan keskitytään siihen pohjatyöhön, joka mahdollistaa tällaisten sovellusten luomisen. Mutta juuri tämän vuoksi semanttista verkkoa luodaan: se antaa tietokonesovelluksille mahdollisuuden toimia älykkäämmin ja luo pohjan tällaisten sovellusten toiminnalle.

2.5. Semanttinen rikkaus

Tässä tutkielmassa on kyse semanttisen verkon, yhden rajallisen tietojenkäsittelyn sovellusalueen, kehityksestä ja tämänhetkisestä tilasta sekä avoimesta datasta. Tiedon portaat on erinomainen tapa jäsentää semanttisen verkon käsitettä sen sisältämän semanttisen datan kautta. Vaikka se ei yksi yhteen sitä kuvaakaan, niin yllä kuvattu yhteys ja viittaukset semanttisen datan tutkimukseen kertovat tiedon portaiden ja semanttisen datan jakamien käsitteiden läheisyydestä. Semanttisen verkon keskeinen motivaatio on, että tietokoneohjelmista pyritään tekemään entistä älykkäämpiä. Niiden pitäisi siis ymmärtää asioita verkotetussa maailmassa ja jopa viisastua ymmärtämästään. Näkökulma semanttiseen verkkoon on usein teknis-orientoitunut, eikä tämä tutkielma pysty eikä pyri välttämään tätä tarkastelukulmaa. On hyvä siis muistaa ja muistuttaa, että semanttinen verkko pohjaa tutkimukseen ja käsitteisiin, jotka ovat olleet olemassa jo kauan ennen internetiä.

Tiedon portaita kiivetessä datan merkitys syvenee. Erityisesti dataa käyttävät sovellukset hyötyvät, jos ihmisen tietämystä voidaan siirtää osaksi dataa. Liitettäessä dataan aina syvempiä merkityksiä sen semanttinen rikkaus (eng. semantic richness) kasvaa. Käytän semanttisen rikkauden käsitettä tässä tutkielmassa, sillä se mielestäni kuvaa hyvin portaittain tapahtuvaa tiedon jalostamista. Semanttinen rikkaus käsitteenä ei ole vakiintunut, mutta sitä on käyttänyt julkaisuissaan ainakin Sabou (2006) ja Knublauch (2004).

Datan laadun parantaminen luo sille lisäarvoa. Avoimen eri ideologioiden kohdalla eri toimijoilla on toisistaan poikkeavat syyt tuottaa arvoa yleiseen käyttöön, sillä jokin syy on oltava tällaiseen resursseja käyttävään toimintaan. Julkisen rahoituksen varassa toimivat tutkijat antavat tulokset yhteiseen käyttöön. Avoimen lähdekoodin sovelluksia tuottavat yritykset (sitä tuotetaan paljon myös puhtaasti yhteisöllisissä yhteenliittymissä) tuottavat myös lisäpalveluita ilmaisen sovelluksen ympärille. On erittäin olennaista kysyä, kenen intresseissä on tuottaa raakadataan lisäarvoa ja miksi. Tähän kysymykseen palaan tutkielman lopulla, kun on selvillä, missä määrin lisäarvoa on avoimeen dataan tuotettu.

3. Semanttinen data

Tilastolliset taulukot ovat pitkälle jäsennettyä dataa ja lakikirjat vähemmän jäsennettyä (teknisessä mielessä, toisesta näkökulmasta voisi argumentoida aivan päinvastaista), mutta molemmat ovat ymmärrettävissä niin, että ne sisältävät meille merkityksellistä informaatiota tietyssä muodossa. Taulukko-ohjelmissa on usein sarakkeiden ja rivien otsakkeet kertomassa, miten solussa oleva data pitäisi lukea. Kirjoissa sisällysluettelo kertoo, mitä kirja sisältää. Sovellusten taustalla olevissa tietokantaratkaisuissa on sovelluksen toiminnan vuoksi tarpeellisia kuvauksia eli skeemoja. Kuvaukset ovat usein metatietoa, jota liitetään datan yhteyteen sen merkityksen esittämiseksi. Semanttinen data on sellaista dataa, joka sisältää ennalta sovittujen kaltaisia metakuvauksia. Tässä ja seuraavissa luvuissa lähdetään liikkeelle semanttisesti köyhästä datasta ja löydetään keinot tuottaa siitä yhä rikkaampaa semanttista dataa ja lopulta luvussa 4 tästä datasta luodaan semanttinen verkko. Mutta ennen sitä liitetään dataan merkityksiä.

Semanttinen data kuvailee jotakin kohdetta. Se kertoo kohteestaan jotain sellaista, joka olennaisesti liittyy siihen. Esimerkiksi hevosella on sellainen ominaisuus kuin väri. Kun me sanomme, että hevonen on ruskea, se kertoo jotain kyseisestä hevosesta, siis kuvailee sitä. Semanttinen data on periaatteessa mitä hyvänsä dataa, jota on lisätty kuvailtavan kohteen yhteyteen erillisenä datana tai kuvattavana olevan sisällön yhteyteen lomittain. Verkkosivulle usein liitetään semanttista dataa sen sisällön kanssa lomittain, jotta dokumentin merkitys avautuu esimerkiksi hakukoneille.

Tällaista dataa datasta kutsutaan metatiedoksi. Tiedon portaiden käsitteistöä käyttämällä semanttinen data sijoittuu portailla siten, että semanttisella datalla täydennetty raakadata on portaikon tieto-askelmalla. Se on jotain enemmän kuin raakadataa tai informaatiota, mutta kuvailu ei tee datasta itsestään älykästä. Älykkyys syntyy agenttisovelluksissa, jotka dataa käyttävät, tai vasta ihmisissä, jotka agenttien löydöksiä arvioivat vielä laajemmassa kontekstissa.

Semanttinen data liitetään perinteisesti datan semanttisen mallintamisen (eng. semantic data model) yhteyteen. Jokainen maailmaa mallintava sovellus joutuu kuvaamaan datan sovelluksen tasolla ja liittämään siihen merkityksiä. Tietokannat ovat tästä ilmeinen esimerkki. Kuvausinformaatio on tietokantamalleissa datasta erillinen kuvaus. Tämän tutkielman kohteena on internetistä löytyvä semanttinen data, jossa kuvaus on kiinteämpi osa dataa tai tavoitteena oleva data.

Tässä luvussa käydään läpi semanttisen merkityksen liittäminen dataan. Osana tätä merkitysten antamista kuvataan keinot, joilla tuotetaan linkkejä data-alkioiden ja tietokantojen välillä. Tällaisten linkitysten muodostaman semanttisen verkon esittelen tarkemmin luvussa 4. Teen tässä yhteydessä eron käyttöyhteydessään kuvaillun datan ja sellaisen datan välillä, joka kuvataan, jotta sitä voidaan käyttää sen ulkopuolella määritellyissä tarkoituksissa.

3.1. Sisällön ymmärtäminen

Miksi sisällön ymmärtäminen on tärkeää, ja kuka tai mikä sitä yrittää ymmärtää? Voidaan sanoa, että ihminen ymmärtää tietoa oppimisprosessin kautta ja oppii jonkun päämäärän saavuttamiseksi. Semanttisella datalla pyritään antamaan mahdollisuus sellaisten sovellusten tekemiseksi, jotka tehokkaasti avustavat ihmistä löytämään tarvitsemansa informaation. Tällaisia sovelluksia sanotaan agenteiksi, sillä ne toimivat käyttäjän aloitteesta jonkin päämäärän saavuttamiseksi, mutta ovat toimiessaan laajasti autonomisia. Nykyisin tiedon hakeminen toimii pääosin hakukoneiden avulla, jollaisia ovat mm. Google ja Bing. Hakukoneet eivät ole erityisen älykkäitä, ne esittävät hakualgoritmiensa mukaisia hakutuloksia, joiden pitäisi vastata annettuja hakuehtoja. Semanttisen datan pitäisi mahdollistaa myös semanttista informaatiota ymmärtävät hakukoneet, ja voi olla, että näissä käyttötapauksissa saadaan ensimmäisenä hyöty semanttisesta datasta.

Semanttista dataa hyödyntävät agentit eivät ole kuitenkaan hakukoneiden korvikkeita. Vielä enemmän datasta saadaan hyötyä, jos agentit kykenevät päättelemään sisällöstä asioita, jotka saavat ne antamaan älykkäämpiä vastauksia esitettyihin kysymyksiin. Semanttisen verkon agenttien toiminnassa on kyse tiedonhausta ja sääntöjen sovittamisesta löydettyyn dataan.

3.2. Tiedonhakuesimerkki

Kuvitellaan seuraava yksinkertaistettu, mutta tavallinen tiedonhaku internetissä: Verkon käyttäjä on ostamassa hevosta ja etsii itselleen sopivaa kohdetta. Hän selaa myynti-ilmoituksia internetissä olevassa verkkopalvelussa. Hän löytää kiinnostavan, sopivan hintaisen ja muilta ominaisuuksiltaan halutun hevosen ilmoituksen. Hevosen tietojen yhteydessä on mainittu sen nimi. Nimen perusteella käyttäjä avaa selaimessaan yhteisöllisen verkkosivuston ja hakee sieltä hevosen tiedot – ja saa nähtäväkseen mm. kuvia hevosesta ja lain vaatiman ja hevoselle annetun rekisteröintitunnuksen. Tämän jälkeen hän hakee selaimellaan lain velvoittaman rekisterin ylläpitäjän sivulta tietoja hevosesta sen rekisterinumeron perusteella, jolloin hänelle selviää mm. hevosen kantakirjaukset.

Hevosen hankkimisesta kiinnostunut henkilö käytti esimerkkitapauksessa hyväkseen kolmea tietovarastoa: myynti-ilmoituksia, yhteisöllisen sivuston keräämää tietokantaa ja lain määräämää virallista hevosrekisteriä. Jokainen näistä oli verkkopalvelu, eikä esimerkiksi painettu luettelo tai tiedustelu puhelimitse joltakin asiakaspalvelijalta. Henkilöllä on ollut käytettävissään sellaista tietoa ja ymmärrystä, joka on auttanut häntä selvittämään ilmoituksen hevosesta lisätietoja: ymmärrys siitä, että hevonen on löydettävissä kahdesta muusta datavarastosta, toisaalta nimen perusteella ja toisaalta hän on ymmärtänyt rekisteritunnuksen merkityksen hevosen identifioijana. Hän on tiennyt, mistä verkko-osoitteesta hän käy hakemassa tiedot, tai on käyttänyt hakukonetta. Kuvitellaan sama näkymä tietokonesovellusten ja erityisesti hakukoneiden näkökulmasta.

Selain on sovellus, joka hakee myynti-ilmoituksen. Selain näyttää kirjaimia ja numeroita, ehkä kuvia näytöllä kuitenkaan ymmärtämättä, mitä merkityksiä myynti-ilmoituksessa on – selain toimii

siis datan tasolla. Samalla tavalla hakukone on hakenut sivun, tallentanut sen tietokantaan, ehkä etsinyt avainsanoja, mutta yhtä kaikki, indeksoi sivun hakutuloksiinsa sanojen perusteella. Sivun metatieto on voinut kertoa dokumentin kielen ja sitä kuvaavia asiasanoja. Tämä toistuu hakukoneen toimesta kaikkien kolmen esimerkkisivuston ja niistä löydettyjen dokumentin osalta. Hevosen nimi ei merkitse hakukoneelle niitä sanoja enempää, joista nimi muodostuu. Merkkijono "rekisteritunnus" ei sano mitään. Jokaisen kolmen palvelun hevonen on hakukoneelle eri hevonen, se ei osaa yhdistää niitä toisiinsa. Hakutuloksista ihminen löytää merkityksiä, joita tietokone ei niissä ymmärrä olevan, ja pystyy siksi yhdistämään hakuosumia toisiinsa. Voidaan tietysti luoda älykkyyttä sovelluksiin niin, että ne ymmärtäisivät paremmin kontekstia, jossa tieto esiintyy, mutta tämä ratkaisu lisää jotain sovellus- eikä datatasolle. Semanttisella datalla tuotetaan lisäarvoa datalle. Jos halutaan aidosti parempia hakutuloksia, on tietoa kuvattava paremmin ja annettava päättelyyn kykenevien sovelluksien hakea sieltä informaatiota. Tällainen sovellus voi olla hakukone tai agenttisovellus.

Agentti ymmärtää merkityksiä toisin kuin selain. Hyvä agenttisovellus kykenisi tekemään oikealla tavalla kuvatun datan perusteella samat päätelmät kuin ihminen ja esittämään halutut tiedot ilman, että käyttäjä niitä hakee tai edes agentilta pyytää. Tarpeeksi älykäs agenttisovellus ymmärtää myös, ettei käyttäjä ole pelkästään ostamassa hevosta, vaan tarvitsee esimerkiksi hevosen omistamiseen liittyvän tallipaikan jostain palvelua tarjoavasta yrityksestä. Data on kuitenkin kuvattava agentin ymmärtämään muotoon, ja sen on ymmärrettävä kuvattujen käsitteiden välillä olevia riippuvuuksia. Tarvitaan metatietoa.

3.3. Metatieto

Metatieto kertoo jotain kohteesta, jota sillä kuvataan. Kohde voi olla digitaalinen dokumentti, kuten verkkosivu, mp3-tiedosto, tekstidokumentti tai digitaalinen valokuva; se voi olla myös ihminen, kirja kirjaston hyllyssä tai tietokannan taulun merkitysten kuvailu. Metatieto määritellään usein lyhyesti niin, että se on dataa datasta. Se, millaista tämä dataa kuvaava data on, selitetään metatietomallissa. Tasoja on siis kolme, 1) kohde, 2) data jolla siihen liitetään informaatiota ja 3) säännöt, joiden mukaan tämä informaatio luodaan. Aivan kuin metatieto kertoo jotain datasta, metatietomalli kertoo jotain metatiedosta. Metatietomallilla saavutetaan se hyöty, että tagittamiselle tai sitä vastaavalle metatiedon keräämiselle sovitaan yhteiset säännöt. Jos ja kun tällaiset säännöt johtavat hyvin jäsennettyyn mallin perusteella tehtyyn konkreettiseen metatietokuvaukseen, seuraa siitä edelleen, että metatiedon käsittely mahdollistuu myös tietokonesovelluksille – ne osaavat nyt lukea tiedon, ja eri sovellukset ymmärtävät sen yhdessä sovitulla tavalla.

Useat metatietomallit ovat tuttuja kaikille. Tällä hetkellä näkyvin internetissä oleva tapa kiinnittää metatietoa on tagittaminen, jolla kuvataan usein yhtä verkkosivua tai sen sisältöä. Jokainen sovellusohjelmoija tuntee tietokannan skeema-käsitteen, jolla kuvataan tietokannan

sisältö. Ne kuvausmenetelmät, joita tässä tutkielmassa käsitellään, ovat tavallista tagittamista tai tietokantaskeemoja tarkempaan ja laajempaan käyttöön tehtyjä määrityksiä. Näiden kahden ääripään väliin jää malleja, joissa pyritään löytämään kompromisseja molemmista maailmoista. Esimerkki laadukkaasta metatietomallista on RDF-määrittely. Samoja hyötyjä etsiviä kevyempiä malleja ovat semanttinen tagittaminen ja mikroformaatit.

Edellä on kuvattu hevosen ostotapahtumaan liittyvä haaste tiedon hyödyntämisessä, johon semanttisella datalla ja verkolla pyritään vastaamaan. Semanttisen verkon säännöt pyrkivät lopputulokseen, jossa selain tai hakukone ymmärtää, mistä tietystä hevosesta on kyse, ja löytää itsenäisesti sen tiedon hevosesta, jonka ihmisälyllä varustettu käyttäjä tietäisi siihen liittyvän. Ja paljon tätä enemmän: hyvä agenttisovellus löytää hevosesta käytyjä keskusteluita, ilmoituksia muilta markkinapaikoilta, vakuutustietoja, päättelee hevosen riskin joidenkin perinnöllisten sairauksien osalta, ja esittää muut mahdolliset myytävänä olevat hevoset, joista käyttäjä voisi olla kiinnostunut. Kaikki tämä tapahtuu ilman, että käyttäjän tarvitsee tietää tällaista informaatiota olevan olemassa. Tähän tavoitteeseen pääseminen vaatii kolme asiaa: hevoseen on kiinnitettävä semanttista dataa kaikkialla siellä, missä se internetissä esiintyy, semanttisen datan perusteella on voitava tehdä johtopäätöksiä, ja kolmanneksi hevosella on oltava sellainen semanttinen informaatio, jonka perusteella se on tunnistettavissa samaksi hevoseksi eri tietovarastoissa. Siis lyhyesti: on oltava olemassa rikasta semanttista dataa ja semanttinen verkko.

3.3.1. Metatieto tietokantamalleissa

Tietokoneohjelmista puhuttaessa data on tietueita tallennettuna jonkun taustajärjestelmän avulla: useimmiten riveinä tietokantaan tietokantasovelluksen avulla tai tiedostoina tiedostojärjestelmään. Käytännössä tietokantasovellukset ovat sovellusten pääasiallinen tallennusjärjestelmä, ja näissä edelleen relaatiokannat ovat yleisempiä kuin oliotietokannat. Tietokanta voidaan hahmottaa usean taulukon muodostamana datavarastona. Tietokannan taulukoihin tallennetaan rivejä, joissa on yksi tai useampia sarakkeita. Yhteen sarakkeeseen menee yksi tietokantasovelluksen kannalta atominen informaatio. Tietokannassa olevalla informaatiolla on metatietoa: taulukolle on annettu nimi, sarakkeella on nimi ja solulla tietotyyppi. Metatietokuvausta, joka kuvaa tietokannan taulukot, niiden rakennetta ja niiden välisiä yhteyksiä, sanotaan tietokannan skeemaksi.

Skeeman esittämiseksi tarvitaan metatiedon tietomalli, ja tällainen malli on esimerkiksi ER-malli (eng. Entity-relationship model). Tietokantasovellus tarjoaa rajoitetusti toimintoja tällaiseen skeemaan pohjautuen. Skeemaan perustuen se antaa mahdollisuuden tallentaa, lukea, päivittää ja poistaa rivejä ja tehdä näihin hakuja. Hakutulos on taulukkomuotoista dataa ilman skeemaa. Hammer (1978) osoitti jo 1978 eron tietokannan loogisen rakenteen ja semanttisen rakenteen välillä. Hän myös esitti metatietomallin, jonka tarkoitus oli tuottaa lisää kuvauksia dataan niin, että se olisi paremmin hyödynnettävissä. Relaatiokantojen skeema ei siis ole riittävä ratkaisu kaikkiin metatietotarpeisiin.

Oliotietokannat mainitaan ensimmäisen kerran vuonna 1985 ainakin kolmessa eri lähteessä (Atwood, 1985; Derrett et al., 1985; Maier et al., 1985). Oliotietokannat esittävät datan oliomuodossa, ja se saa merkityksensä luokkamäärittelyissä, joita voi pitää oliotietokannan metatiedon tietomallina. Luokka on perusrakenne useissa ohjelmointiin liittyvissä kielissä kuten Java tai UML (UML itsessään on metatietomalli, jolla voidaan kuvata mm. Java-sovelluksen rakenne), ja niillä kuvataan jonkin kohteen ominaisuudet ja toiminnot. Tällainen määrittely sopii hyvin nykyään yleistyneen olioparadigmaan perustuvien ohjelmointikielien ohessa käytettäväksi. Oliotietokannat eivät ole silti yleistyneet, mutta esimerkiksi UML-kieli on.

Tietokantaskeema on yhden sovelluksen sisäiseen toimintaan tarkoitettu metatietokuvaus. Sitä voi hyödyntää tietokannan ulkopuolella, mutta vain sen käyttöön liittyen. Tietokantaskeemalla ei voi kuvata verkkodokumenttia tai muita sellaisia kohteita. Ei ole myöskään mitään yhteistä sääntöä sille, millä tavalla jokin tieto tallennetaan tietokantaan – kaksi täsmälleen samanlaista sovellusta voi toimia kahden erilaisen tietokantarakenteen varassa. Tietokannat on toisaalta suunniteltu lähinnä yhden sovelluksen käyttöön, vaikka pääsyn sellaiseen voi avata internetiin, mikä entisestään rajaa niiden yleisempää käyttöä esimerkiksi julkisen datan jakamiseksi.

3.3.2. Metatieto verkkosivulla

Usein tietokanta on verkkosovelluksen taustalla ja sen sisältämästä datasta luodaan verkkosivuja. Verkkosivu ei ole hyvin määritelty dokumentti vaan usein se sisältää navigaatiorakenteita, mainoksia ja muita osia varsinaisen sisällön lisäksi. Varsinainen sisältö on useimmiten artikkeli, uutinen, myynti-ilmoitus tai sosiaalisen median sovellus. Verkkosivulla tarkoitan tällaista keskeistä sisältöä, en samalle sivulle liitettyjä muita elementtejä. Verkkosivuja on olemassa valtava määrä, ja relevantin tiedon löytäminen on haastavaa. Hakupalvelut hakevat valtavia määriä sivuja, indeksoivat niitä sen sisällön perusteella ja antavat käyttöön hakusovelluksen, jolla voi etsiä sivuja, joiden sisällöstä on mahdollisesti kiinnostunut. Verkkosivut on toteutettu HTML-kielellä. HTML-muotoisesta dokumentista selain luo näytölle dokumenttia vastaavan graafisen esityksen. HTML-dokumentteihin voidaan löydettävyyden parantamiseksi lisätä metatietoja kuten kieli, asiasanoja ja kirjoittaja – tämä on osa HTML-dokumentin metatietomallia. Nämä eivät näy graafisessa esityksessä, mutta hakukoneet näkevät ne ja näitä tietoja käytetään hyväksi hakusovelluksissa hakutulosten relevanssin arvioimiseksi.

Myös HTML-dokumenttien tapauksessa on merkitys liitettävä dokumentin yhteyteen, jotta tietokoneohjelma voi ymmärtää datan merkityksen. Merkityksiä voi liittää toinen tietokoneohjelma jollain päättelylogiikalla (sovellusten luoma metatieto), tai voidaan käyttää ihmisen tietämystä ja antaa ihmisen liittää haluamansa tieto (ihmisten luoma metatieto). Valitaanpa kumpi vaihtoehto hyvänsä, on luotava yhteiset säännöt sille, miten tieto liitetään. Verkkosivujen metatietomalli antaa mahdollisuuden lisätä asiasanoja osaksi dokumenttia, mutta varsinkin sosiaalisessa internetissä on yleistynyt mahdollisuus kiinnittää tagittamalla asiasanoja dokumentin yhteyteen.

3.3.3. Tagittaminen

Tagittamisella tarkoitetaan sitä, että dokumentin sisällön yhteyteen liitetään asiasanoja, joiden avulla se löydetään helpommin. Asiasanojen lisäämisen tarkoitus on antaa mahdollisuus kohdistaa hakuehtoja sanoihin, jotka ovat dokumentin kannalta relevantteja tai niitä ei muuten dokumentista ole löydettävissä. Tagitettaessa metatieto kiinnitetään dokumenttiin sen rinnalla olevina sanoina tai lauseina, harvemmin erillisenä dokumenttina, tai metatieto on dokumentin omissa metatiedoissa erillisinä kenttinä. Tagittamisen yhteydessä on tunnistettava kaksi erilaista järjestelmää: hierarkkinen taksonomia ja verkon sisältöjen luokittelutarpeen synnyttämä ei-hierarkkinen folksonomia.

Taksonomia on vanha käsite, ja Linnén kasvi- ja eläinkunnan luokittelu on siitä tunnetuimpia esimerkkejä. Hess ja Kushmerick (2003) ovat antaneet esimerkin, jossa taksonomialla pyritään kiinnittämään semanttista metatietoa internetin resurssiin ja joka kuvaa hyvin vaikeuksia tuottaa semanttisesti tarkkaa dataa tällä menetelmällä. Folksonomia on uudempi termi, jota Andersonin (2007) mukaan käytti ensimmäisenä Thomas Vander Val blogissaan vuonna 2004. Folksonomia on puutteellinen semanttisen merkityksen tuottamisessa, kuten Mathesin (2004) aiheen käsittelystä käy ilmi. Tagittamisella voidaan nähdä yhteys semanttiseen verkkoon, ja esimerkiksi Xu (2006) pitää tagittamista askeleena kohti semanttista verkkoa. Usein näillä asiasanoilla on tarkoitus mahdollistaa relevanttien hakutulosten löytäminen suuresta tietomäärästä. Tällaisia sanalistoja voi määritellä, tai ne voivat syntyä sitä mukaa, kun tageja tuotetaan. Tagittaminen on ehkä ensimmäinen askel kohti metatiedon antamista verkkosisällöille, mutta semanttisen tiedon ja semanttisen verkon osalta siihen täytyy suhtautua kriittisesti. Se on lähinnä Web 2.0 -termi ja sosiaalisen internetin käsite.

Sanastoista tapahtuvassa tagittamisessa on se ongelma, ettei dokumentille annettu sana kerro tietokoneohjelmalle, mikä sanan merkitys on. Dokumenttiin liitetty sana ”Tolstoi” kertoo aiheyhteydestä tai muusta syystä valistuneelle lukijalle, että dokumentti liittyy venäläiseen kirjailijaan. Tietokonesovellus ei tätä ymmärrä, ja metatietomallien tarkoituksena onkin saada sovellukset ymmärtämään, että annetulla metatiedolla on merkitys `kirjailija`.

Semanttinen tagittaminen tarkoittaa sellaisen ”sanan” liittämistä dokumenttiin, jossa on merkitys mukana (Hedden 2008). Semanttinen tagittaminen ei tarkoita, että tagitetusta sisällöstä tulisi osa semanttista verkkoa. Semanttinen verkko koostuu semanttisista dokumenteista, jotka on tehty myöhemmin tässä tutkielmassa esiteltävien määrittelyiden mukaisesti.

Eräs tunnetuimmista ja käytetyimmistä semanttiseen tagittamiseen tehdyistä metatietosanastostandardeista on Dublin Core, jossa alun perin määritellään 15 yleistä dokumentin metatietoa, mm. kirjailija, teoksen nimi ja julkaisuvuosi. Metasanastostandardit määrittelevät, mitä tietoja kuvataan. Ne myös usein ohjeistavat, miten tiedot kuvataan. Dublin Core, jota käytetään mm. HTML-dokumenttien, kuvien ja muiden tiedostojen kuvailemiseen, voidaan käyttötarkoituksesta riippuen esittää myös semanttisen datan muodossa. Silloin käytetään semanttista tagittamista tarkoituksena tuottaa semanttisen verkon dokumentti dokumentin yhteyteen.

3.4. Määrittelyt

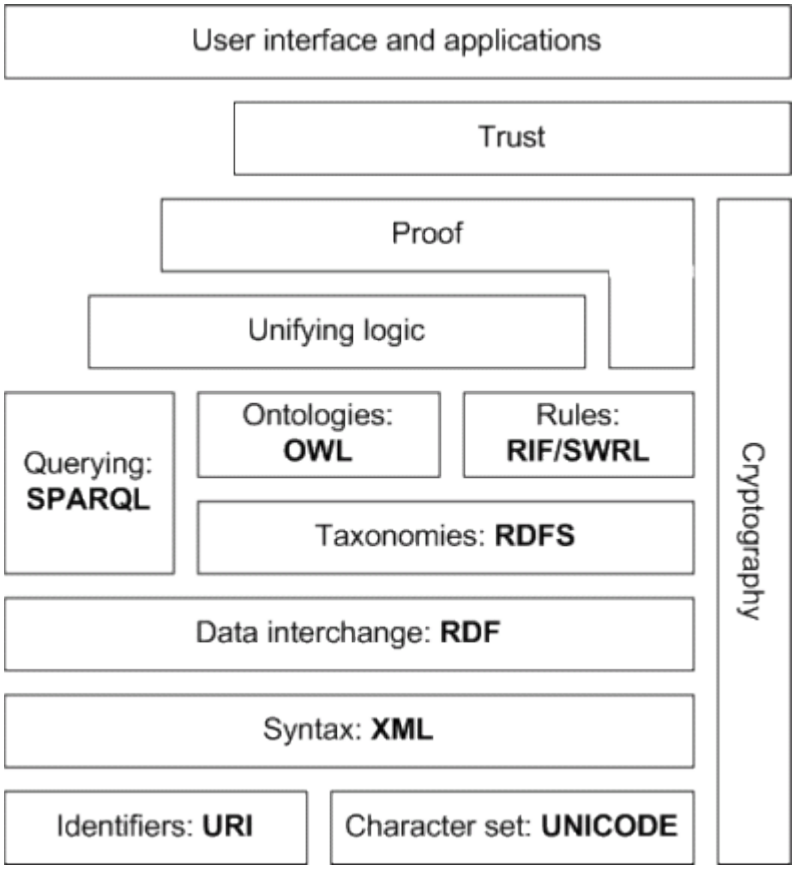
Semanttisen informaation kiinnittäminen voidaan toteuttaa monella tavalla, mutta on syytä määritellä metatietomalli. Näin tehtäessä mahdollistetaan metatiedon mahdollisimman laaja käyttö. Metatietomalleja määritellään erilaisissa yhteistyöjärjestöissä, joista Tim Berners-Leen johtama W3-organisaatio on keskeinen semanttisen verkon ja sitä edeltävien tekniikoiden osalta. W3:n missio on "... to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure the long-term growth of the Web" ("About W3", 2012).

Metatietomalleja on useita eri tarkoituksiin. W3:n tärkeimmät semanttisen verkon metatietomallit kuvataan RDF- (Resource Description Framework), RDFS- (Resourse Description Framework Schema), OWL- (Web Ontology Language) ja RDF/a-määrittelyissä. Lisäksi on olemassa mikroformaatit, jotka määritellään yhteisöllisenä projekteina ilman, että jokin taho keskitetysti johtaa määrittelyä. Semanttinen tagittaminen puolestaan ei ole yleistynyt käsitteenä semanttisen verkon yhteydessä, jossa puhutaan laajemmin kuvauksista ja ontologioista. Semanttisella datalla tässä tutkielmassa eri metatietomallien esittelyn jälkeen tutkitaan vain RDF:n ja sen päälle rakennettujen määrittelyiden mukaan tehtyjä laajempia metatietokuvauksia.

3.4.1. RDF

RDF on W3-standardi (Manola & Miller, 2004) ja sen ensimmäinen versio on vuodelta 1999. Siinä on kolme keskeistä osiota. Se on tietomalli: se kertoo, miten datan kuvataan. Se määrää, miten kaikille kuvauksen kohteille ja niissä käytetyille termeille annetaan yksikäsitteinen tunniste internetiin soveltuvin menetelmin. Siihen sisältyy myös tyyppisysteemin perusta, joka antaa mahdollisuuden tyypittää kaikki siinä käytetyt tiedon osat.

RDF-määrittely on semanttisen datan keskeisin standardi, ja siitä on johdettu täydentäen ja laajentaen aina uusien tarpeiden mukaan kehittyneempiä malleja. Ei ole liioittelua sanoa, että koko semanttisen verkon pohjana on RDF, ja se näkyy esimerkiksi semanttisen verkon teknologiapinossa, joka on esitetty kuvassa 1, jossa RDF on dataperusta, jonka päälle muut tekniikat sijoittuvat. Sen alle sijoittuvat määrittely kuvaavat RDF:n käyttämiä käsitteitä ja tekniikoita.



Kuva 1: Semanttisen verkon teknologiapino. Kuva muuttuu teknologioiden kehittyessä (”Semantic Web Stack,” 2013).

Semanttinen data semanttisen verkon yhteydessä tarkoittaa metatiedon esittämistä ns. RDF-triploilla, joiden tarkoitus on pilkkoa tieto erillisiksi palasiksi. Datan kuvauksen metamallina on RDF-määrittely, jossa luodaan säännöt sille, mitä tällaiset triplat ovat ja miten ne esitetään. Semanttinen data kuvataan siis toisin kuin skeemat relaatio- tai oliotietokannoissa, ja sen esitysmuoto on vastaavasti toinen – triplan abstrakti esitysmuoto on graafi. Uuden määrittelyn oikeutus ja syy ovat sellaiset tarpeet, joihin relaatiokantoihin perustuvat sovellukset eivät voi vastata. Toiseksi yhteinen määrittely antaa yhteisen tavan toimia, jonka päälle rakentuu laajemmin ymmärretty ja hyödynnettävissä oleva toteutus lopullisena tavoitteena olevasta semanttisesta verkosta. Tavoitteena on luoda sellainen internetin taso, jossa voidaan esittää tietokoneille ymmärrettäviä dokumentteja, joita voidaan linkittää toisiinsa ja joiden sisältöä kuvataan eri sovellusten luomissa dokumenteissa yhteisesti sovitulla tavalla. Rinnasteinen verkko olisi tietokantojen muodostama verkko. Tällaiseen tietokantasovelluksia ei ole suunniteltu, eivätkä ne tarjoa tällaiseen sovelluspalveluita. RDF sen sijaan määrittelee yhteisesti sovitun dokumenttirakenteen, joka toteuttaa tietovarastojen sovellusten käyttämän internetin ja joka pyrkii olemaan niin joustava, että sillä voitaisiin kuvata lähes mitä hyvänsä. RDF on määrittely, jolla on

alun perin haluttu kuvata internetistä löytyvä resurssi. Siis sellainen dokumentti, joka on haettavissa http-osoitteen osoittamasta sijainnista.

Otetaan RDF-määrittelyn rinnalle esimerkiksi Microsoftin toimistosovelluksilla tehtävä Excel-dokumentti, joka esittää taulukkomuotoista dataa. Yksinkertainen Excel-dokumentti sisältää metatietoa tiedostossa olevissa käyttäjälle näkymättömissä otsakkeissa. Näissä voi olla dokumentin kirjoittaja tai muita tietoja: usein juuri sellaisia, joita Dublic Core -standardi kuvaa. RDF kuvaa dokumentin metatiedot RDF-triploilla. Tripla-nimitys tulee siitä, että yksi atominen metatieto kuvataan kertomalla sen sisältö kolmella kentällä. Triplalla voidaan esimerkiksi sanoa, että hevosen Isla J Brave rekisterinumero on 88-2122. Tämä esitettäisiin triplalla

”Isla J Brave” ”rekisterinumero” ”88-2122”.

Tällaista merkintää kutsutaan sen osien merkityksen mukaan subjekti-predikaatti-objekti -triplaksi (SPO-tripla) tai joskus vain SPO:ksi. Subjektina on ”Isla J Brave”, predikaattina on ”rekisterinumero” ja objektina on ”88-2122”. Excel-dokumentin osalta RDF tripla voisi kertoa, että dokumentin ”Myynti-2001.xls” ”kirjoittaja” on ”Erkki Esimerkki”. RDF pilkkoo tällä tavalla tietämyksen sellaisiin palasiin, että palasella voidaan esittää mikä hyvänsä fakta ja että niillä on kuitenkin sellainen rakenne, että sovellukset voivat alkaa toimia niiden sisältämän tietämyksen perusteella. RDF-määrittelyn mukaisen metatiedon rikkaus ei perustu sen monimutkaisuuteen, vaan faktojen sisältämään tietoon.

RDF ei määrittele pelkästään SPO-rakennetta. Sen toinen keskeinen konsepti on triplojen osien esittäminen URI-osoitteina, joskaan tämä ei määritelmän mukaan ole pakollista. URI-osoite on mielivaltainen mutta tietyn syntaksin mukainen merkkijono, jonka tarkoitus on antaa yksilöllinen tunniste jollekin triplan osalle. URL-osoitteet, jotka ovat URI-määrittelyn mukaisia, ovat yleensä http-alkuisia osoitteita, joita seuraamalla löytää olemassa olevan resurssin internetistä. RDF-määrittelyssä erityisesti sanoudutaan irti tästä käytännöstä; tunnisteilla voidaan kuvata myös kohde jota ei ole internetissä, esimerkiksi tietty ihminen. Kun esimerkki triplamme esitetään rikkaamman informaation avulla, se voisi näyttää esimerkiksi alla esitetyltä. Tapana on erottaa URI-referenssi kulmasuluilla ja literaali lainausmerkeillä:

<http://heppa.hippos.fi/heppa/horse/FamilyInfo,desc_name.$DirectLink.sdirect?sp=l9491885661447240 75&sp=Shorse/FamilyInfo>

<http://heppa.hippos.fi/sanasto/rekisterinumero>

”88-2122”.

Http-alkuisiin URI-referensseihin palataan myöhemmin linkitetyn datan yhteydessä. Nyt riittää se huomio URI-osoitteista, että ne eivät välttämättä johda internetissä olevaan resurssiin, mutta

voivat niin tehdä. Seuraamalla esimerkissä subjektina olevaa http-osoitetta pääsee rekisterin pitäjän verkkosivulle, jossa on hevosen Isla J Brave rekisteritiedot. Olen valinnut sen tähän esimerkkinä siitä, että http-osoite voi johtaa johonkin mielekkääseen dokumenttiin. Predikaattina oleva tieto on myös http-osoite, mutta se ei osoita olemassa olevaan resurssiin vaan identifioi käsitteen, tässä tapauksessa käsite on ”rekisterinumero”. Jos käsite ”rekisterinumero” halutaan määritellä, niin linkki voi johtaa sen määrittelydokumenttiin.

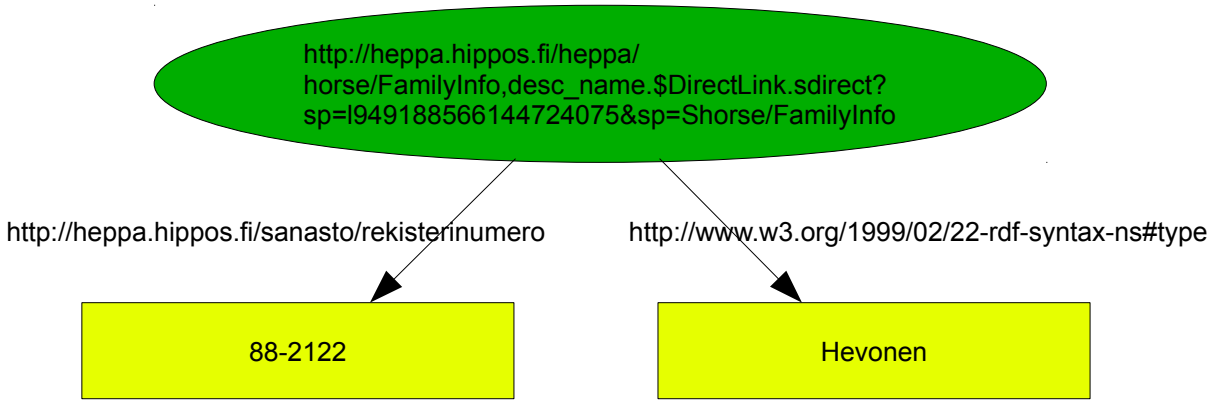
Kolmas RDF:n merkittävistä konsepteista on sen tyyppi-predikaatti (rdf:type). Tyyppi on yksi harvoista predikaateista, joita RDF määrittelee sen omassa sanastossaan. Tällä predikaatilla voidaan antaa subjektin, objektin tai predikaatin tyyppi. Se voi saada arvokseen jonkun RDF-määrittelyn mukaisen tyypin, mutta ennen kaikkea se voi olla mikä hyvänsä määritelty tyyppi. Tämä antaa mahdollisuuden laajentaa RDF-kuvaus käsittämään tarkkaan määriteltynä kaikki ne mahdolliset tyypit, joita ikinä jonkun kohdealueen triplojen luomiseksi tarvitaan. On puolestaan RDF-triploja käsittelevästä sovelluksesta kiinni, miten hyvin se pystyy toimimaan tyyppien perusteella. Edellä mainitun hevosen rekisterinumeron lisäksi voitaisiin määritellä siis hevosen tyyppi triplalla:

<http://heppa.hippos.fi/heppa/horse/FamilyInfo,desc_name.$DirectLink.sdirect?sp=l9491885661447240 75&sp=Shorse/FamilyInfo>

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

”Hevonen”.

RDF:n triplat ovat abstrakti käsite. Triplojen esitysmuotona on graafi. Isla J Braveen liittyviä faktoja Hippos ry:n rekisteritiedoista voidaan esittää graafilla, kuten kuvan 2 esimerkissä.



Kuva 2: RDF-graafi esimerkistä Isla J Brave.

RDF-mallilla ratkaistaan suurin osa metatiedon kuvaustarpeista. Triplat, URI-predikaatit ja rdf:type-predikaatti mahdollistaa hyvän kuvauksen kohdealueesta. Toisaalta se ei vielä selitä, millainen käsite on kirjailija, Excel-taulukko, rekisterinumero tai hevonen. Tarvitaan RDF-skeema

(RDFS), joka kertoo, miten esimerkiksi käsitteen hevonen voi esitellä. Esittelyt toteutetaan tekemällä luokkia ja kertomalla faktat näitä luokkia koskien.

3.4.2. RDFS

RDF:n hyvin alkeellisen sanaston laajentamisen mahdollistaa tätä tarkoitusta varten annettu RDF-skeema RDFS (Brickley & Guha, 2004). Kun jonkun sovellusalueen on kuvattava sen käyttämissä triploissa olevat tietotyypit, nämä voidaan määritellä RDFS:ssä annettujen sääntöjen mukaan. Yleisesti voidaan sanoa, että RDFS luo tyyppisysteemin RDF:lle. XML-kieli ja sen XML-skeema ovat RDF-kieltä ja RDFS-skeemaa vastaavat rakenteet, joiden erona on se, että XML-skeema rajoittaa XML-dokumentin rakenteen ja RDF-skeema antaa informaatiota väitteen tulkitsemiseksi (Brickley & Guha, 2004). RDF määrittelee tarkasti sen, millä tavalla tietämys esitetään. Se antaa jopa mahdollisuuden kertoa jonkin esityksessä käytetyn URI:n tyypin ja esittelee Property-määrittelyn predikaateille, mutta yhtä kaikki se antaa edelleen vain syntaksin datan esittämiselle, kuten taulu tai luokka antavat muissa referensseinä esitellyissä järjestelmissä. Tarvitaan semanttisesti rikkaampaa tietoa, jotta sovelluksiin voidaan ohjelmoida näennäistä älykkyyttä ja jotta ne ovat hyödyllisempiä käyttäjilleen.

RDFS käyttää triploja kuvatakseen triploja. Tällaiset rekursiiviset mallit ovat metakielissä (jotka kuvataan metametakielillä) tavallisia. RDFS määrittelee tietyt predikaatit, joilla voidaan kuvata triplojen ominaisuuksia. Näitä ominaisuuksia ovat mm. Resoure, Class, Literal, subClassOf, subProperty ja range. Monet skeeman konseptit ovat samankaltaisia kuin olio-ohjelmoinnin paradigmat: esimerkiksi olio, aliluokka ja (luokan-) ominaisuus – ne eivät kuitenkaan ole samoja. RDFS antaa toki mahdollisuuden sanoa, että Ori on Hevosen alikäsite – luokka, jos niin halutaan ajatella ja RDFS käyttää tätä termiä – mutta voidaan myös kertoa, että jos jokin ominaisuus on tyyppiä Nimi, on ominaisuudella aina myös tyyppi Etunimi ja Sukunimi.

Isla J Braven tapauksessa määriteltäisiin luokka Hevonen, joka olisi sen tyyppi. Jonkun tahon on tehtävä tällaiset määrittelyt käyttäen RDF- ja RDFS-työkaluja. Kun jokin perinnöllisiä sairauksia hevosilla tutkiva taho haluaa julkaista datan semanttisena datana, se voi käyttää näitä olemassa olevia käsitteitä ja laajentaa niitä RDFS:n mahdollistamin keinoin – ja tehdä kokonaan uusia. RDF-skeemat ovat yhteinen sopimus jonkun sovellusalueen käsitteistöstä. Kun sovelluskehittäjä toteuttaa ohjelman etsimään yksittäisen hevosen sukusiitoshistoriasta aiheutuvia geneettisiä riskejä sairastua tiettyihin perinnöllisiin sairauksiin, hänellä on hyvin kuvattu data, jota ohjelma käyttää. Sovellus tuottaa sellaista tietoa, jota ei ollut olemassa ennen datan kuvaamista ja jota siitä olisi erittäin hankala löytää ilman tällaista vaihetta. Datan kuvaaminen ei toki ole pieni haaste sekään.

3.4.3. OWL

Verkon ontologiakieli OWL on tarkoitettu rikkaan ja monimutkaisen tietämyksen esittämiseksi asioista, joukoista asioita ja näiden välisistä relaatioista. OWL on logiikkaan perustuva kieli, jota

päättelyyn erikoistuneet sovellukset kykenevät hyödyntämään varmentaakseen asioiden oikean tilan tai löytääkseen uutta tietoa päättelyprosessin tuloksena (Hitzler et al., 2012).

Semanttinen data on tarkoitettu päättelyä tekeville agenttisovelluksille. On muistettava, että mikään tietokonesovellus ei itsessään ole älykäs, vaan se ymmärtää maailmasta juuri niin paljon kuin se on opetettu ymmärtämään. Maailmassa asioiden välillä on monimutkaisia suhteita, paljon haastavampia kuin RDFS voi kuvata. Siksi tarvitaan yhä pidemmälle meneviä kuvauskieliä semanttisen informaation antamiseksi dataan, jotta loogiseen päättelyyn kykenevät ohjelmat voivat sovittaa sääntöjä dataan ja löytää sieltä parempia vastauksia. Semanttista dataa voi hyvin tuottaa ilman ontologiakieltä ja kuvata hyvin kaikki ne luokat ja näiden väliset suhteet, joita kohteena olevasta datasta on löydettävissä. Tämä on jo huikea edistysaskel taulukoihin, relaatiokantoihin ja tagittamiseen verrattuna. Tällaiset ontologiat ovat kuitenkin tästä seuraava tapa tuottaa jo tietämystä hipovaa informaatiota dataan. Ontologiat käsitellään omassa luvussaan, sillä yksi määrittely ei enää kykene kuvaamaan, mistä niissä on kyse.

3.4.4. Mikroformaatit (μF), mikrodata ja RDF/a

Mikroformaatit ovat kolmas tapa tuottaa metatietoa dokumentin yhteyteen. Mikroformaatit antavat tavan tuottaa dokumentin omalla rakenteella siihen kuvailevaa metatietoa. Esimerkiksi verkkosivustot on tehty HTML-kuvauskielellä, joka itsessään on jo metakieli. Koska HTML-dokumentti on laajennettavissa, voidaan sitä laajentaa semanttisella kuvailulla. Mikroformaatit ovat tällainen laajennus. Mikroformaatit eroavat tagittamisesta ja metatietotietueista siten, että ne ovat lomittainen osa dokumenttia eikä erillinen tieto dokumentin ohessa tai itsenäinen dokumentti. Mikroformaatit ovat yhteisöllinen projekti, joka toimii verkkosivustolla microformats.org. Suhteessa W3:n semanttisen datan määrityksiin mikroformaatit ovat huomattavasti helpompi tapa tuottaa metatietoa HTML-dokumentin osaksi, ja se on jopa saanut hieman huomiota selainvalmistajilta, jotka rajoitetusti tukevat tällaisia kuvauksia dokumenteissa.

Mikrodata on suurten hakukoneiden yhteinen projekti verkkosivustojen metatiedon parantamiseksi. Niiden schema.org-sivustolla määrittämät skeemat ovat sellaisia, että paitsi hakukoneet, myös muutamat selaimet tukevat niiden käyttämistä. Mikroformaatit ovat edeltäneet mikrodatamallia, mutta toisin kuin mikroformaatit, mikrodatan skeemat on määritetty myös RDF-kielellä.

RDF/a on W3-määrittely (Adida et al., 2012). Toisin kuin kaksi edellistä, sillä on tarkoitus kuvata myös muita resursseja kuin HTML-dokumentteja. RDF/a-määrittelystä on myös suppeampi versio, jota voidaan verrata mikrodataan. Koska W3:n standardit saavat hakukoneilta, sovelluksilta ja sosiaalisen median palveluilta paremman hyväksynnän, tämä todennäköisesti tulee olemaan voittaja näistä kolmesta tekniikasta.

Kaikki kolme määrittelyä ovat olemassa kahdesta hyvästä syystä: metatiedon tuottaminen skeemoineen erillisenä dokumenttina on haastavaa. HTML-dokumentti on myös oma erikoistunut

dokumenttimuotonsa, jossa on valmiina sellainen rakenteinen rakenne, jonne informaatiota on mahdollista lisätä. Erityisesti RDF/a-määrittelyn mukaan tehty verkkosivu on muunnettavissa semanttisen verkon dokumentiksi – verkkosivuksi, jolla on sitä kuvaava laadukas metainformaatio mukana. RDF/a ei riko semanttisen kuvausinformaation mallia, jota W3 muuten ajaa eteenpäin toisissa määrittelyprojekteissa.

3.5. Ontologiat

Ontologia on käsitteellistämisen määrittely (Gruber, 1995). Gruber toteaa aiheellisesti, että ontologia-käsite herättää kiistoja. Hänen mukaansa se kertoo, miten me kuvaamme käsitteitä ja niiden välisiä yhteyksiä agenttien toimintaympäristössä. Agenttien ymmärtämät dokumentit ovat semanttisen verkon keskeisin tavoite, ja semanttiseen verkkoon tuotetut Gruberin kuvausta vastaavat ontologiat ovat semanttisen verkon keskeisiä rakenteita. Yksi ontologia kuvaa annetulla kielellä yksittäisen kohdealueen merkitykset. Ontologia tässä suppeassa muodossaan kertoo, millainen dokumentti syntyy, kun data ja siihen liittyvä tietämys kuvataan.

Ontologioita on monilla eri tieteen osa-alueilla, semanttisesta verkosta puhuttaessa on kyse web-ontologioista. Siis ontologiakuvauksista, joilla on tarkoitus tuottaa lisäarvoa verkossa olevien tietovarastojen metakuvauksiin. OWL on yksi tällainen kuvaus, jolla tuotetaan lisäarvoa RDF-tietueisiin.

3.5.1. Laajennettu RDF-skeema

Skeeman yhteydessä olemme kuvanneet sanastojen (vocabulary) luomista. Ontologia ja sanasto ovat sellaiset käsitteet, että niiden käytöstä ei ole yhtenäistä linjaa. Tässä tutkielmassa sanastolla tarkoitetaan sellaista kohdealueen kuvausta, joka tapahtuu RDF-skeeman mahdollistamilla työkaluilla; se siis sisältää lähinnä kohdealueen keskeisten käsitteiden luokittelua. Ontologia liittää mukaan sellaisia tietoja, jotka luovat käsitteiden välille sääntöjä, joita sovellukset voivat käyttää tuottaessaan vastauksia kohdealuetta koskeviin kysymyksiin. Ontologioissa on myös sanastoa joka luodaan ja laajennetaan RDFS:n antamilla työkaluilla.

Qin (2001) tekee eron selväksi esitellessään yhden tietovarannon sanaston laajentamisen ontologiaksi: ”Compared to the original semantic model of GEM controlled vocabulary, the major difference between the two models lies in the values added through deeper semantics in describing digital objects, both conceptually and relationally.” Qin kirjoittaa syvemmästä semantiikasta; itse käytän semanttisen rikkauden käsitettä.

3.5.2. Sanasto vs. ontologia

Sanaston ja ontologian käsitteiden suhde on sellainen, että ontologiassa annetaan sanasto, jolla jokin kohde voidaan kuvata, ja arvot, joita kuvauksissa käytetään. Semanttisen verkon pinossa (ks. kuva

1) ontologiataso rakennetaan RDF- ja RDFS-määrittelyiden päälle OWL-kielen avulla. Siksi tässä yhteydessä voidaan myös esittää OWL niin, että sen ontologiasanasto on laajennus RDF- ja RDFS-sanastoihin. Todellisuudessa OWL on itsenäisempi kokonaisuus, vaikka se usein esitetään ja mielletään kuvatun kaltaiseksi laajennukseksi.

DL-kielet (Description Logic) tuovat formaaliin loogiseen päättelyyn liittyvän semanttisen informaation tietämyksenhallintajärjestelmiin. OWL on yksi tällainen järjestelmä, joka toteuttaa osan DL-kielestä. Tällainen loogisen päättelyn formalisoinnin esittely on oma tieteenalansa ja tämän tutkielman ulkopuolella. OWL-kielen osalta on mielenkiintoista huomata, että siinä sekoittuvat RDF- ja RDFS-määrittelyt ja DL-kielen tutkimus. OWL sisältää kaksi määrittelyä: OWL DL ja OWL FULL. Myöhempi OWL 2 esittelee lisää alimäärittelyitä. Kun RDF ja RDFS ovat hyvin suoraviivaisia ja yksinkertaisia toteutukseltaan ja kelpaavat siten pohjaksi kaikille OWL-variaatioille, on loogisen päättelyn alueella jo huomattavasti suurempi määrä erilaisia toteutuksia. OWL-variaatioita tarvitaan, kun päättelysovellusten rajoitteet täytyy dokumentoida yhteisiksi sopimuksiksi.

3.5.3. Suljettu maailma dokumentissa

Ontologiakuvauksen perusteella tehty dokumentti on eräänlainen itsenäinen artifakti: se sisältää säännöt ja mahdollisesti datan jostakin tietämysalueesta. Tunnettuja ja käytetyimpiä ontologiakuvauksia ovat Dublin Core ja FOAF (Friend Of A Friend). Dublin Coren sovellusaluetta ovat erilaiset elektroniset dokumentit, joiden sisältöä sillä kuvaillaan. Se mahdollistaa agenteille paremmat tiedot dokumenteista ja niiden välisistä suhteista. FOAF kuvaa henkilöitä ja näiden välisiä suhteita. Se mahdollistaa hajautetun sosiaalisen verkoston syntymisen semanttiseen verkkoon. OWL-dokumenteista onkin erotettava kaksi osa-aluetta: ensinnäkin sovellusalueen ominaisuudet, luokat ja tietotyypit, ja näiden ominaisuudet ja riippuvuudet toisistaan, ja toiseksi niissä annetaan kohdealueen oliot. Näille kahdelle ontologian eri puoliskolle ei ole omia termejään, vaikka kyseessä on selvästi käytännön ontologiaesimerkkien perusteella kaksi eri osa aluetta.

Usein nämä dokumentit erotetaan omiksi kokonaisuuksikseen selkeyden vuoksi, ja toisaalta käytännön syistä usein ontologiakuvaus on erillinen dokumentti, jonka perusteella on tarkoitettu, että muut luovat instanssit, joihin ontologiakuvauksen säännöt voidaan sovittaa. Kun molemmat osat ontologiasta kuvataan yhdessä dokumentissa, on käytäntönä, että ensin kuvataan ontologian säännöt ja näiden jälkeen kuvataan instanssit. Ehkä voitaisiin puhua ontologiasanastosta ja ontologiamaailmasta. Ero voitaisiin kuvata esimerkiksi niin, että sanasto antaa asioita, joita maailmassa voi olla. Säännöt kertovat, miten nämä potentiaalisesti vuorovaikuttavat keskenään. Mutta simulaatio maailmasta on olemassa vasta, kun instanssit on määritelty ja ne hakevat paikkansa maailmassa ja alkavat vuorovaikuttaa sen sääntöjen mukaan.

3.5.4. Päättelykyselyt

Miten tietokoneohjelma voi ymmärtää asioita? Handschuh (2007) kuvaa ontologioiden ja agenttien välistä yhteyttä ja näiden toimintaan liittyviä prosesseja. Jotta saataisiin täysi hyöty RDF:n kaltaisilla kuvauksilla täydennetyistä tietovarastoista, voidaan lisätä tiedon yhteyteen sääntöjä siitä, mitä johtopäätöksiä voidaan tehdä tiedon sisällöstä. RDF:n säännöin ja skeemoin voidaan antaa sisällöstä tarkkaan määritelty metatietokuvaus, mutta ei ole ennalta saneltua, mitä yksittäinen tietokoneohjelma sillä tekee. Tarkastellaan esimerkkinä rekisteriä hevosista. Voidaan kuvata tarkkaan se, että hevosen käsite sisältää käsitteet isä ja emä tai jälkeläiset – ja millainen käsite on kyseessä. Ihminen muodostaa tietämyksensä perusteella heti käsityksen siitä, miten nämä yksittäiset tiedot voi käyttää löytääkseen yksittäisen hevosen sukupuun myös sen isovanhempiin asti.

Ontologia kuvaa termejä ja niiden välisiä suhteita. OWL-ontologia määrittelee sen, miten suhteita voidaan kuvata, se ei kuvaa yksittäisen tietovarannon suhteita. OWL on siis abstraktio tai metakuvaus, joka antaa ohjeet, millä tavoin voidaan kuvata päättelysääntöjä. Kun tietokoneohjelma lukee tällä tavalla tehdyn kuvauksen, se voi tietää tai olla tietämättä, mitä se luetuilla säännöillä tekee. Mutta se voi myös käsitellä säännöt, sillä se osaa jäsentää ne. OWL:n tapauksessa säännöt muodostetaan RDF-määrittelyn mukaan triploina ja käyttäen hyväksi RDF- ja RDFS-määrittelyitä.

3.5.5. Laajennettavuus

Ontologiakuvaukset ovat itsessään laajennuksia RDF- ja RDFS-määrittelyistä. Siitä seuraa suoraan, että ontologiakuvausten tekniikka mahdollistaa olemassa olevien ontologioiden käyttämisen osana uutta ontologiaa (aggregointi) ja niiden laajentamisen (eng. extension). Tämä tarkoittaa sitä, että uusi ontologia kartuttaa ontologiakuvausten joukkoa ja sillä tavalla jäsentää edelleen sitä, millaista dataa internetissä on tarjolla. Suomessa yleisiä ja avoimeen käyttöön tarkoitettuja ontologiakuvauksia on tehty FinnONTO-projektissa (Hyvönen et al., 2007).

Ontologioiden luomiseen on olemassa työkaluja. Se, mitä ontologian pitäisi sisältää, on kokonaan toinen asia. Kuka määrittelee keskeiset käsitteet jonkun kohdealueen osalta? Kuka määrittelee esimerkiksi suomalaisen hevosaiheisen keskeisen sanaston? Yleinen suomalainen asiasanasto sisältää hyvin suppean sanaston tästä aihepiiristä: se tuntee sanan hevonen mutta ei esimerkiksi sanaa ori.

4. Semanttinen verkko

Semanttinen verkko syntyy dokumenteissa olevista linkityksistä dokumenttiin itseensä tai toisiin dokumentteihin. Semanttisesta verkosta tietoa hakevat agentit voivat seurata näitä linkityksiä tavoitteidensa saavuttamiseksi. Agenttien liikkumiseen verkossa ja sen käyttämiseksi tarvitaan verkkoon sovelluskerroksia. Tarvitaan myös staattisten dokumenttien lisäksi dynaamiset dokumentit mahdollistavat määrittelyt.

Semanttisen verkon luomiseksi tarvitaan paitsi yhteisiä sopimuksia myös niistä johdettuja dokumentteja, jotka luovat konkreettisen semanttisen verkon, ja työkaluja, joilla nämä dokumentit tuotetaan. Edelleen tarvitaan työkaluja dokumenttien lukemiseksi ja sovelluksia, jotka dokumenttien perusteella tuottavat hyötyä käyttäjälleen. Edellä on kuvattu keskeisimmät yhteiset sopimukset, joiden varaan semanttinen verkko rakentuu: RDF-, RDFS- ja OWL-määrittelyt. Tässä luvussa keskitytään näiden määrittelyiden pohjalta syntyneisiin työkalutarpeisiin, joita tarvitaan semanttisen verkon dokumenttien tuottamiseksi. Semanttisen verkon dokumentilla (SVD) tarkoitetaan sellaista tiedostoa, joka sisältää RDF-määrittelyn mukaisen graafin tekstimuodossa. Dokumentit voivat olla staattisia tai dynaamisia, jolloin niitä tuotetaan semanttisiin dokumentteihin erikoistuneiden tietokantasovellusten tai verkkosovellusten avulla. Työkalut, joihin lasketaan valmiiden loppukäyttäjälle annettavien sovellusten lisäksi myös sovellusohjelmoijan käyttämät sovelluskirjastot, ovat olennaisia, kun teoriasta luodaan käytännön sovelluksia. Semanttisen verkon idea tarvitsee semanttisen verkon toteuttamiseksi työkalut. Laajasti tarkasteltuna myös olemassa olevat skeemat ja ontologiakuvaukset ovat työkaluja: ne mahdollistavat jonkin datan kuvailemisen valmiilla sanastolla, tai niitä laajentamalla saadaan oma sanasto aikaan nopeammin.

4.1. Linkitetty data

Semanttinen verkko muodostuu semanttisen verkon dokumenteissa olevista linkeistä toisiin semanttisen verkon dokumentteihin. Tämä verkko ei eroa HTML-dokumenttien välisen navigaation toteuttavista linkeistä. Linkkien muoto on molemmissa IRI-määrittelyn (Internationalized Resource Identifier) mukainen. Neljä kirjainyhdistelmää (IRI, URI, URL ja URN) esiintyy kuvattaessa internetin dokumenttien välisiä linkityksiä, ja ne on tässä syytä selvittää, sillä semanttisen verkon dokumentit voivat sisältää linkin näköisiä arvoja triploissa ilman, että nämä ovat toisiin dokumentteihin johtavia linkkejä.

4.1.1. IRI, URI, URL ja URN

Kolme lyhennettä, jotka viittaavat internetin dokumenttien väliseen linkitykseen. ovat IRI, URI ja URL. Neljäs lyhenne, URN, on tarkoitettu kuvaamaan resurssia, joka ei ole internetistä ladattavissa. Toisilleen rinnakkaiset käsitteet ovat URL (uniform resource locator) ja URN (uniform resource

name). Erona näillä kahdella on se, että kun ensin mainittu kuvaa resurssin, jonka voi hakea internetistä seuraamalla annettua osoitetta, jälkimmäinen kuvaa resurssin, jota ei voi hakea internetistä. URL-osoitteet ovat useimmiten http-alkuisia. Semanttisen verkon määrittelyssä, sen sijaan että olisi käytetty URN-osoitteita kuvaamaan resursseja, jotka eivät ole olemassa verkossa, käytetään http-skeemalla alkavia URL-osoitteita näissäkin tapauksissa. Tämä aiheuttaa hämminkiä. Kuuluuko http-osoitteen RDF-triplassa olla oikea dokumentti verkossa vai ei? Vastaus on, ettei tarvitse, mutta se voi olla. URI, joka on kattokäsite URL ja URN-osoitteille, ei määrää, että jonkin tietyn skeeman (http, ftp, file) mukainen osoite osoittaa oikeaan dokumenttiin, vaikka käytännön myötä sellainen tulkinta on syntynyt.

IRI-osoite (Internationalized Resource Identifiers) on myöhempi, vuonna 2005 hyväksytty määrittely URI-osoitteista, joka sallii erikoismerkkien (esimerkiksi skandinaavisten kirjainten) käyttämisen ISO-10646-merkistöstä. IRI-osoitteet ovat URI-osoitteiden sijasta käytössä jo OWL-dokumenteissa ja kaikissa tulevissa määrittelyissä.

Semanttisen verkon dokumenteissa on URI-osoitteita, jotka http-skeeman käytön vuoksi sekoitetaan URL-osoitteisiin. RDF-määrittelyssä on kuvattu URI-osoitteiden käyttöä ja lisätty liite, jossa selvennetään entisestään osoitteiden käyttöä RDF:ssä. Selvennyksen voi tiivistää seuraavasti: osoitteet muistuttavat verkkodokumenttien linkkejä, mutta niitä ei tule ymmärtää sellaisina eikä olettaa, että niiden samankaltaisuus heijastaisi käsitteellisiä yhtäläisyyksiä. Poikkeuksena siis se, että ne voivat olla URL-osoitteita. Se, mitä osoitetta seuraamalla tapahtuu, on selvitettävissä ainoastaan kokeilemalla.

Jos semanttisessa dokumentissa olevan URI-muotoisen osoitteen osoittamasta internetin palvelusta saa resurssin, on oletettava, että myös se on semanttisen verkon dokumentti. On olemassa käytäntö siitä, voidaanko URI:ssa olevasta valinnaisesta risuaitamerkillä ”#” erotetusta tiedon osan kuvaavasta tunnisteesta vielä tehdä johtopäätöksiä sen suhteen, onko haluttu tieto osa haettua dokumenttia. URI voi siis viitata myös semanttisen verkon dokumentin rakenteeseen, vaikka sellaista päätelmää alkuperäisessä URI-dokumentissa varoitetaan tekemästä. Näin semanttinen verkko määrittelee uudestaan käsitteitä itselleen paremmin sopivaksi. Tässä tapauksessa mahdollistetaan viittaaminen dokumentin aligraafiin koko dokumentin sijasta.

Yksittäinen hevonen ei ole internet-dokumentti, mutta sille voidaan antaa yksikäsitteinen URI, esimerkiksi http://hippos.fi/hevonen/#123456. Hevosen emän käsite voidaan ilmaista literaalin ”emä” sijasta Hippoksen pitämän kuvitteellisen käsitehakemiston URI:lla http://rdf.hippos.fi/sanasto/1.0/ema. Hevosen emän arvo on URI ja samalla URL sen emään, esimerkiksi arvo http://rdf.hippos.fi/hevonen/#234567. Tällöin samassa tietovarastossa data alkaa linkittyä keskenään. Hevoselle voidaan merkitä syntymäpaikka jostain maantieteellisen paikannimien hakemistosta, kuten suomalaisesta SUO-ontologiasta, jossa Tampereen arvo olisi http://www.yso.fi/onto/kunnat/k837. Tällöin muodostuu linkki ensimmäisestä semanttisesta tietokannasta toiseen ja semanttisen verkon perusrakenne on valmis.

4.1.2. Verkottuminen

Semanttisessa verkossa voidaan nähdä kolme linkittämisen tasoa: yhden dokumentin sisällä on linkityksiä, dokumenttiin voidaan linkittää ja dokumentista on linkityksiä muihin dokumentteihin. Tim Berners-Lee on määritellyt uuden termin: linkitetty data (Linked Data). Hän antaa tämän termin sisällön kolmella säännöllä:

- http-osoitteita käytetään kuvaamaan asioita, ei enää dokumenttien välisiä suhteita
- tuo http-osoite palauttaa verkosta dataa sen kuvaamasta kohteesta
- data sisältää lisää http-osoitteita, jotka linkittävät uusiin kohteisiin.

Näiden kolmen ehdon ei tarvitse olla voimassa samanaikaisesti. Data on linkitettyä dataa, kun jompikumpi jälkimmäisistä ehdoista toteutuu.

Tim Berners-Lee piti puheen Ted-konferenssissa vuonna 2009. Puheessaan hän pyysi avaamaan datan ja toisaalta kehotti ihmisiä vaatimaan dataa käyttöönsä. Hän ei puhunut tuossa lyhyessä (Ted-konferenssin esitykset ovat lyhyitä ja ne ovat aihettaan popularisoivia puheenvuoroja) esityksessä tekniikasta mitään, koska tekniikka on jo valmiina. RDF-määrittelyssä subjektille, predikaatille ja objektille voidaan antaa literaalisen arvon sijasta URI, joka viittaa toiseen dokumenttiin.

Rakenne on sama kuin hypertekstidokumenttien linkitys internetissä, mutta nyt vapaamuotoisten dokumenttien sijaan on linkitetty keskenään hyvin määriteltyä dataa. Vuonna 2009 Tim Berners-Lee tiesi, että tekniikka on valmiina, että dataa on olemassa valtavia määriä, ja että esteet ovat käytännön toimissa ja asenteissa. Asenteista hyvä esimerkki Suomesta on vuonna 2012 käyty lyhyt julkinen keskustelu maanmittauslaitoksen osalta. Valtiovarainministeriö vastusti karttojen avaamista julkiseksi ilmaiseksi avoimeksi dataksi, mutta hallitusohjelmaan kirjattu linjaus ”julkisin varoin tuotettuja tietovarantoja avataan kansalaisten ja yritysten käyttöön” voitti. Kyseessä on mielestäni merkittävä linjaus, joka antaa kansalaisille mahdollisuuden vaatia yhteiskunnan keräämää dataa ja hyödyntää sitä sekä yleishyödyllisissä että kaupallisissa palveluissa.

4.2. Semanttisen verkon dokumentti

Sellaiset termit kuin Web1.0, Web2.0 ja Web3.0 ovat vaikeasti määriteltävissä. O'Reilly (2007), Millard ja Ross (2006) ja Murugesan (2007) yrittävät kuvata käsitteitä Web1.0 ja Web2.0, ja heiltä kaikilta löytyy saman suuntaisia ajatuksia sen sisällöstä. He näkevät erojen syntyvän staattisen ja dynaamisen eroihin, yhteisöllisyyteen ja sisällön jakamiseen liittyen. Keinotekoista ja jälkikäteen määriteltyä termiä Web1.0 ei olisi olemassa ilman, että on ollut tarpeen esitellä termi Web2.0. Semanttisen datan kontekstissa määrittelyiden epämääräisyys tulee esiin. Web1.0 viittaa usein internetin ensimmäiseen vaiheeseen, jossa staattiset dokumentit linkittyivät keskenään. Web2.0 mahdollistui sovellusten tuottaessa dynaamisia dokumentteja, jotka luodaan jonkun datavaraston

päälle. Web2.0 on dataorientoitunut internet. Semanttinen verkko on toisinaan (Hendler, 2009) nimetty internetin kolmanneksi askeleeksi, siinä mielessä Web3.0 kuvaa semanttista verkkoa. Semanttisen verkon dokumentit voivat kuitenkin olla staattisia tai dynaamisia, ja vaikka niitä voidaan luoda yhteisöllisesti, se ei ole niille leimallista. Semanttinen verkko on tarkoitettu agenttien käyttöön, ja se tarjoaa sekä staattisia että dynaamisia dokumentteja, jotka ovat enemmän tai vähemmän täydellisiä ontologioita jostakin kohteesta. Semanttisen verkon dokumentin voi dynaamisten tekniikoiden vuoksi nähdä Web2.0-dokumentin kehittyneempänä muotona, mutta semanttinen verkko ei ole jatkoa Web2.0-internetille eikä pohjaudu sen keskeisille tekniikoille.

RDF-muotoiset semanttisen verkon dokumentit sisältävät tripla-graafeja johonkin sovelluskäytön mahdollistamaan esitysmuotoon tallennettuna. Esitysmuodosta käytetään usein nimitystä koodaus: annettu abstrakti esitys, kuten tässä tapauksessa graafi, koodataan tekstimuotoiseksi esitykseksi eli dokumentiksi. Triplat voidaan kirjoittaa dokumentiksi käyttäen XML-syntaksia, ja määrittely nimeltään RFD/XML on keskeisin RDF-dokumentin esitysmuoto. OWL-määrittely, joka sisältää RDF-triplat datan esitysmuotona, antaa mahdollisuuden kirjoittaa SVD:n usealla eri syntaksilla, mutta OWL/XML-syntaksi on ainoa toteutuksilta vaadittu muoto – muut ovat valinnaisia koodauksia. Olennaista näissä eri syntaksien mukaan kirjoitetuissa dokumenteissa on se, että dokumentin sisältö ja merkitys eivät muutu dokumenttien välillä: ainoastaan luku- ja kirjoitustapa muuttuu. Kaikki esitysmuodot, joita RDF- ja OWL-dokumenteissa määritellään, ovat tekstimuotoisia; siksi RDF- ja OWL-dokumentit voidaan tuottaa tavallisella tekstieditorilla ja ne ovat myös ihmiselle luettavassa muodossa sellaisenaan.

Kooltaan pienen SVD:n tekemiseksi ei siis tarvita enempää kuin ymmärrys dokumentin sisällöstä ja ne säännöt, joilla sisältö kirjoitetaan – sekä valinnainen tekstieditori. Tärkeämpää tietysti on, että tarkka syntaksi antaa mahdollisuuden tehdä tietokoneohjelma, joka osaa lukea ja jäsentää dokumentin sisällön ja jolla voidaan kirjoittaa tavallista tekstieditoria erikoistuneempi ohjelma tietyn muotoisen dokumentin kirjoittamiseksi. XML-syntaksi on abstrakti määritelmä, jolle löytyy työkaluja kaikille tarpeellisille ohjelmointikielille. On olemassa paljon XML-editoreita, jotka avustavat XML-dokumenttien kirjoittamisessa; Altovan XMLSpy on esimerkki tällaisesta työkalusta. XML-syntaksilla tuotetusta dokumentista käytetään joskus nimitystä sovellus: siis RDF:n ja OWL:n XML-syntaksin mukaiset dokumentit ovat XML-sovelluksia. Useille tällaisille sovelluksille on olemassa yleisimmillä ohjelmointikielillä tehtyjä kirjastoja, jotka osaavat jäsentää geneeristä XML-jäsennintä tarkemmin juuri tietyn XML-sovelluksen sisältöä. Toisaalta sovellukset, kuten mainittu Altovan XMLSpy, avustavat myös RDF-dokumenttien kirjoituksessa. Ihminen voi kirjoittaa näitä dokumentteja siis useilla eri tavoilla ja eri ohjelmat avustavat siinä. RDF-editoreita ei juurikaan ole, sillä XML-editorit täyttävät tämän tarpeen, tai triploja ei kirjoiteta dokumenteiksi, kuten triplavarastojen yhteydessä myöhemmin käy ilmi. OWL-dokumentti sisältää enemmän merkityksiä ja OWL-määrittely antaa enemmän sääntöjä ontologioiden rakentamiseksi. Sovellukset tarvitsevat kirjastoja dokumenttien käsittelyyn. Java-ohjelmointikielelle tehty OWL API (Horridge

2009) on yksi tällainen kirjasto, jolla voidaan lukea ja kirjoittaa ainakin XML-muotoisia RDF- ja OWL-dokumentteja.

Lähtökohtaisesti semanttiset dokumentit ovat saatavilla internetissä. Tavoite on antaa verkossa tietoa etsiville agenttisovelluksille mahdollisuus löytää tietoa dokumenteista – sovellukset, jotka sitä dokumenteiksi tallentavat, ovat välttämättömiä, mutta eivät ensisijaisia tavoitteisiin nähden. Jotta agentit löytävät tiedon verkossa, täytyy olla olemassa sellainen verkkopalvelu, jota ne voivat käyttää. Staattiset dokumentit ovat haettavissa sellaisenaan, ja niihin osoitetaan URL-osoitteilla. Myös dynaamiset dokumentit haetaan URL-osoitteista. Haettaessa dynaamisia SVD:ja täytyy kyselykielen lisäksi olla yhteinen sopimus siitä, millä tavalla agentit tekevät kyselyitä internetissä sovellukselle, jota kutsun semanttisen verkon solmuksi. Ensin esittelen kuitenkin triplavaraston, sovelluksen, jollaisen avulla semanttisen verkon dokumentit tallennettaan.

4.3. Triplavarasto

Triplavarasto on tietokantamainen tallennusjärjestelmä triploille. Triplavarasto ei tallenna triploja dokumenttimuodossa vaan triplojen käsittelyn kannalta järkevässä muodossa. Triplavarasto toteuttaa jonkun sellaisen sovellusrajapinnan, että triploja voidaan käsitellä mielekkäällä tavalla, ja ilmeisin esimerkki tästä on, että se kykenee tuottamaan tallennetuista triploista RDF-dokumentin. Monet verkkosivut tuotetaan käyttäen käyttäjältä tulleita parametreja verkkosivun tuottavassa sovelluksessa, hakemalla rajattua tietoa tietokannasta ja esittämällä vain relevantti informaatio lopputuotoksena. Myös RDF-muotoiset dokumentit voidaan tuottaa ohjelmallisesti rajaten luotavan dokumentin sisältämiä triploja.

Triplavaraston toteuttamiseksi tarvitaan jokin tietokantaratkaisu, jolla triplat talletetaan. RDF-dokumenttia ei ole suunniteltu tehokkuusnäkökulmasta: tieto ei ole sellaisessa muodossa, että sitä hyödyntävät algoritmit olisivat läheskään aina tehokkaita. Relaatiotietokantojen palvelut voitaisiin toteuttaa taulukkomuotoisia tekstitiedostoja käyttäen, mutta tämä ei ole niiden tarkoitus tai se tarve, johon ne ovat syntyneet. Semanttisilla tietokannoilla haetaan samoja etuja, joita saadaan relaatiokannoilla: tuotetut dokumentit ovat dynaamisia, hakukieli on tarkoituksenmukainen, mukaan voidaan liittää tietoturvaominaisuuksia ja niiden toiminta skaalautuu tallennetun tiedon ja metatiedon määrän kasvaessa. Listaa voisi jatkaa, mutta jo näiden esimerkkien perusteella on selvää, että staattinen dokumentti ei voi olla ainoa tapa tallentaa ja käsitellä triploja ja ontologioita verkossa.

Triplavaraston sisäisen toiminnallisuuden toteuttaminen on vapaasti päätettävissä, sitä eivät sido mitkään määrittelyt, ja triplavarastosovelluksen toimittajat tekevät aina oman toteutuksensa. Olennaista on, että tieto ei ole enää yksi dokumentti eikä triplavarastosovellus välttämättä edes tallenna tietoa triploina ymmärrettävässä muodossa. Ei ole mitenkään itsestään selvää, että triplavaraston tuottama dokumentti olisi validi RDF-dokumentti. Yhtä lailla se voisi tarjota vain

taulukkomuotoisen dokumentin. Itse asiassa useat RDF-triplavarastot on rakennettu konventionaalisten relaatiokantojen päälle niin, että taulukkomuotoisesta datasta tehdään muunnos RDF-dokumenttiksi. Osa toteutuksista on tehty vain triploja varten, mutta niidenkin sisäinen toteutus ja tietorakenteet ovat sovelluksen toimittajan päätettävissä. Tarkoituksenmukaista on, että triplavarastosovelluksen lopullinen tuotos on RDF-dokumentti kuten relaatiotietokantojen tuotos on taulukko. Kuten relaatiokantoja, myös triplavarastoja käytetään useimmiten kyselykielen eikä sovellusrajapinnan kautta. Triplavarastojen kyselykielenä on SPARQL.

4.3.1. SPARQL

SPARQL-kyselykieli on RDF-triplavarastojen hakukyselyissä käytetty kieli. Se ei kerro, millä tavalla triplavarasto pitäisi toteuttaa eikä edes sitä, millainen sovellus ottaa vastaan kyselyitä ja palauttaa tuloksia niiden perusteella. SPARQL määrää, miten kyselylause on muodostettava. Relaatiotietokannat perustuvat vakiintuneeseen tieteelliseen tutkimukseen ja useiden vuosikymmenten käytännön kokemuksiin. Relaatiotietokantatoteutuksista tunnetaan yleisesti ainakin MySQL, Postgres, MSSQL, IBM DB2, Terabase, Sybase ja Oraclen relaatiotietokantatoteutukset. Niiden kaikkien keskeinen kyselykieli on SQL, johon usein tehdään toteutuskohtaisia laajennuksia. Niiltä odotetaan, että ne toteuttaisivat kokonaisuudessaan vähintään SQL-määrittelyn mukaiset ominaisuudet, missä ne epäonnistuvat joiltain osin. SPARQL on määrittely, jonka triplavarastojen tekijät pyrkivät toteuttamaan, mutta jossa jo lähtökohtaisesti osa ominaisuuksista on määrätty valinnaisesti toteutettaviksi.

SPARQL-kyselykieli (rekursiivinen lyhenne englanninkielen sanoista SPARQL Protocol and RDF Query Language) on W3-määrittely. Määrittelydokumentin ensimmäinen versio on julkaistu vuonna 2004, ja vuonna 2008 siitä on tullut voimaan virallinen suositus. Semanttisen verkon aikajanalla se ajoittuu siis selvästi RDF- ja OWL-standardien julkaisun jälkeen alkaneeksi määrittelytyöksi. SPARQL on semanttisen verkon teknologiapinossa asetettu RDF-määrittelyn päälle, RDFS- ja OWL-kerrosten viereen näiden korkuiseksi. Tällä asettelulla pyritään kertomaan, että se on kyselykieli, jonka kohteena ovat RDF-triplat ja että se toteuttaa jotain sellaista, jonka ilmaisuvoima on vähemmän kuin agenteille tarjotut loogisen päättelyn kuvaukset tai muu laajempi agenttien tekemä päättely. Kyselykielenä se sisältää kuitenkin tällaisille palveluille tyypillisiä kyselyihin kirjoitettavia hakuehtoja. Sen luonteesta kertoo myös hakulauseen (CONSTRUCT) lisäksi määrittelystä löytyvä kysymyslause (ASK), jolla voidaan saada vastauksia yksinkertaisiin päättelyä vaativiin kysymyksiin.

Segaran ja muut (2009) esittävät semanttisen verkon sovellusohjelmointia käsittelevässä kirjassaan esimerkkikoodin muistiin ladattavasta triplavarastosovelluksesta. He esittelevät mm. hakusovelluksen, jonka koodi on erittäin lyhyt, noin 60 riviä, ja heidän oman arvionsa mukaan se kykenisi käsittelemään kymmenien tuhansien triplojen muodostamaa triplavarastoa (triple store). Heidän käyttämänsä kyselykieli on samankaltainen SPARQL-kielen kanssa, minkä he myös itse

toteavat. Sen jälkeen kun W3 määritteli semanttisen verkon triplat, syntyi näiden käsittelyä varten useita erilaisia ja eri lähestymistavan omaavia ratkaisuja. Osassa ratkaisuja on SQL-kielen kaltainen luonnollisen kyselykielen rakenne, osa ratkaisuista käyttää RDF/XML-muotoiseen dokumenttiin sopivaa xpath-kyselykieltä. Haase ja muut (2004) vertasivat kuutta kyselykieltä toisiinsa vuonna 2004, siis samana vuonna kun SPARQL-määrittelyn ensimmäinen versio julkaistiin kommentoitavaksi. Kaikki hänen tekemänsä huomiot ovat edustettuna SPARQL-kielessä:

- Tarvitaan yksi yhteisesti hyväksytty ja jaettu määrittely.
- Kyselykielenä voidaan käyttää SQL:n kaltaista luonnollista hakukieltä.
- RDF-skeema tulee olla huomioituna.
- Kyselykielessä täytyy olla mukana ryhmittely ja aggregaatit, ja sen täytyy tukea osittaisia hakuosumia.

Muitakin tavoitteita on listattu sekä Haasen toimesta että ”RDF Data Access Use Cases and Requirements” -dokumentissa vuodelta 2005, jossa SPARQL:lle asetettavia vaatimuksia hahmotellaan. Suuri osa SPARQL:n tavoitteita liittyy samoihin käsitteisiin kuin SQL-kielen tavoitteet, ja se onkin osin hämäävän samankaltainen. Se on kuitenkin olennaisesti erilainen kieli, sillä RDF-triplat ja niiden skeemat sisältävät sellaista informaatiota, jota relaatioskeemoissa ei ole. Lisäksi aivan olennainen ero on, että RDF on rekursiivinen tietorakenne (graafi), kun taas relaatiokantojen eräs heikkous on niiden kyvyttömyys käsitellä hierarkkista tietoa.

SPARQL-kyselykielen neljä perusoperaatiota ovat SELECT-, CONSTRUCT-, ASK- ja DESCRIBE-kyselyt. Rakenteina SELECT ja CONSTRUCT ovat samankaltaisia; SELECT palauttaa hakutuloksen taulukkomuodossa, kun CONSTRUCT-kyselyn tulos on RDF-graafi. ASK palauttaa esitettyyn kysymykseen kyllä- tai ei-vastauksen, ja DESCRIBE antaa tuloksen, jonka tarkoitus on kuvata tulosjoukkoa. SPARQL tukee neljää erilaista hakulausetta ja näissä käytettäviä hakulauseen ominaisuuksia – vain osa on toteutettuna yksittäisissä SPARQL-sovelluksissa.

4.3.2. Triplavarastojen rajoitukset

Kaikki triplavarastot eivät toteuta kaikkia SPARQL-määrittelyssä esiteltyjä ominaisuuksia. Tämä ei ole välttämätöntä määrittelyn vaatimusten täyttämiseksi: tavallista W3-määrittelyille on, että osa ominaisuuksista katsotaan pakollisiksi ja osa valinnaisiksi. Lisäksi toteutusten filosofian mukaan kaikki triplavarastot eivät toteuta SPARQL-kyselypintaa lainkaan, vaan saattavat tarjota muun sovelluskirjaston tarjoaman rajapinnan triplojen käsittelyyn. Koska SPARQL ei määrittele päivitysoperaatioita, se ei edes voi olla ainoa rajapinta triplojen käsittelyyn. SPARQL-sovellus on hyvin suppea toiminnoiltaan. Se sisältää ainoastaan sellaiset tietovarastoon tehtävät kyselyt, jotka etsivät ja lukevat tietoa. Se ei ole tämänkään ominaisuutensa mukaan rinnastettavissa esimerkiksi

SQL-kyselykieleen, joka sisältää kaikki tietoalkioihin kohdistuvat operaatiot (luonti-, luku-, päivitys- ja poisto-operaatiot). SPARQL-kyselyihin on SPARUL-laajennus joka sisältää nämäkin operaatiot ja josta on tullut tätä tutkielma kirjoittaessani joulukuussa 2012 W3-suositus – se on voitu toteuttaa olemassa oleviin sovelluksiin ainoastaan ehdotetun dokumentin perusteella. On hyvä huomata, että SQL- ja SPARQL-kielten eroon on selvä syy: SPARQL:n tarkoitus on tarjota korkean tason rajapinta semanttisen verkon työkaluksi, ei määritellä RDF-triplavarastosovellusta, jollainen jokaisen SPARQL-kyselyn taustalla on.

4.4. Semanttisen verkon solmu

SPARQL-kyselykielen toteuttavat sovellukset voivat mahdollistaa kyselyiden kohteena olevan semanttisen tietokannan teknisenä verkkopalveluna (SPARQL endpoint). Itse kutsun tällaista verkkopalvelua semanttisen verkon solmuksi. Verkkopalveluna se eroaa sellaisista palveluista kuin verkkokauppa tai kuluttajalle tarjottu pankkipalvelu siten, että se on sovelluksille suunnattu tekninen verkkopalvelu. Se eroaa relaatiokannoista niin, ettei relaatiotietokantoja ole koskaan tarkoitettukaan tällaiseen julkiseen ja avoimeen käyttöön, vaan lähinnä yksittäisten sovellusten suljetuksi taustajärjestelmäksi. Relaatiotietokantaa käyttää verkkosovellus, SPARQL-verkkopalvelua käyttää puolestaan tietoa etsivä agenttisovellus. Semanttisen verkon solmun sisältämä data sisältää sellaisia elementtejä, joiden avulla se voi linkittyä muihin semanttisen verkon solmuihin, mikä on relaatiokannoista kokonaan puuttuva suunnitteluperiaate.

Semanttisen verkon solmun tarkoitus on tuottaa rajapinta, jota agentit voivat käyttää osana sovelluksen tarkoituksen täyttämiseksi. SPARQL-verkkopalvelu on määritelty niin, että se tarjoaa määrätyssä http-osoitteessa ja kuvatulla protokollalla mahdollisuuden tehdä standardoidulla kyselykielellä kyselyitä triplavarastoon – ja saada RDF-dokumentti kyselyn vastauksena. Muita tällaisia teknisiä verkkopalveluita, jotka eivät siis ole semanttisen verkon palveluita, on yhä kasvavassa määrin internetissä, ja niitä toteutetaan eri tarkoituksiin erilaisin tekniikoin. Mainittakoon vaikka SOAP- tai REST-tekniikoin toteutetut palvelut esimerkkeinä tällaisista. Yleisellä tasolla semanttisen verkon SPARQL-solmut eivät eroa tällaisista palveluiden toteutuksista, mutta ne palvelevat vain kaikki yhtä tarkoitusta – semanttisen verkon olemassaoloa.

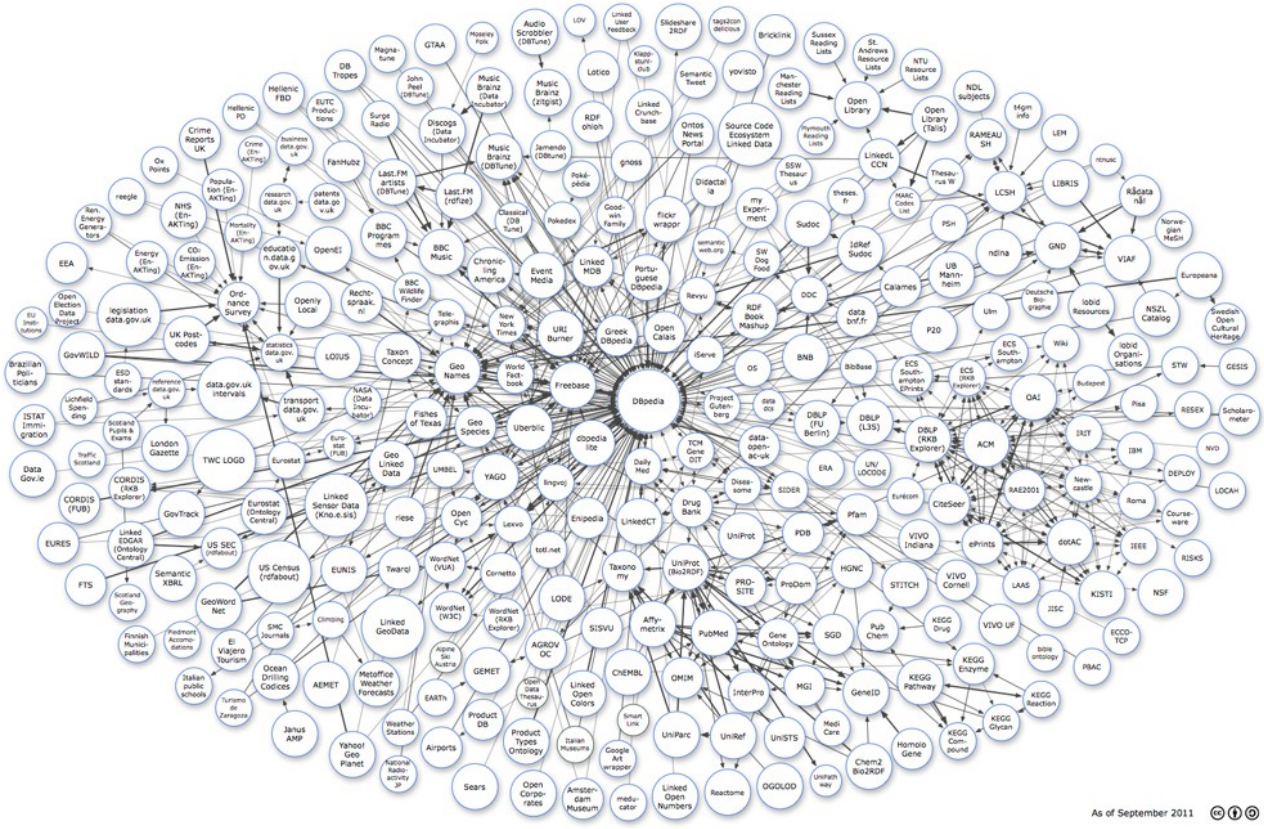
SPARQL tarkoittaa sekä edellä esiteltyä kyselykielen määrittelyä että määrittelyä kyselyiden tekemiseksi internetin yli (”protocol” lyhenteessä kuvaa termin tätä ulottuvuutta). W3 on tuottanut erillisen dokumentin sen määrittämiseksi, miten kyselyt tehdään internetin yli siten, että SPARQL-asiakassovellus voi tehdä ja saada vastauksia SPARQL-palvelinsovellukselta. Tämä määrittelydokumentin nimi on SPARQL Protocol for RDF, ja joskus protokollaosasta käytetään lyhennettä SPROT johtuen dokumentin nimestä. SPARQL-lyhenne on yleisemmin käytetty ja asiayhteydestä selviää, tarkoitetaanko kyselykieltä vai protokollaa. Internetin yli tehtävä kysely eroaa triplavaraston paikallisesta käsittelystä siten, että paikallisia sovellusrajapintoja ei ole

määritelty. Tällaisiakin määrittelyitä on olemassa, esimerkiksi XML-dokumentin käsittelyyn. SPARQL-protokollan mukaisia viestejä voidaan välittää sitomalla ne joko HTTP- tai SOAP-viestinvälitysprotokollilla. Yleistä mielenkiintoa tällä määrittelyllä ei ole kuin sovelluskehittäjille. Jos triplavaraston yhteyteen ei toteuteta teknisen verkkopalvelun toteuttavaa osaa, ne eivät voi toimia semanttisen verkon solmuina. Data voi olla toki muutoin tarjolla, mutta se on sitä yleisestä toteutuksesta poikkeavalla tavalla, mikä heikentää sen käytettävyyttä.

4.5. Julkaiseminen

Jos joku taho, vaikkapa Hippos ry, haluaisi tarttua Berners-Leen haasteeseen ja julkaista hevosrekisterinsä avoimena linkitettynä datana, niin kuinka se voi tämän tavoitteen saavuttaa? Sen olisi vähintään luotava RDF-dokumentti, jossa on sen rekisteriä vastaavat tiedot triploina. Jotta dokumentti olisi mitenkään mielekäs, sen yhteyteen on tuotettava skeema-dokumentti (RDFS-muodossa) kertomaan, miten dokumentti on tehty ja miten siinä annetut triplat ja niiden osat – subjekti, predikaatti ja objekti – pitäisi tulkita. Hippoksen tapauksessa nykyinen järjestelmä on sellainen, että tietovarastona käytetään relaatiotietokantaa. Tämän jälkeen on luotava datan käytölle säännöt, eli valittava sopivin lisenssi tai luotava oma. Sitten semanttisen verkon dokumentin voi julkaista internetiin haettavaksi jonkin http-osoitteen osoittamasta sijainnista. Myös skeemat ja muut metatietokuvaukset julkaistaan ja niihin osoitetaan linkittämällä ne dokumentin metatiedoissa.

Voi sanoa, ettei tehtävän määrä ole vähäinen ja että se sisältää hyvin erikoistuneita työvaiheita, joihin tarvitaan erityisosaamista. Vielä ei ole edes perustettu teknistä SPARQL-solmua tai tehty syvempää OWL-tasoista kuvausta kohdemaailmasta. Omalta osaltaan tämän tehtävän vaativuus selittänee, miksi semanttinen verkko mahtuu vielä Cyganiakin ja Jentzschin (2011) tekemässä vuoden 2011 tilannetta esittävässä kuvassa 3 yhteen kertasilmäyksellä luettavaan kuvaan, jossa ei ole pienten toimijoiden tuottamia semanttisen verkon tietokantoja. Olisi kuitenkin tärkeää, että myös pienet toimijat ja varsinkin yhteisölliset toimijat voisivat tuottaa semanttiseen verkkoon sisältöä.



Kuva 3: Semanttinen verkko vuonna 2011.

5. Avoimen datan laatu

Avoin data ei ole homogeeninen kokonaisuus. Se on käsite, jonka yleisimmän nykymerkityksen synty löytyy tiedeyhteisön parissa, mutta jota on alkanut värittää erilaiset avoimen ideologian ja hyvien hallintotapojen siihen tuomat merkitykset. Avoin data on toisaalta avoimen ideologian ympärille syntyneiden liikkeiden ja sitä kautta valtioiden ja näiden hallinnoimien globaaleiden instituutioiden keräämää ja julkaisemaa dataa. Toisaalta termi on vahvasti mukana siinä prosessissa, jossa W3:n työryhmissä tuotetaan metatietokuvausten määrittelyitä mille hyvänsä datalle – suljetulle tai avoimelle. On olemassa myös useita yhteisöllisiä projekteja, jotka perustuvat avoimen informaation ja tiedon tuottamiseen. Näistä tunnetuin on Wikipedia-projekti.

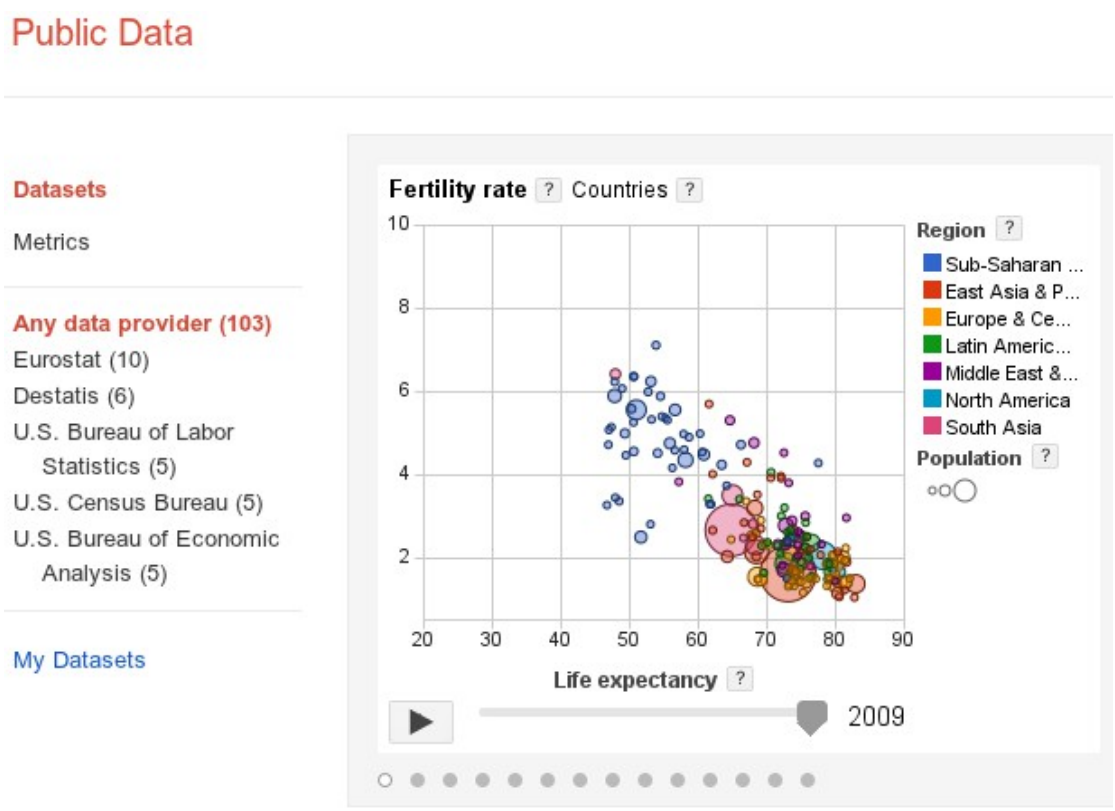
Samanaikaisesti on siis käynnissä useita erilaisia hankkeita. On olemassa paine avoimen datan määrän lisäämiseksi avaamalla julkisten instituutioiden suljettua dataa internetistä vapaasti saatavaksi. Samalla on olemassa tämän datan laadun parantamiseen tähtääviä aloitteita. Lisäksi yhteisöt jalostavat jo omaa dataansa, mistä erinomainen esimerkki on Wikipedian artikkelien jalostaminen Dbpedia-projektissa (Auer, 2007). Tässä luvussa pyrin vastaamaan kysymyksiin siitä, mihin haasteisiin laadun parantamisella yritetään vastata ja miten toteutuneiden datavarastojen laatua voidaan arvioida. Lopuksi esitän kysymyksen siitä, mikä on valtioiden ja vastaavien instituutioiden julkaiseman avoimen datan laatu. Vastaan kysymykseen laadusta arvioimalla datan olemassa olevalla mittarilla.

5.1. Hyödynnettävyyden haasteet

Datan hyödynnettävyys on aina käyttötapauskohtaista. Hans Rosling käyttää kuuluisiksi tulleissa esityksissään aikasarjoitettua tilastodataa ja muodostaa tästä kaksi- tai kolmiulotteisen esityksen, joista esimerkiksi eliniänodote ja lasten lukumäärä perheessä -esitys on ansainnut hänelle nimityksen ”tilastoguru”. Rosling käyttää esityksissään erityisesti YK:n keräämää dataa, joka on vapaasti käytettävissä. Ilman avointa lisenssiä esitykset olisi toki mahdollista järjestää, mutta data ei olisi yhtä helposti hyödynnettävissä kuin se nyt on.

Dataa käyttävät sovellukset olettavat, että data on jossakin järjestäytyneessä muodossa – se on informaatiota. On olemassa tiedostomuotoja, jotka on tarkoitettu yhden sovelluksen käyttöön; tällaisesta ehkä tunnetuin esimerkki on Microsoftin toimisto-ohjelmiston Excel-taulukko. Erityisesti tilastosovellukset käyttävät perustellusti omia tiedostomuotojaan, jotka mahdollistavat tehokkaan tilastollisen analyysin ja soveltuvat juuri sen yhden ohjelman tarpeisiin. Hyödynnettävyyden kannalta tässä on kuitenkin ilmeinen epäkohta, kun data ei ole ehkä lainkaan muiden sovellusten luettavissa teknisten tai lisenssiongelmien vuoksi. On olemassa toisia tiedostomuotoja, joiden tarkoitus on tehdä datasta yleiskäyttöisempää. Tällaisista muodosta CSV ja XML ovat tunnetuimmat. Esimerkiksi CSV-muodossa olevaa dataa hyödyntämällä voidaan luoda Google

Public Datan tapaisia palveluita, jotka visualisoivat dataan käyttäjän valitsemin ehdoin (ks. kuva 4). Tällaisessakin palvelussa dataa pitää kuvata, mutta data itsessään on helposti hyödynnettävissä.



Kuva 4: Google Public Data -visualisointityökalu.

Agentit ovat erityinen sovellusjoukko, jonka tarkoituksena on hyödyntää dataa. Jotta ne voivat hyödyntää sitä, niiden on voitava saada raakadatasta informaatiota sen merkityksestä ja muodostaa tästä jonkinasteista tietoa. XML-dokumentti ja siihen liitetty skeema antavat mahdollisuuden tuottaa kuvailtua dataa. Se, että RDF-graafi voidaan tallentaa XML-dokumenttina, ei tarkoita, että sen sisältö ymmärrettäisiin XML-määrittelyihin. RDF-dokumentin avaava sovellus muodostaa siitä sellaisen tietorakenteen, joka se todellisuudessa on: RDF-graafi. RDF-graafilla ja sen triplojen muodostamisen säännöillä on eroa verrattuna XML-dokumenttiin. Tätä eroa ovat verkkoartikkeleissa selittäneet niin Berners-Lee (1998), Sequeda (2012) ja Tauberer (2008) kuin Decker (2000) vertaillessaan myös sisällön rikkautta eri tekniikoilla. Heidän mukaansa asia on ymmärrettävissä niin, että XML määrittelee puun, kun taas RDF graafin. RDF-skeemamalli on joustavampi kuin vastaava XML-määrittely mahdollistaen uusien (hyvin määriteltyjen) triplojen lisäämisen joustavammin RDF-dokumenttiin kuin elementtien XML-dokumenttiin. Viimeiseksi URI-muodon käyttäminen RDF-triplojen osissa antaa globaalit tunnisteet, jollaista XML-määrittelyissä ei ole saatu kattavasti tuotettua. XML- tai CSV-dokumentteihin nähden RDF-

dokumentti on siis paremmin kuvattua dataa ja dokumentti on muunneltavampi ilman, että agentit menettävät kykynsä hyödyntää sitä. Se on myös sisällöltään yksiselitteisempi globaalisti yksilöivien tunnisteidensa vuoksi. Semanttisen rikkauden käsite on tässä erityisen hyödyllinen – RDF-dokumentti on aina XML-dokumenttia rikkaampi semanttisella ulottuvuudellaan.

RDF-määrittelyn URI-osoitteiden käytöstä on se seuraus, että data linkittyy kuin itsestään toisiin datavarastoihin. Enää sovelluksen ei tarvitse yhdistää kahdesta tai useammasta erillisestä datajoukosta tietoja toisiinsa jonkin näiden ulkopuolella pääteltävissä olevan säännön mukaan. Datassa itsessään on sisäänrakennettuna nämä linkit, jolloin kaikesta semanttisesta datasta tulee automaattisesti verkko, josta saadaan kaikki verkosta syntyvät hyödyt.

Internet on maailman suurin datavarasto. Se, millaista dataa sieltä on, ei ole tiedossa. Emme tiedä, onko jossakin tiettyyn tarpeeseen dataa ja jos on, niin mistä se on löydettävissä. Hakukoneet auttavat tässä ongelmassa, mutta niiden kapasiteetti on rajallinen. RDF-muotoisessa datassa olevat linkit auttavat löytämään lisää dataa. Joitakin hakukoneita semanttiselle datalle on yritetty luoda, Swoogle esimerkkinä tällaisesta. Julkisten instituutioiden toimesta on alettu perustaa avoimen datan portaaleja, joissa on sekä datajoukkoja että linkkejä avoimen datan lähteisiin. Juuri näiltä sivuilta etsin tutkimukseeni dataa arvioidakseni näissä portaaleissa julkaistun avoimen datan laadun.

5.2. Laadun arviointi

Avoin data on niin laaja käsite, että sen alla julkaistuilla datajoukoilla ei vaikuta olevan mitään yhteistä. Jos halutaan arvioida datan laatua, on analysoitava ne tekijät, joista laatu syntyy. Pipino (2002) antaa 16 eri ulottuvuutta, joilla mitata datan laatua. Tällaisia ovat mm. datan saamisen helppous, datan sopivuus käsillä olevan ongelman ratkaisuissa, datan täydellisyys ja virheettömyys, kuinka arvostettua data on, jne. Avoimen datankin voisi arvioida näillä kriteereillä, ja monessa tilanteessa tuleekin tehdä niin: jos toimittaja käyttää dataa artikkelia kirjoittaessaan (datajournalismi on nouseva journalismin laji), hän joutuu arvioimaan mm. sen luotettavuutta. Semanttisen verkon kohdalla konteksti on toinen. Semanttinen verkko syntyy, kun data täyttää tietyt sille asetetut ehdot, ja dataa tulee arvioida tässä kontekstissa näitä ehtoja tarkastelemalla. Ehtoja syntyy edellä esitellyistä hyödynnettävyyden haasteista.

Hyödynnettävyyden jäsentämiseksi ja sen arvioimiseksi laadullisella asteikolla on olemassa oleva arviointikehys. Sellaisen on esittänyt Berners-Lee (2010) ”kannustaakseen valtioita avaamaan dataa”. Arviointikehyksessä annetaan datajoukoille yhdestä viiteen tähteä siten, että viisi tähteä on laadukkainta dataa. Arviointikehyksestä on hyvä esitys Berners-Leen artikkelissa ja tälle arviointikehykselle omistetulla verkkosivulla. Lyhyesti kuvattuna laadun arviointi tapahtuu alla kuvattujen sääntöjen mukaan. Jokainen askel lisää jotain edellisiin vaatimuksiin sisällyttäen sen implisiittisesti itseensä. Asteikko on kuvattu taulukossa 1.

Kategoria	**Ehdot kategoriaan pääsemiseksi**
Yksi tähti *	Data on julkaistu missä hyvänsä muodossa, mutta avoimella lisenssillä.
Kaksi tähteä **	Data on julkaistu jossakin koneluettavassa muodossa.
Kolme tähteä ***	Data on julkaisu jossakin avoimessa koneluettavassa muodossa.
Neljä tähteä ****	Data on kuvattu RDF- tai SPARQL-määritysten mukaan, jotta siihen voidaan osoittaa muualta (voidaan linkittää dataan).
Viisi tähteä *****	Datassa on linkkejä muihin datajoukkoihin.

Taulukko 1: Viiden tähden avoimen datan laadullinen luokittelu.

Viiden kategorian luokittelu on suppea verrattuna esimerkiksi Pipininon 16-kohtaiseen mittariin. Se on kuitenkin käyttöyhteydessään tarkoituksenmukainen. Iso osa julkisesta datasta on suljettua, ja sen tuominen internetiin avoimella lisenssillä on jo iso askel kohti avoimuutta. Kahden ja kolmen tähden ero on pieni, ja suuri osa datasta on muunnettavissa kahdesta kolmen tähden dataksi – ja se on sallittua vaaditun avoimen lisenssin mukaan. Neljä tähteä saa, kun tekee kuvauksen RDF-määrittelyn mukaan – edes skeemaa ei tarvitse tehdä – siihen voi linkittää ja siitä tulee osa semanttista verkkoa. Viides tähti liittää datajoukon aktiivisesti osaksi semanttista verkkoa niin, että linkitykset tekevät siitä aidosti semanttisen verkon solmun.

Kritiikkiä tätä luokittelua kohtaan on osoitettu juuri siksi, että se ei vaadi tosiasiassa kuvaamaan dataa, vaan ainostaan noudattamaan annettua syntaksia. RDF:n perusmäärittely ei sisällä linkitysten ja tyyppi-predikaatin lisäksi juurikaan syventävää informaatiota, vaan tällainen luodaan RDFS- ja OWL-määrittelyiden mukaisilla kuvauksilla. Semanttisen rikkauden syvenemisen voisi ottaa huomioon tähdistössä tai se voisi olla mukana yksinomaan RDF-skeemojen mukaisten datajoukkojen semanttista rikkautta arvioivaa kehystä.

Kahden ja kolmen tähden ero on merkittävä mutta ei käytännönläheinen. Kun data on avointa, mutta siihen ei pääse käsiksi kuin yhdellä sovelluksella, puhutaan Vendor locking -tyyppisestä suojaamisesta. Tosiasiassa moni tiedostomuoto, kuten Excel, on jopa Microsoftin työkaluilla muutettavissa avoimeen muotoon, jolloin tällaista lukitusta ei tosiasiassa ole. Tällaiset formaatit

pitäisi mielestäni hyväksyä kolmen tähden luokaan. Tässä tutkielmassa noudatan kuitenkin yleistä käytäntöä ja lasken Excel-sovelluksen XLS-tiedostot kahden tähden arvoisiksi.

5.3. Datan kerääminen

Arvioitavaksi tarvittavaa dataa kerättiin vuoden 2012 joulukuusta vuoden 2013 helmikuuhun ulottuvalla ajanjaksolla. Avoimen datan sivustoja käytiin läpi noin 20 valtion ja kuuden globaalin instituution osalta. Lopulta lähempään tarkasteluun valittiin yhteensä 20 avoimen datan sivustoa, joiden kattavuus sekä valtioiden koon että sijainnin puolesta oli riittävä. Valtioista valituiksi tulivat Yhdysvallat, Uusi-Seelanti, Tsekki, Iso-Britannia, Kanada, Ranska, Norja, Kenia, Australia, Alankomaat, Espanja, Brasilia, Intia ja Singapore. Muista julkisista instituutioista tarkasteltiin Maailman pankin, Euroopan Unionin ja Yhdistyneiden Kansakuntien portaaleita. Erikoistapauksina mukana on kaupungeista Buenos Aires (Argentiina), kansalaisjärjestön ylläpitämä PublicData.eu ja yksityisen henkilön ylläpitämä sivu Ruotsista. Kerätty aineisto on liitteenä 1.

Alustava läpikäynti osoitti, että avoimen datan sivustot ovat selvästi samaan lopputulokseen tähtääviä mutta toteuttavat sen yksityiskohdissaan usein toisistaan poikkeavalla tavalla. Tarkasteluun valittuja sivustoja on perustettu vuosien 2009 – 2012 aikana, joten ne ovat eri kehitysvaiheissa. Sivustoja tutkittaessa kävi ilmi, että useat sivustot käyttävät kahta avoimen datan julkaisuun tarkoitettua sovellusta, Socrates- ja CPAN-sovelluksia. Voi olettaa, että nämä sovellukset myös ohjaavat sitä, millaista dataa sivustoilla julkaistaan, jolloin sivustot muistuttavat toisiaan myös sisällön suhteen, ja näin myös näyttää olevan.

Kielimuuri tulee vastaan joidenkin sivustojen osalta. Vaikka Google translator auttaa joissain tilanteissa, kuten lisenssin ymmärtämisessä, esimerkiksi Italian sivustoa oli käytännössä mahdoton analysoida. Tätä vaikeutti sekin, että kyseisellä sivustoilla oli rikkinäisiä linkkejä – eikä Italian sivusto ollut poikkeus tässä suhteessa. Toisaalta samojen kahden sovelluksen yhdenmukainen käyttö mahdollisti esimerkiksi espanjankielisten sivustojen analysoinnin, sillä kielten välillä olevien yhtenäisten sanarunkojen ja sovellusten käytäntöjen vuoksi datan löysi samasta kohdasta navigaatiorakennetta. Sovellukset tuottavat myös datajoukoista yhdenmukaisia koosteita, kuten niiden tyypit ja lukumäärät, joiden tulkintaan käytetty kieli ei vaikuta.

Tavoitteena on vastata kysymykseen: mihin avattu data sijoittuu viiden tähden asteikolla arvioituna? Portaaleja arvioidessa kävi ilmi, että arvioinnin tekeminen ei ole helppoa. Joillakin portaaleilla on tuhansia tai kymmeniätuhansia datajoukkoja, ja useilla niistä on myös paljon dataa, joka löytyy jostain muualta kuin kyseisestä palvelusta. Jokaista palvelussa olevaa datajoukkoa ei ole voitu kategorisoida. Portaalien sisällöstä käy kuitenkin ilmi tietoja, joiden perusteella voi tehdä johtopäätöksiä datajoukkojen yleisestä laadusta. CPAN- ja Socrates-sovellukset koostavat ja esittävät informaatiota datajoukoista. Socrates-sovellus soveltuu parhaiten taulukkomuodossa olevan datan jakamiseen, ja siksi Socrates-sovellusta käyttäville sivustoille on tuotu lähinnä tällaista

dataa, kun muilla sivuilla on esimerkiksi karttatasoja ja PDF-dokumentteja. Socrates toki mahdollistaa myös muiden dokumenttien tuomisen sivuille ja näiden lataamisen palvelusta. Tällaista dataa Socrates-sovelluksella toteutetuilla sivuistoilla on kuitenkin vähän lukuun ottamatta Yhdysvaltojen portaalia.

Samainen Socrates-sovellus tarjoaa myös vientitoiminnot, joissa datan saa ladattua XLS-, XML-, CSV- tai RDF-muodossa – tällöin lähes kaikki näillä sivustoilla oleva data on ainakin kolmen tähden arvoista. CPAN-sovelluksessa annettuihin datajoukkoihin – ovatpa nämä palvelusta ladattavissa tai linkitettyinä – annetaan datajoukon tyyppi. CPAN tuottaa suoraan koosteita datajoukkojen sisällöistä, jolloin näkyvillä on datajoukkojen määrä ja tyyppi. Useilla sivustoilla dataa on niin vähän, että se on vielä selattavissa läpi ja kokonaiskuva ja yksittäisten datajoukkojen laatu selviää suoraan. Joillakin sivustoilla, kuten Kanadan portaalissa, datan alkuperä kategorisoidaan. Kun kategorian sisältä löytyy pitkällisen etsimisen jälkeen vain tietyn laatuista dataa, voi tehdä sen johtopäätöksen, että kategoriassa on pääosin laadultaan samanlaista sisältöä. Erityisesti tilastollista dataa on paljon, ja se on palveluissa johdonmukaisesti joko kahden tähden tai kolmen tähden dataa.

5.4. Tulokset

Taulukossa 2 on yleistiedot portaaleista, ja taulukossa 3 on datan laadun arviointiin tarvittavat tiedot. Millaista laatua tutkimustulosten perusteella valtiot julkaisevat avoimen datan käsitteen alla? Arvioin tuloksia viiden tähden luokittelun avulla, aina jokaisen tähden osalta erikseen. Sellaiset johtopäätökset laadusta, jotka eivät ole tällä asteikolla tulkittavissa, käyn läpi johtopäätösten yhteydessä.

Sivustoja tutkittaessa nousi esiin myös muita sellaisia tietoja, jotka katsoin hyödylliseksi ottaa mukaan analyysiaineistoon. Kahteen taulukkoon on koostettu alla luetellut tiedot:

- Valtio tai instituutio, julkaisuvuosi ja osoite.
- Datajoukkojen lukumäärä.
- Dokumenttien muoto. Lähinnä mainittu SHP-, XLS-, PDF-, HTML-, XML-, CSV- ja RDF-muodot, vaikka muitakin vastaavia on sivustoilla voinut olla. Näillä dokumenttimuodoilla on merkitystä arvioitaessa yhden, kahden ja kolmen tähden kategoriaan sijoittumista ja arvioitaessa yleisintä datajoukon laatua. SHP on karttatasotiedosto. XLS on Excel-sovelluksen tiedosto. PDF on ei-rakenteinen dokumentti. HTML, XML ja CSV ovat avoimia rakenteisia dokumentteja.
- Sivuston päivitystieto (päivitetäänkö sitä säännöllisesti – siis esimerkiksi viikoittain).
- Merkintä siitä, onko sivustolla yhden, kahden, kolmen, neljän ja viiden tähden dataa – jokaisesta merkintä erikseen.
- Arvio siitä, mikä on datan keskimääräinen laatu. Myös korkein löydetty laatu on merkitty

erilliseen sarakkeeseen.

- Arvio siitä, onko sivustolla käytetty CPAN- tai Socrates-sovellusta vai jotakin yksittäistä toteutusta.
- Mitä lisenssiehtoja sivustolla on käytetty? Jos sivustolla on dataa, jonka lisenssi ei käy ilmi tai se ei ole avoin, se on mainittu.

#	Valtio tai instituutio	Verkko-osoite	Jul-kaisu	Lisenssi	Formaatit	Päivitys	Sovellus
01	Maailman pankki	data.worldbank.org finances.worldbank.org	4/2010	Yleensä avoin Lueteltu erikois-tapaukset	Excel PDF CSV XML	Usein	Socrates + Oma
02	Euroopan unioni (EU)	open-data.europa.eu	2012	Aineistokohtainen Yleisesti avoin	Excel CSV Tilastot	Usein	CKAN
03	PublicData.eu	publicdata.eu	6/2011	Tuntematon Ylesesti avoin	Kaikki	Usein	CKAN
04	Yhdistyneet Kansakunnat (YK)	data.un.org	2/2008	Kaikki avointa	XML CSV	Usein	Oma
05	Yhdysvallat	data.gov	5/2009	Aineistokohtainen	Excel XML CSV SHP KML	Usein	Socrates + Oma
06	Uusi-Seelanti	data.govt.nz	10/2009	Joitakin epäselviä Yleisesti avoin	Excel PDF HTML SHP API	Usein	Oma
07	Tsekki	opendata.cz	2011	Tuntematon Avoin data	?	?	CKAN
08	Iso-Britannia	data.gov.uk	9/2009	Tuntematon Avoin data	Excel PDF XML CSV RDF	Usein	CKAN
09	Kanada	data.gc.ca	9/2012	Avoin data	CSV XML KML RDF	Usein	Oma
10	Ranska	data.gouv.fr	12/2011	Avoin data	XLS XML CSV SHP	Usein	Oma
11	Norja	data.norge.no	4/2012	Avoin data	XLS CSV XML MUU	Epä-säännöllisesti	Oma
12	Ruotsi	opengov.se 1)	?	Avoin data	?	?	Oma
13	Kenia	opendata.go.ke	7/2011	Avoin data	XLS CSV	Säännöllisesti	Socrates
14	Australia	data.gov.au	3/2011	Avoin data	XLS CSV SHP	Usein	Oma

#	Valtio tai instituutio	Verkko-osoite	Jul-kaisu	Lisenssi	Formaatit	Päivitys	Sovellus
15	Alankomaat	data.overheid.nl	10/2011	Avoin data	PDF XLS CSV RDF	Usein	CKAN
16	Espanja	datos.gob.es	10/2011	Avoin data	HTML CSV PDF XLS SHP	Usein	Oma
17	Brasilia	dados.gov.br	5/2012	Ei avoin lisenssi	XML PDF CSV	Ei selviä	CKAN
18	Argentiina, Buenos Aires	data.buenosaires.gob.ar	?/2012	Avoin data	CSV XLS	Päivitetään	CKAN
19	Intia	data.gov.in	10/2012	Ei mainita sivuilla	XLS CSV HTML	Päivitetään	Oma
20	Singapore	data.gov.sg	6/2011	Rajoittava	CSV XLS	XML	Oma

Taulukko 2: Avoimen datan portaaleiden ja datan yleispiirteet.

1) Ruotsin kohdalta on otettu yksityisen toimijan ylläpitämä sivusto. Virallinen sivusto on kokoelma API-osoitteita.
2) Argentiina kohdalla mukaan on otettu Buenos Airesin sivusto. Argentiinalla ei ole valtion perustamaa sivustoa.

#	Valtio tai instituutio	*	**	***	****	*****	Keskim. laatu	Korkein laatu	Data-joukkoja	SPARQL
01	Maailman pankki	On	On	On	On	Ei	***	***	1200/50 1)	Ei
02	Euroopan unioni (EU)	On	On	On	Ei	Ei	***	***	5850 2)	Ei
03	PublicData.eu	On	On	On	On	Ei	? 3)	**** 4)	>16000	Ei
04	Yhdistyneet kansakunnat (YK)	On	Ei	On	Ei	Ei	***	***	34 tietokantaa 5)	Ei
05	Yhdysvallat	On	On	On	On	Ei	***	***	378529	Ei
06	Uusi-Seelanti	On	On	On	Ei	Ei	***	***	2300 6)	Ei
07	Tsekki	On	On	On	On	Ei	**	****	161	On
08	Iso-Britannia	On	On	On	Ei	On	***	*****	>9000	Ei
09	Kanada	On	On	On	On	Ei	***	****	12800 7)	Ei
10	Ranska	On	On	On	Ei	On	**	*****	353300 8)	Ei
11	Norja	On	On	On	Ei	Ei	?	***	100 9)	Ei
12	Ruotsi	On	On	On	On	On	?	*****	110 10)	Ei
13	Kenia	On	Ei	On	On	Ei	***	***	534	Ei
14	Australia	On	On	On	Ei	Ei	***	***	1126	Ei
15	Alankomaat	On	On	On	On	Ei	***	****	5193	On
16	Espanja	On	On	On	Ei	Ei	* 11)	***	640	Ei
17	Brasilia	On	On	On	Ei	Ei	0	***	100	Ei
18	Argentiina, Buenos Aires	On	On	On	Ei	Ei	***	***	80	Ei
19	Intia	On	On	On	Ei	Ei	***	***	115	Ei
20	Singapore 12)	Ei	Ei	Ei	Ei	Ei	0	***	7900	Ei

Taulukko 3: Avoimen datan portaalien datan laatu.

1) 1200 aikasarjaa, 50 datakatalogia joissa voi olla useita datajoukkoja.
2) 5600 datajoukkoa Eurostatilta.
3) Laatu vaihtelee suuresti, koska datajoukkoja on hyvin erilaisista palveluista.
4) Lähinnä Iso-Britannian yksittäisten datajoukkojen vuoksi, mutta enenevässä määrin myös muista maista.
5) Oman ilmoituksen mukaan ”34 databases - 60 million records”.
6) Paljon karttakuvia, satoja datajoukkoja on tällaista materiaalia.
7) 8800 tilastolaitokselta ja 3000 maatalousministeriöltä.
8) 294000 datajoukkoa Excel-dataa.
9) Norjan sivustolla on ainoastaan linkkejä muualle, ja sivustolla on suppea kuvaus data-aineistosta. Linkkien takaa löytyy kolmen tähden dataa. Keskimääräistä laatua on vaikea arvioida.
10) Linkkejä toisille sivustoille. Sivustolla käytetään tähtiluokittelua, ja kaikkiin luokkiin löytyy dataa. Keskimääräistä laatua on vaikea arvioida.
11) Espanjan sivustolla linkitetään HTML-dokumentteihin, joissa saattaa olla tiedosto ladattavaksi.

12) Singaporen lisenssiehto tehnee muuten avoimesta datasta suljettua.

Yhden tähden dataa on lähes kaikissa portaaleissa, vain yhdeltä tämä puuttuu täysin. Ensimmäinen ja minimivaatimus Berners-Leen viiden tähden asteikolla on avoin lisenssi. Singaporen käyttämä lisenssi on selvästi avoimen datan vastainen, se mm. vaatii seuraamaan sitä, onko datajoukko edelleen sivustolla. Vain kymmenen sivustoista ilmoittaa datan olevan yksinomaan sellaisen lisenssin alaista, joka on luokiteltavissa avoimeksi lisenssiksi. Maailmanpankin sivustolla on listattu erikseen rajoitetun lisenssin alla olevat datajoukot, joita ei tutkimuksen aikaan ollut lainkaan – siis Maailmanpankki olisi tosi asiassa yhdestoista täysin avoimella lisenssillä toimiva sivusto, vaikka varaa mahdollisuuden myös suljettuihin lisensseihin. Muut sivustot sekoittavat avoimella ja rajoitetulla lisenssillä olevaa dataa, eivät tuo esiin lisenssiä riittävän selvästi esiin, tai kuten Intian sivusto, eivät sisällä lisenssi-informaatiota lainkaan. Intian osalta oletan, että lisenssi on ennemmin avoin kuin suljettu. Yhden tähden dataa on Singaporea lukuun ottamatta kaikilla sivustoilla, ja muillakin sivustoilla avoin on yleisin lisenssi – tältä osin yhden tähden datan määritelmä täytyy useimmiten. Yhden tähden luokitteluun sisältyy myös oletus siitä, että data on sellainen dokumentti, joka on ”missä tahansa formaatissa”. Usein dataan viitataan luomalla linkki sivulle tai sivustolle, josta se voidaan ladata. Muita portaaleista ladattavia yhden tähden dokumentteja ovat mm. PDF-dokumentit, joita löytyy jokaiselta tutkitulta sivustolta. Myös karttasovelluksien kanssa käytettävä shapefile-muoto on yleinen useilla sivustoilla.

Kahden tähden kategoriaan pääsee rakenteisella dokumentilla, joka on sovelluksille luettavassa muodossa. Excel-sovellus on tästä hyvä esimerkki: sen tuottamat tiedostot voidaan lukea paitsi Excel-sovelluksessa myös useilla avoimen lähdekoodin sovelluksilla. Erityisesti Excel-sovelluksen XLS-tiedostojen yleisyyden vuoksi tällaista dataa löytyy lähes kaikkilta tutkituilta sivustoilta. Tässä kohtaa tulkitsen niin, että jos data on saatavissa sekä XLS-tiedostona että CSV-tiedostona, se on kolmen tähden dataa. Vain jos datajoukko on saatavissa ainoastaan XLS-tiedostona, se on kahden tähden dataa. Socrates-sivustoilta ei löydy aina kahden tähden dataa, sillä sellainen pitäisi olla erikseen linkitettynä datajoukkona – Socrates tuottaa siihen tuodusta raakadatasta aina esimerkiksi CSV- ja RDF-dokumentit. Kahden tähden dataa löytyi 17 tutkitusta portaalista.

Kolmen tähden dataa ovat esimerkiksi kaikki CSV- ja XML-muotoiset tiedostot. Erityisesti Socrates-sovellus antaa mahdollisuuden ladata saman datan useassa eri formaatissa, jolloin data on aina kolmen tähden dataa. Samoin useilla sivuilla taulukkodata on julkaistu Excel-formaatin lisäksi CSV-formaatissa. Portaaleista 19 sisältää tällaisia datajoukkoja tai linkkejä näihin.

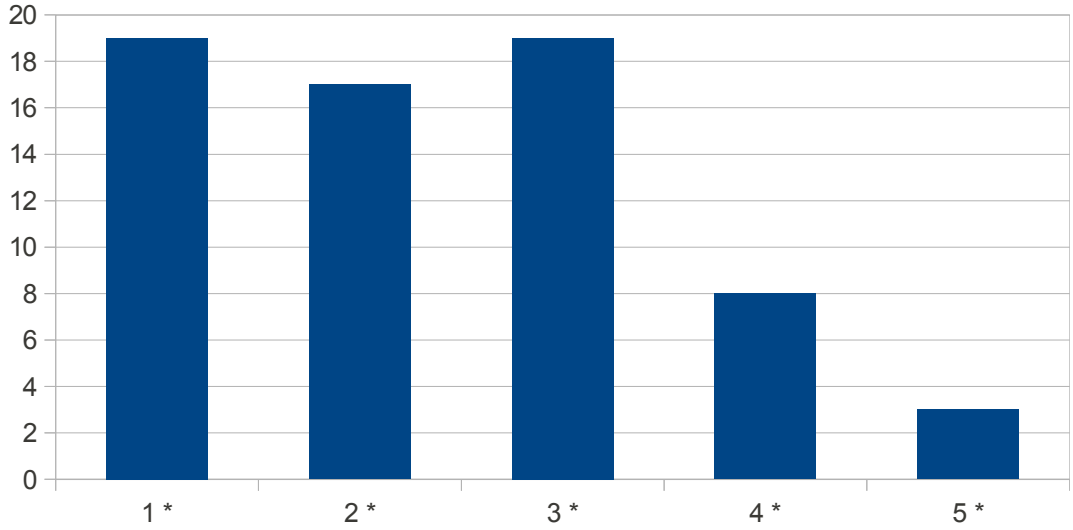
Neljän tähden dataa on vain harvoilla sivustoilla. Socrates-sovellus useimmissa tapauksissa (poikkeuksena Euroopan Unionin open-data.europa.eu-portaali) toteuttaa linkityksen RDF-datajoukkoon. Se antaa jopa hakea yksittäisen RDF-aligraafin dokumentin sisältä. Neljän tähden datassa riitti, että dataan voidaan linkittää, sen ei tarvitse sisältää linkkejä muualle. Muissa kuin Socrates-sovelluksen datajoukkojen tapauksessa neljän tähden dataa on esimerkiksi sellainen, jota

voi kysellä SPARQL-rajapinnan kautta, ja joillakin portaaleilla data on linkitettyä dataa tämän vuoksi. Kahdeksalla sivustolla on neljän tähden dataa.

Viiden tähden dataa on kolmella sivustolla. Yksi näistä (publicdata.eu) linkittää muita lähteitä, jolloin mm. mukana oleva Iso-Britannian viiden tähden data sisältyy sivuston arviointiin. Ruotsin yksityishenkilön ylläpitämä lista linkittää tällaista dataa. Muutoinkin portaaleissa on voitu linkittää sivustoihin, joilta löytyy viiden tähden dataa, vaikka tällaista ei ole tullut tämän tutkimuksen aikana esiin.

6. Johtopäätökset ja pohdintaa

Tutkimustulosten valossa näyttää siltä, että semanttinen verkko ei tällä hetkellä hyödy avoimesta datasta. Tutkimuksen perusteella avoimen datan laatu viiden tähden asteikolla arvioitaessa on heikohkoa, sillä avoimen datan sivustoilta löytyy keskimäärin kolmen tähden dataa, mikä viisiportaisella asteikolla ei liitä dataa vielä mitenkään osaksi semanttista verkkoa. Kuvan 5 diagrammi portaaleista, joilta eri tähtiluokituksen mukaista dataa on löytynyt kertoo, kuinka vähän semanttista dataa on saatavilla avoimen datan portaaleissa.



Kuva 5: Portaalien määrä 20 tutkitusta, josta on löytynyt ainakin yksi kyseisen luokituksen saanut datajoukko.

Avoimen datan ja semanttisen verkon yhteyden tutkiminen ei ollut ainoa tutkielman tavoite. Tutkimuksen aikana esiin nousi kolme keskeistä teemaa. Ensimmäinen on avoimen datan käsitteiden käyttö. Olen pyrkinyt tuomaan esiin asiat käsitteiden takaa. Nämä käsitteet tarvitaan, kun tarkastellaan avoimen datan portaaleita, mutta niiden merkitysten määrittäminen on ollut yhtä tärkeää kuin avoimen datan portaaleita arvioiva tutkimus.

Toisena ja tärkeimpänä tutkimuksen kohteena on avoimen datan laatu, jota tutkimuksen kautta selvitän tähtiasteikolla. Tähtiasteikko ei lopulta ole mielestäni ainoa mittari, jolla arvioida sivustojen datajoukkojen laatua.

Kolmantena teemana tutkimusta tehtäessä kiinnittyi huomio itse sivustojen laatuun. Niiden on tarkoitus olla suurien julkisten instituutioiden dataportaaleita, eivätkä nämä instituutiot ole aina tässä onnistuneet.

Käsitteet avoin data, semanttinen data, linkitetty data ja semanttinen verkko liittyvät toisiinsa – mutta kuinka paljon? Näiden käsitteiden avaamisessa ja niiden takana olevien haasteiden selvittämisestä on hyötyä arvioitaessa sitä, miksi avoin data ei tuota enempää semanttisen verkon

dataa. Semanttisen datan ja semanttisen verkon tuottaminen on mittava työ: pelkästään RDF-triplojen tuottaminen ei ole vaativaa, kuten Socrates-sovellus osoittaa automatisoimalla taulukkomuotoisen datan RDF/XML-koodauksen. Sellaisen skeeman tuottaminen, joka käyttää RDFS-määritelyä tai peräti OWL-käsitteitä, on vaativa työ. Semanttisen verkon dokumenttien linkittäminen toisiinsa vaatii laajempaa koordinointia ja haastavampia teknisiä ratkaisuja kuin Excel-tiedoston ladattavaksi saaminen. Semanttisesta datasta semanttiseksi verkoksi tarvitaan vielä syvempää osaamista ja resursseja – erityisesti jos samalla tuotetaan SPARQL-palveluita. CKAN-sovellus yrittää auttaa tässä tarjoamalla SPARQL-kyselyrajapinnan ja triplavaraston. Avoin data on siis hyvin kaukana semanttisesta datasta ja verkosta, näiden välillä on leveä kuilu.

Avoimen datan analysointi paljasti paljon käyttökelpoista ja hyödynnettävää dataa. Tässä tutkielmassa avoimen datan laatua on lähdetty arvioimaan siitä lähtökohdasta, miten se soveltuu osaksi semanttista verkkoa – siitähän neljäs ja viides tähti tulevat. Semanttinen verkko ei ole kuitenkaan ainoa avoimen datan käyttöyhteys. Viiden tähden asteikko ei sovellu hyvin avoimen datan kokonaisuuden analysointiin, koska se ei nosta kahden ja kolmen tähden datan hyötyjä riittävästi esiin. On nähtävä hyödyt avoimesta datasta myös muille kuin semanttisen verkon sovelluksille. Arvioitaessa dataa tästä lähtökohdasta sen laatuun vaikuttaa kolme ensimmäistä tähteä. Laatua tulisi arvioida siis sen soveltuvuudesta yleisesti perinteiseen sovelluskäyttöön eikä yksinomaan semanttisen verkon agenttien käyttöön. Muuta käyttöä on olemassa, ja tämän käytön osalta voidaan arvioida datan laatua toisesta tarkastelukulmasta.

Arvioitaessa datan soveltuvuutta semanttisena datana tai semanttisen verkon käyttöön luokittelu kolmeen kategoriaan on liiaksi rajoittava. Kolmesta viiteen tähteä sisältyy jollakin tavalla semanttiseen verkkoon. Se, linkitetäänkö dataan ja datasta on toki merkittävää, mutta se ei kerro esimerkiksi semanttisesta rikkaudesta mitään. Semanttinen rikkaus syntyy skeeman laadusta, siitä miten laajasti se antaa mahdollisuuden kuvata dataa. Semanttisen rikkauden mahdollistamiseksi eri määrittelyt laajentavat käytettävissä olevaa kuvauskieltä niin, että RDF, RDFS ja OWL mahdollistavat aina laadukkaammat metatietokuvaukset. Semanttisen laadun mittarin täytyy sisältää tällaisten skeemojen laadun arvioinnin. Tällaista pyrkimystä ei vielä ole havaittavissa. Tähtiluokitus kaipaa korjausta, jossa sen ensimmäiset ja viimeiset kategoriat erotetaan toisistaan, ja niiden kohteena olevalle datalle luodaan uudet laatumittaristot.

Avoin data tekee tuloaan, ja se näyttää tulevan nopeasti. Internet on väline, jolla voidaan nopeasti julkaista suuria määriä dataa, ja useat julkiset instituutiot ovat alkaneet näin tehdä. Tällä hetkellä näyttää olevan käynnissä prosessi, jossa tehdään avoimen datan portaaleita, joissa toisaalta on dataa, mutta joihin myös linkitetään sivustoja tai datajoukkoja muilta portaaleilta. Kehitys voi viedä useaan suuntaan. Toisaalta voi syntyä sellaisia portaaleja, joihin tuodaan data-aineistoa ja Socrates-sovellus on tehty tällaiseen tarkoitukseen. Voi syntyä myös portaaleita, joissa listataan avoimen datan lähteitä – joskin portaaliksi kutsuminen tässä tapauksessa on harhaanjohtavaa, kun portaali ei koosta sisältöä vaan toimii linkkiluettelona ja hakemistona. Yleistyykö näistä

lähtökohdista toinen, vai ottavatko eri toimijat eri lähtökohdista erilaiset avoimen datan strategiat? Olisi hyödyllistä analysoida tähän mennessä kertyneiden kokemusten perusteella eri strategioiden hyödyt ja haitat eri tavoitteisiin nähden. Kolmas lähestymistapa, avoimet rajapinnat, on yksi tässä vertailussa mukaan otettava vaihtoehto – silloin kun se on SPARQL-solmu se täyttää myös semanttisen verkon ehdot. Semanttisen verkon rakentamisen näkökulmasta tärkeä kysymys on, miten instituutiot saadaan tavoittelemaan Berners-Leen asteikon ylimpiä portaita, eikä viiden tähden asteikko ole ainoa eikä välttämättä edes oleellinen tavoitemittari avoimen datan portaaleille.

Valtioilla on motiivi toimia aktiivisesti avoimen datan suhteen. Vaikka väitettä on vielä avoimen datan lyhyen olemassaolon vuoksi vaikea todistaa, on syytä epäillä, että avoimesta datasta voi tulla informaatiotaloudessa valtion menestystä siivittävä tekijä. Avoin data ei ole itseisarvo, vaan sitä hyödyntävät palvelut, kuten sovellukset ja semanttinen verkko, tuottavat lisäarvoa yhteiskunnalle ja siinä toimiville kaupallisille ja ei-kaupallisille tahoille. Laatu, määrä ja lisensointi ovat tässä mielessä keskeisiä laatumittareita.

On olemassa myös avoimen datan käyttöön kannustavia projekteja, kuten mm. Suomessa järjestettävät Apps4-kilpailut (”Apps4Finland”, 2012) tai eurooppalainen Open Data Challenge -kilpailu (Jonathan, 2012). Näissä kilpailuissa, joita järjestetään eri maissa, haastetaan yleisö – lähinnä ohjelmistoalan ammattilaiset – kilpaileman avointa dataa käyttävien sovellusten kehittämisessä. Näiden sovellusten laatu ei ehkä kerro niinkään avoimen datan laadusta, mutta se kertoo siitä laadusta, jota avoimella datalla on saavutettu. Semanttisen verkon hyödyt ovat vielä tuntemattomia. Tärkeintä lopulta on se, minkälaista hyötyä avoimesta datasta seuraa kansalaisille ja yhteiskunnalle.

Viiteluettelo

About W3C. (2012). Retrieved 3, 2013, from http://www.w3.org/Consortium

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applies Systems Analysis, 16*, 3-9.

Adida, B., Herman, I., Sporny, M. & Birbeck, M. (2012). RDFa 1.1 primer. Retrieved, 2013, from http://www.w3.org/TR/xhtml-rdfa-primer

Anderson, P. (2007). What is web 2.0? ideas, technologies and implications for education. *JISC Technology and Standards Watch, 1(*1*)*, 4-26.

Anderson, C. (2009). *Free: The Future of a Radical Price.* Hyperion Books.

Apps4finland 2012. (2012). Retrieved 2013, 3, from http://apps4finland.fi

Atwood, T. (1985). An object-oriented DBMS for design support applications. An Object-Oriented DBMS for Design Support Applications*,* (*Proceedings of the IEEE COMPINT 85*) 299-307.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K. Choi, N. Noy, D. Allemang, K. Lee, L. Nixon, . . . P. Cudré-Mauroux (Eds.), *The Semantic Web* (pp. 722-735). Springer.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2003). *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press.

Berners-Lee, T. (1998). Why RDF model is different from the XML model. Retrieved 3, 2013, from http://www.w3.org/DesignIssues/RDF-XML

Berners-Lee, T. (2010). Linked data - design issues. Retrieved 2013, 2, from http://www.w3.org/DesignIssues/LinkedData.html

Bikakis, N., Tsinaraki, C., Gioldasis, N., Stavrakantonakis, I., & Christodoulakis, S. (2013). The XML and semantic web worlds: Technologies, interoperability and integration: A survey of the state of the art. In I. E. Anagnostopoulos, M. Bieliková, P. Mylonas & N. Tsapatsoulis (Eds.), *Semantic Hyper/Multimedia adaptation* (pp. 319 – 360). Springer, Berlin Heidelberg.

Brickley, D., & Guha, R. V. (2004). RDF schema. Retrieved, 2013, from http://www.w3.org/TR/rdf-schema

Cyganiak, R., & Jentzsch, A. (2011). The linking open data cloud diagram. Retrieved 3, 2013, from http://lod-cloud.net

Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley.

Daniel, J. W. (2002). Current patent practice. Retrieved 3, 2013, from http://www.w3.org/TR/2002/NOTE-patent-practice-20020124

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., . . . Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *Internet Computing, IEEE, 4*(5), 63-73.

Derrett, N., Kent, W., & Lyngbaek, P. (1985). Some aspects of operations in an object-oriented database. *IEEE Computer Society, 8*(4), 66-74.

Gruber, T. (2008). What is an ontology. *Encyclopedia of Database Systems.*Springer-Verlag.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies, 43* (5), 907 – 928.

Haase, P., Broekstra, J., Eberhart, A., & Volz, R. (2004). A comparison of RDF query languages. *The Semantic Web–ISWC 2004,* , 502-517.

Hammer, M., & McLeod, D. (1978). The semantic data model: A modelling mechanism for data base applications. *Proceedings of the 1978 ACM SIGMOD International Conference on Management of Data,* Austin, Texas. 26-36.

Handschuh, S. (2007). Semantic annotation of resources in the semantic web. In R. Studer, S. Grimm & A. Abecker (Eds.), *Semantic Web Services* (pp. 135–155). Springer, Berlin Heidelberg.

Hedden, H. (2008, Semantic tagging. *Econtent, October 2008*, 38–43.

Hendler, J. (2009). Web 3.0 emerging. *Computer, 42*(1), 111-113.

Heß, A., & Kushmerick, N. (2003). Learning to attach semantic metadata to web services. In D. Fensel, K. Sycara & J. Mylopoulos (Eds.), *The Semantic Web - ISWC 2003* (pp. 258-273). Springer, Berlin Heidelberg.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F. & Rudolph, S. (2012). OWL 2 web ontology language primer. Retrieved, 2013, from http://www.w3.org/TR/owl2-primer

Horridge, M., & Bechhofer, S. (2009). The OWL API: A java API for working with OWL 2 ontologies. *Proc.of OWL Experiences and Directions, 2009.*

Hyvönen, E., Viljanen, K., Mäkelä, E., Kauppinen, T., Ruotsalo, T., Valkeäpää, O., . . . Kurki, J.

(2007). Elements of a National Semantic Web Infrastructure – Case Study Finland on the Semantic Web. *Proceedings of the First International Semantic Computing Conference,* Irvine, California.

Jonathan, G. (2012). The open data challenge – organising Europe's biggest open data competition. Retrieved 3, 2013, from http://lod2.okfn.org/2012/11/07/the-open-data-challenge-organising-europes-biggest-open-data-competition

Kappel, G., Pröll, B., Reich, S., & Retschitzegger, W. (2006). *Web Engineering: The Discipline of Systematic Development of Web Applications,* Wiley.

Kashyap, V., & Sheth, A. (1996). Semantic and schematic similarities between database objects: A context-based approach. *The VLDB Journal, 5*(4), 276-304.

Knublauch, H. (2004). Ontology-driven software development in the context of the semantic web: An example scenario with Protege/OWL. *International Workshop on the Model-Driven Semantic Web,* Monterey, CA..

Levy, S. (2001). *Hackers: Heroes of the Computer Revolution.* (4th ed.) Penguin Books New York.

Maier, D., Otis, A., & Purdy, A. (1985). Object-oriented database development at servio logic. *Database Engineering, 18*(4), 58-65.

Manola, F., & Miller, E. (2004). RDF primer. Retrieved, 2012, from http://www.w3.org/TR/rdf-primer

Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication, 47*(10).

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence,* A.K. Peters

Millard, D. E., & Ross, M. (2006). Web 2.0: Hypertext by any other name? *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia,* 27–30.

Murugesan, S. (2007). Understanding web 2.0. *IT Professional, 9*(4), 34–41.

Open data. (2013). Retrieved 3, 2013, from http://en.wikipedia.org/wiki/Open_data

Open definition. (2013). Retrieved 3, 2013, from http://opendefinition.org/okd

O'Reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies,* (1), 17.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218.

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: The case of GEM. *Information Research, 6(*2*).*

Sabou, M., Lopez, V., Motta, E., & Uren, V. (2006). Ontology selection: Ontology evaluation on the real semantic web. In: *15th International World Wide Web Conference* (pp. 23-26). Edinburgh, Scotland.

Segaran, T., Evans, C., & Taylor, J. (2009). *Programming the Semantic Web* O'Reilly Media.

Semantic web stack. (2013). Retrieved 3, 2013, from http://en.wikipedia.org/wiki/Semantic_Web_Stack

Sequeda, J. (2012). Introduction to: RDF vs XML. Retrieved 3, 2013, from

http://semanticweb.com/introduction-to-rdf-vs-xml_b31071

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *Intelligent Systems, IEEE, 21*(3), 96–101.

Spencer, N. (2012). How much data is created every minute? Retrieved 4, 2013, from http://www.visualnews.com/2012/06/19/how-much-data-created-every-minute

Tauberer, J. (2008). What is RDF and what is it good for? Retrieved 3, 2012, from http://www.rdfabout.com/intro

Terzi, E., Vakali, A., & Hacid, M. S. (2003). Knowledge Representation, Ontologies, and the Semantic Web⋆. In *Web Technologies and Applications: 5th Asia-Pacific Web Conference, APWeb 2003, Xian, China, April 23-25, 2002, Proceedings* (Vol. 5, p. 382). Springer.

Underwood, G. J. (2011). Preserving electronic information for future generations. *Can.L.Libr.Rev., 36*, 112.

Wagner, R. P. (2003). Information wants to be free: Intellectual property and the mythologies of control. *Columbia Law Review, 103*, 3-22.

Xu, Z., Fu, Y., Mao, J., & Su, D. (2006). *Towards the semantic web: Collaborative tag suggestions. Collaborative Web Tagging Workshop at WWW2006,* Edinburgh, Scotland.

Liite 1. Tutkimusaineisto.

#00 Valtion tai instituution nimi.	
Sivun ylläpidosta vastaava taho, jos tiedossa.	
Sivun osoite.	
Julkaistu	Julkaisu kuukausi ja vuosi jos tiedossa.
Datajoukkoja	Sivustolla oleva datajoukko tai siellä listattu datajoukko. Datajoukon käsite ei ole yksiselitteinen, jolloin annettua lukumäärää on tarkennettu.
Päivitys	Päivitys tieheys vapaasti kuvattuna, esimerkiksi usein, kuukausittain, harvoin.
Formaatit	Yleisimmät sellaiset tiedostomuodot lueteltu, joilla merkitystä tähdityksessä: Excel, PDF, SHP, CSV ja XML. API-rajapinta mainittu erikseen, jos sellainen on. Sivustolla voi olla muita verrannollisia tiedostomuotoja, joita ei ole lueteltu.
Keskimääräinen laatu	Arvioitu keskimääräinen laatu datajoukkojen tyypin ja määrän mukaan.
Korkein laatu	Löydetty korkein laatu.
*	On/Ei tämän laatuista dataa.
**	On/Ei tämän laatuista dataa.
***	On/Ei tämän laatuista dataa.
****	On/Ei tämän laatuista dataa.
*****	On/Ei tämän laatuista dataa.
Lisenssi	Avoin, tai jos poikkeuksia, niin mainittu miten poikkeaa.
SPARQL-rajapinta	On/Ei ole SPARQL-rajapintaa.
Muita huomioita, kuten käytetty sovellus. Tarkennukset yllä mainittuihin tietoihin. Muuta laadun kannalta olennaista tai erityisiä huomioita.	

#01 Maailman pankki	
World Bank's Development Data Group	
data.worldbank.org Myös muita osoitteita, joissa Maaiman pankin keskeistä dataa: finances.worldbank.org databank.worldbank.org	
Julkaistu	4/2010
Datajoukkoja	1200 indikaattoria (*), 50 datakatalogia (**)
Mistä asti dataa	Aikasarjoja ainakin 1960 luvulta asti.
Päivitys	1/2013. Aikasarjoja päivitetään säännöllisesti.
Formaatit	Excel, CSV, PDF, JSON, XML, API, RDF, SHP
Keskimääräinen laatu	***

Korkein laatu	***
*	On.
**	On.
***	On.
****	On (***)
*****	Ei.
Lisenssi	http://data.worldbank.org/summary-terms-of-use Avoin käyttö, lisenssiä ei erikseen nimetty. Jotkin datajoukot voivat olla toisin ehdoin, mutta toistaiseksi sellaisia ei ole listattuna tätä varten osoitetulla sivulla. Linkit, joiden osoittamilla sivuilla nämä listataan ovat: http://go.worldbank.org/OJC02YMLA0 http://go.worldbank.org/R6942GMMH0
SPARQL-rajapinta	Ei

(*) Indikaattorit ovat tilastotietoja kuten BKT tai lukutaitoisten osuus väestöstä.
(**) Yhdessä datakatalogissa voi olla useita tietokantoja.
(***) Socrata-sovellus käytössä finances-osiossa. Tämä tuottaa RDF/XML-muotoisia dokumentteja. Dokumenteissa on käytössä mm. URI http://finances.worldbank.org/resource/wc6g-9zmq, johon laitamalla päätteeksi.rdf, saa RDF-dokumentin. Siis neljän tähden dataa.

Vain Socrateen “rowId”-rakenne skeema käytössä. Ei todellisia sisällön skeemoja. Ei OWL-kuvauksia.

Data on Maailmanpankin hallinnoimilla sivuilla, pääasiassa kahdella sivulla.

Iso osa sivustolla olevista dokumenteista voi olla koostettu taulukoina saatavasta avoimesta datasta.

Maailman pankki kerää globaalia tilastollista dataa ja julkaisee sitä eri formaateissa. Tarjolla on myös Excel-, PDF-, RDF- tai CSV-muotoisia raportteja generoiva verkkotyökalu. Työkalut eivät aina toimi, linkit eivät kaikki toimi.

Yhteenvetona, Maailmanpankin vahvuus on Socrates -ovelluksen käytössä, jossa on hyödynnetty RDF-dokumentiin linkittäminen. Siksi Maailmanpankkilla on myös neljän tähden dataa, vaikka useimmat datat ovat kahden tai kolmen tähden dataa.

#02 Euroopan unioni (EU)	
The European Commission Data Portal	
open-data.europa.eu	
Julkaistu	2012
Datajoukkoja	5850
Mistä asti dataa	Tilastosarjoja.
Viimeisin päivitys	Päivitetään usein.
Formaatit	Excel, CSV, tilasto-ohjelmaformaatit, RDF, XML.
Keskimääräinen laatu	***
Korkein laatu	***
*	On
**	On
***	On

****	Ei
*****	Ei
Lisenssi	Datajoukkokohtainen. Esimerkkejä. Creative Commons Attribution (Open Data) Europa Legal Notice (Open Data) Eurostat Copyright/Licence Policy (Open Data)
SPARQL-rajapinta	Ei. (*)

"The European Commission Data Portal provides access to open public data from the European Commission. It also provides access to data of other Union institutions, bodies, offices and agencies at their request."

Eurostat 5600 datajoukkoa, eli portaaliin on pääosin tuotu tilastodataa eurostatilta. Datajoukot ovat ***-laatua.

(*) Sisältää CKAN-sovelluksen SPARQL-integraation, joka tarjoaa linkitettyä dataa. Se antaa kuvaukset datajoukoista RDF-muodossa.

RDF-esimerkki: http://open-data.europa.eu/open-data/data/dataset/zOMHvCgeFXW4I8vnOlekyA.rdf.
Kuvaa yhden datajoukon, joka voidaan hakea pakattuna zip-tiedostona:
https://circabc.europa.eu/d/a/workspace/SpacesStore/e383ce72-ce09-408a-abe2-783d38b33d83/AgriculturalVegetableCatalogue.zip
Kuvattuun datajoukkoon näyttäisi olevan tarkoitus linkittää. Linkit ovat rikki, mikä tekee siitä 3 tähden dataa.
Kuvaus sisältää ontologian http://ec.europa.eu/open-data/ontologies/ec-odp käytön ja se käyttää myös Dublin Core -laajennusta.
Itse datajoukosta ei ole linkityksiä muualle. Siihen on rikkinäiset linkit.

Kaikki 11 RDF-datajoukkoa ovat samasta lähteestä, jokin eu virasto joka toimii ruokaturvallisuuden parissa. Kaikki data on .zip-tiedostoissa, mikä tarkasti ottaen tekisi tästä datajoukosta yhden tähden dataa.

CKAN-sovellus.

Yhteenvetona open-data.europa.eu on lähinnä eurostatin datajoukkojen varassa. Muista datajoukoista food&health RDF-datat ovat lupaavia, mutta koska ne eivät ole linkitettyä dataa, niin niille annetaan kolme tähteä. Portaaliin tulee kuitenkin jatkuvasti uusia datajoukkoja. Sitä seikkaa, että CKAN-sovellus tarjoaa RDF-muotoisia kuvauksia datajoukoista, ei lueta sivuston datajoukkojen laaduksi; nämä kuvaukset ovat neljän tähden dataa.

#03 PublicData.eu	
Open Knowledge Foundation	
publicdata.eu	
Julkaistu	6/2011 (Beta)
Datajoukkoja	Ainakin 16000
Dataa ennen julkaisua	On
Viimeisin päivitys	
Formaatit	Excel, CSV, RDF, TXT, HTML, PDF, ja monet muut formaatit
Keskimääräinen laatu	? (1)
Korkein laatu	***** (2)
*	On
**	On

***	On
****	On
*****	Ei
Lisenssi	License Not Specified Creative Commons Attribution UK Open Government Licence (OGL) [OpenData] etc...
SPARQL-rajapinta	Joissain seteissä.

(1) Datan laatu vaihtelee.
(2) RDF datajoukkoja Iso-Britanniasta, Italiasta ja Hollannista. Määrä kasvoi tutkimuksen aikana. Suurin osa on Iso-Britannian avoimen datan portaalista.

Kerää dataa eri lähteistä Euroopan Unionin alueella. EU-rahoitteinen: http://ckan.org/case-studies/publicdata-eu.

Linkit rikki, tulee http-autentikaatio haettaessa joitain datajoukkoja.
RDF tehdään joissain tapauksissa csv2rdf-muuntimella, joka tuottaa N-Triples -koodauksen. Linkit eivät näytä johtavan mihinkään.
Ei voida selvittää linkitystä rikki olemista RDF-settiin tai setistä.
Muutamia RDF-settejä tarjotaan SPARQL-solmun kauttaa, jolloin ne ovat linkitettäviä. Näitä on esimerkiksi Italian kohdalla; Italian avoin portaali jätettiin tutkimuksesta pois, sillä se oli erittäin vaikeasti arvioitavissa ja sisälsi paljon rikkinäisiä linkkejä.

Yhteenvetona: Tämä palvelu ei tarkoa omaa dataa vaan yrittää olla katalogi useaan eurooppalaiseen avoimen datan lähteeseen. Tämä palvelu koostaa siis dataa ja Euroopan Unionin alueella ja sen perusteella voisi tutkia datan laatua Euroopan Unionin alueella – jos se olisi siitä kattava oton, mitä se ei ole.

#04 Yhdistyneet Kansakunnat (YK)	
The United Nations Statistics Division (UNSD) of the Department of Economic and Social Affairs (DESA)	
data.un.org	
Julkaistu	2/2008
Datajoukkoja	34 databases - 60 million records
Viimeisin päivitys	1/2013
Formaatit	XML, CSV
Keskimääräinen laatu	***
Korkein laatu	***
*	On
**	Ei
***	On (*)
****	Ei
*****	Ei
Lisenssi	All data and metadata provided on UNdata’s website are available free of charge and may be copied freely, duplicated and further distributed provided that UNdata is cited as the reference.
SPARQL-rajapinta	Ei

“For the time being we are not planning to implement other web services or RDF format for download as we do not have enough capacity and resources for this.” Jokin oma sovellus. Sovelluksen avulla saa ladattua hakutuloksen CSV- ja XML-muodossa. Palvelussa ei siis ole kuin yhden tähden dataa. (*) Ladattua saa kolmen tähden dataa, mutta koska lataukseen ei ole suoraa linkkiä ja data pakataan zip-tiedostoksi, se ei tiukasti arvioiden ole kolmen tähden dataa.

#05 Yhdysvallat	
Hosted by the General Services Administration	
data.gov	
Julkaistu	5/2009
Datajoukkoja	378,529 raw and geospatial datasets. Iso osa linkkejä muualle.
Päivitettävä	Päivitetään.
Formaatit	CSV, Excel, XML, SHP, KLM.
Keskimääräinen laatu	***
Korkein laatu	***
*	On.
**	On.
***	On. (*)
****	On. (**)
*****	? (***)
Lisenssi	Lisenssitietoja ei sivustolla. Viitataan aineiston lähteeseen.
SPARQL -rajapinta	Ei.
Socrates. Lähinnä vain linkitettyjä datajoukkoja– 25 datajoukkoa itse Socrates-sovelluksessa. Geodatalle oma osio. (*) Paljon RDF/XML-muotoon talletettua semantista dataa ilman linkityksiä. (**) Linkitettynä Socraten kautta neljän tähden dataa. (***) Datajoukokja on valtava määrä ja siinä voi olla viiden tähden dataa. Monesti linkitettynä on RDF-dokumentti; mm. 2.3Gt työttymyysdata. Socrates voi tuottaa RDF-joukkoja, mutta itse Socratesissa on vain muutama kymmenen settiä. Näistä ei löytynyt ulos linkittäviä joukkoja.	

#06 Uusi-Seelanti	
Department of Internal Affairs.	
data.govt.nz	
julkaistu	10/2009
Datajoukkoja	2300
Päivitys	Päivitetään
Formaatit	API, CSV, Excel, PDF, XML, SHP, KML, Image
Keskimääräinen laatu	***
Korkein laatu	***

*	On.
**	On.
***	On.
****	Ei.
*****	Ei.
Lisenssi	Creative Commons Attribution 3.0 New Zealand licence ja erilaisia versioita CC-lisensseistä. Joitakin epäselviä lisenssejä.
SPARQL-rajapinta	Ei
Paljon karttakuvia ja karttatietoja (useita satoja datajoukkoja). Tilastodata PDF- ja Excel-muodossa (noin 100 settiä) Oma sovellus.	

#07 Tsekki	
Charles University in Prague; University of Economics, Prague	
opendata.cz	
Julkaistu	2011
Datajoukkoja	161
Formaatit	Excel, CSV, XML
Keskimääräinen laatu	**
Korkein laatu	***
*	On.
**	On.
***	On.
****	On. (*)
*****	Ei.
Lisenssi	Ei tietoa. [OpenData] logo sivuilla. Osassa silti “License Not Specified”.
SPARQL -rajapinta	On.
(*) Muutama SPARQL-kyseltävä RDF-dokumentti. CKAN-sovellus. On palveluna ckan.net osoitteessa. SPARQL-solmu löytyy opendata.cz sivulta. Kieli tsekki.	

#08 Iso-Britannia	
UK Government	
data.gov.uk	
Julkaistu	9/2009
Datajoukkoja	Yli 9500
Päivitys	Päivitetään.
Formaatit	CSV, XML, Excel, PDF, HTML, RDF

Keskimääräinen laatu	***
Korkein laatu	*****
*	On
**	On
***	On
****	Ei
*****	On
Lisenssi	UK Open Government Licence (OGL) 75% datajoukoista Muu lisenssi, joka rajoittaa käyttöä. 25% datasjoukoista.
SPARQL-rajapinta	Ei
RDF-data on organisaatioista ja näiden palkkatiedoista. Erityisesti henkilöt on pyritty kuvaamaan osin FOAF-skeeman mukaisin tiedoin. CKAN-sovellus, mutta portaalissa paljon muutakin sisältöä. Mm. sovelluskehitykselle on oma osio. Linked Data -linkki ei johda CKAN-sovelluksen Virtuoso -kyselylomakkeelle, vaan kertoo Linked Data-projektista. Sitä hoitaa oma työryhmä : http://data.gov.uk/linked-data/UKGovLD. Sivustolla luokitellaan viiden tähden kategorisoinnilla data; annetaan datajoukolle “Openness score”. Sen mukaan RDF-joukkoja on seuraavasti: nollan tähden joukkoja lisenssoinnin tai epämääräisten tietojen mukaan on 42 kappaletta, kolmen tähden RDF dataa 53 joukkoa, neljän tähden dataa ei ole ja viiden tähden joukkoja on 96 kappaletta.	

#09 Kanada	
?	
data.gc.ca	
Julkaistu	9/2012
Datajoukkoja	12764 (8800 tilastolaitokselta ja 3000 maatalousministeriöltä) Geospatiaalisia 260296
Päivitys	Päivitetään
Formaatit	CSV, XML, KML, RDF
Keskimääräinen laatu	***
Korkein laatu	****
*	On
**	On
***	On
****	On
*****	Ei
Lisenssi	Government of Canada Open Data Licence Agreement
SPARQL-rajapinta	Ei
Oma sovellus.	

Ainoa RDF-datajoukko on sanakirja-ontologia.

#10 Ranska	
Pääministeriö.	
data.gouv.fr	
Julkaistu	12/2011
Datajoukkoja	353288
Päivitys	Jatkuvaa
Formaatit	XLS (294 000 datajoukkoa), CSV, ODF, XML, SHP, RDF
Keskimääräinen laatu	**
Korkein laatu	*****
*	On
**	On
***	On
****	Ei
*****	On (*)
Lisenssi	Open Lisence
SPARQL-rajapinta	Ei. (**)
(*) Kansalliskirjasto tarjoaa yhden RDF-datajoukon. (**) http://data.bnf.fr/semanticweb on SPARQL-rajapinta, mutta se löytyy seuraamalla sivustolle linkitettyjä toisia palveluita. CKAN-sovelluksen kaltainen sovellus.	

#11 Norja	
?	
data.norge.no	
Julkaistu	4/2012
Datajoukkoja	Alle 100. Kaikki linkkejä.
Päivitys	Vähäistä ja epäsäännöllistä.
Formaatit	Linkit johtavat eri formaatteihin
Keskimääräinen laatu	? (*)
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei

Lisenssi	Norsk lisens for offentlige data (NLOD)
SPARQL-rajapinta	Ei.
(*) Sivustolta linkitetään toisiin palveluihin ja keskimääräistä laatua on vaikea arvioida. Tarkoituksena oli ottaa mallia Iso-Britannian palvelusta – mutta tavoitteesta jäädään kauas (http://data.norge.no/blogg/2010/09/reuse-of-public-sector-information-the-norwegian-story) Oma sovellus.	

#12 Ruotsi	
Yhden henkilön keräämä lista (Peter Krantz, freelance konsultti)	
opengov.se	
Julkaistu	?
Datajoukkoja	110 linkkiä.
Päivitys	?
Formaatit	Kaikki formaatit
Keskimääräinen laatu	?
Korkein laatu	*****
*	On
**	On
***	On
****	On
*****	On
Lisenssi	Aineistokohtainen
SPARQL-rajapinta	Ei.
On myös opendata.se, jossa on lueteltuina eri API -osoitteita. Vain vähän dataa. Ruosissa on valittu rajapintalähtöiset toteutukset.	

#13 Kenia	
Kenya ICT Board	
opendata.go.ke	
Julkaistu	7/2011
Datajoukkoja	534
Päivitys	Säännöllistä.
Formaatit	CSV, Excel, RDF, PDF
Keskimääräinen laatu	***
Korkein laatu	****
*	On.
**	Ei.

***	On.
****	On.
*****	Ei
Lisenssi	Public Domain (Creative Commons)
SPARQL -rajapinta	Ei
Socrates-sovellus. Socrates-sovellus toteutettu niin, että data-aineistoa on jopa neljän tähden laatuista. On PDF-dataa, mutta ei esimerkiksi Excel-tiedostoja jotka eivät olisi myös avoimessa formaatissa. Siksi ei ole kahden tähden dataa lainkaan. Socratesin luomat RDF-setit eivät sisällä linkityksiä muualle.	

#14 Australia	
The Department of Finance and Deregulation	
data.gov.au	
Julkaistu	3/2011
Datajoukkoja	1126
Päivitys	Päivitetään
Formaatit	CSV, Excel, SHP, ODT, PDF, HTML
Keskimääräinen laatu	***
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei
Lisenssi	Creative Commons - Attribution 3.0 Australia (CC BY 3.0)
SPARQL-rajapinta	Ei.
Sivustolla on myös erikseen linkkejä muihin sivustoihin, joista dataa voi löytyä. Jokin oma sovellus. Sivustolle on linkitetty joitakin HTML-sivuja, joissa on TABLE-elementeissä dataa. Tällainen data pitäisi tarkoitukseen tehdyllä sovelluksella tuoda sivulta, ja tällainen avoin data on kahden tähden dataa. Sitä on sivuilla hyvin vähän. Muista sivuista poiketen Excel-muodossa ei ole lainakaan dataa, joka ei olisi myös CSV-muodossa. Tarkkaan ottaen nämä ovat linkkejä, ja sivustolla ei olisi tällöin lainkaan kahden tähden dataa.	

#15 Alankomaat
?
data.overheid.nl

Julkaistu	10/2011
Datajoukkoja	5193
Päivitys	Jatkuvasti
Formaatit	CSV, RDF, PDF, RDF
Keskimääräinen laatu	***
Korkein laatu	****
*	On
**	On
***	On
****	On
*****	Ei
Lisenssi	Public Domain.
SPARQL-rajapinta	On
Käyttää tähtimittaria aineiston kategorisoinnissa. Linked Open Data Sterren *** (In een open formaat) (5059) ** (Gestructureerde data) (17) **** (RDF en SPARQL) (12) * (Online; open licentie) (4) Hollannin kielellä. CKAN-sovellus.	

#16 Espanja	
The Secretary of State for Telecommunications and the Information Society (SETSI), which is apart of the Ministry of Industry, Energy and Tourism, is responsible for directly managing the service.	
datos.gob.es	
Julkaistu	10/2011
Datajoukkoja	640
Päivitys	Jatkuvasti
Formaatit	HTML, CSV, PDF, Excel, SHP
Keskimääräinen laatu	*
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei
Lisenssi	Public Domain
SPARQL-rajapinta	Ei

Paljon HTML- ja XHTML-sisältöä. HTML-sisältö voi tarkoittaa linkkiä sivulle, josta saa ZIP-tiedoston, jonka sisällä on XLS-tiedostoja. Jokin oma sovellus.

#17 Brasilia	
?	
dados.gov.br	
Julkaistu	5/2012
Datajoukkoja	100
Päivitys	Ei selviä.
Formaatit	XML, PDF, CSV
Keskimääräinen laatu	*
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei
Lisenssi	OpenData, mutta suurin osa dataa, jolla ei ole avoin lisenssi
SPARQL-rajapinta	Ei
Tiedot voivat vääristyä kielemuurin vuoksi. Sivusto portugaliksi. CKAN-sovellus.	

#18 Argentiina, Buenos Aires	
?	
data.buenosaires.gob.ar	
Julkaistu	?/2012
Datajoukkoja	80
Päivitys	Päivitetään
Formaatit	CSV, PDF, Excel
Keskimääräinen laatu	***
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei

Lisenssi	[OpenData] logo sivulla.
SPARQL-rajapinta	Ei
Argentiinan valtiolla ei ole avoimen datan sivustoa, arvioitavana on Buenos Airesin sivusto. CKAN-sovellus.	

#19 Intia	
National Informatics Centre (NIC) , DeitY, MoCIT Government of India.	
data.gov.in	
Julkaistu	10/2012
Datajoukkoja	115
Päivitys	Päivitetään
Formaatit	Excel, CSV, HTML
Keskimääräinen laatu	***
Korkein laatu	***
*	On
**	On
***	On
****	Ei
*****	Ei
Lisenssi	Ei mainita sivuilla.
SPARQL-rajapinta	Ei
Oma sovellus.	

#20 Singapore	
?	
data.gov.sg	
Julkaistu	6/2011
Datajoukkoja	7937
Päivitys	Päivitetään.
Formaatit	Excel, DOCX, PDF, SHP, KML, CSV
Keskimääräinen laatu	**
Korkein laatu	***
*	On (*)
**	On (*)
***	On (*)
****	Ei

*****	Ei
Lisenssi	Register for commercial use... Muutenkin erittäin haastavat ehdot. On lakattava käyttämästä datajoukkoja, jos ne poistuvat kyseiseltä sivustolta.
SPARQL-rajapinta	Ei
(*) Vain jos lisenssiehto jotenkin olisi tulkittavissa avoimeksi, mitä se ei todennäköisesti ole. 2800 datajoukkoa tilastolaitoksen Excel-tiedostoja. Paljon linkkejä. Jokin oma sovellus.	