

hyväksymispäivä arvosana

arvostelija

Semanttisen webin tekniikoiden soveltaminen Wikisovelluksissa

Olli Lyytinen

Helsinki 31.3.2008

Pro Gradu

HELSINGIN YLIOPISTO

Tietojenkäsittelytieteen laitos

Tiedekunta/Osasto — Fakultet/Sektion — Faculty		Laitos — Institution — Department	
Matemaattis-luonnontieteellinen		Tietojenkäsittelytieteen laitos	
Tekijä — Författare — Author Olli Lyytinen			
Työn nimi — Arbetets titel — Title Semanttisen webin tekniikoiden soveltaminen Wikisovelluksissa			
Oppiaine — Läroämne — Subject			
Työn laji — Arbetets art — Level Pro Gradu	Aika — Datum — Month and year 31.3.2008	Sivumäärä — Sidoantal — Number of pages 51	
Tiivistelmä — Referat — Abstract Perinteistä Internetiä on viime aikoina kehitetty kahteen eri suuntaan. Erityisesti W3-konsortio ajaa niin sanottua semanttista webbiä, toista kehityssuuntausta kutsutaan usein nimellä Web 2.0. Tässä tutkielmassa tarkastellaan, miten näiden kahden suuntauksen menetelmiä voidaan käyttää yhdessä wikijärjestelmissä. Erityisesti tarkastellaan yhtä netin suosituimmista sivustoista, avointa tietosanakirjaa Wikipediää, joka edustaa Web 2.0 suuntausta. Semanttisen webin tekniikoiden soveltaminen toisi lukuisia hyödyllisiä lisäominaisuuksia Wikipediaan, esimerkiksi paremman hakuliittymän ja mahdollisuuden käyttää tietoa uudelleen niin Wikipedian sisällä kuin muissakin järjestelmissä. Tutkielmassa testattiin erilaisia tapoja liittää semanttisessa webissä käytettävää RDF-muotoista metatietoa Wikipediaan. Testauksessa käytettiin saatavilla olevia valmiita semanttisia järjestelmiä ja Wikipedian todellista aineistoa. Testeissä havaittiin, että järjestelmissä on vielä lukuisia puutteita ja Wikipedian nykyinen rakenne ja käytännöt eivät täysin ole yhteensopivia semanttisten wikijärjestelmien kanssa. Tästä huolimatta testattujen järjestelmien integroiminen Wikipediaan olisi jo nyt mahdollista ja toisi joitakin hyödyllisiä lisäominaisuuksia Wikipediaan.			
Avainsanat — Nyckelord — Keywords semanttinen web, web 2.0, wiki, Wikipedia			
Säilytyspaikka — Förvaringsställe — Where deposited			
Muita tietoja — Övriga uppgifter — Additional information			

Sisältö

1	Johdanto	1
2	Wikisovellukset ja Wikipedia	3
3	Semanttinen web	7
3.1	Perusidea	8
3.2	RDF-tietomalli	9
3.3	Ontologiat ja RDF:n laajennokset	11
4	Semanttinen wiki	13
4.1	Toteutus ja artikkelien editointi	14
4.2	Hyöty Wikipedian kannalta	15
5	Tiedon louhinta Wikipediasta	17
6	Semantiikan lisääminen Wikipediaan	21
7	Semanttisten wikijärjestelmien testaus	24
7.1	Testimenetelmät	24
7.2	Yleiset havainnot	25
7.3	Monipaikkaiset relaatiot ja taulukot	29
7.4	Subjektin paikallisuus	33
7.5	Wikipedian kategoriat	35
7.6	Ontologioihin liittyvät havainnot	36
8	Tiedonlouhintamenetelmien testaus	38
8.1	Testimenetelmät	38
8.2	Havainnot DBPediasta	40
8.3	Havainnot YAGO:sta	44
9	Johtopäätökset	45

1 Johdanto

Viime vuosien aikana yhdeksi Internetin suosituimmista sivustoista on noussut avoin tietosanakirja Wikipedia. Wikipedia, kaikkien muiden wikisovellusten tapaan, sallii kenen tahansa muuttaa sivujen sisältöä ja luoda uusia sivuja. Voidaan pitää hie-
man yllättävänä, että näin avoimella järjestelmällä on pystytty tuottamaan kattava tietosanakirja, joka haastaa perinteiset painetut ja verkkotietosanakirjat. Wikipediassa on toki ongelmia, ja sitä kritisoidaankin usein esimerkiksi juuri tiedon luotettavuuden takia. Joka tapauksessa Wikipedia sisältää huomattavan määrän avointa hyödyllistä tietoa.

Wikipedian tieto on avointa, ja se on käytettävissä lisenssillä, joka muistuttaa avoimen ohjelmistokehityksen lisensoijaa. Tämä tarkoittaa, että tietoa voidaan hyödyntää muissa sovelluksissa. Tällä hetkellä on kehitteillä, ja osittain tuotantokäytössäkin, kaksi suuntausta, joissa tiedon uudelleenkäyttöä Internetissä pyritään edistämään. Semanttinen web [BHL01] on W3-konsortion (W3C) visio siitä, miten Internetiä tulisi kehittää. Semanttisessa webissä pääpainona on annotoida sisältöä, eli lisätä tekstiin sellaisia semanttisia lisämääreitä, että koneellisesti on mahdollista ymmärtää tai jäsentää, mitä tekstissä sanotaan. Kun tietokone ymmärtää tekstin sisällön, voidaan tietoa hyödyntää uusilla tavoilla, myös sovelluksissa, joihin tietoa ei alun perin oltu suunniteltu. Toinen suuntaus, Web 2.0, ei ole minkään yksittäisen tahon kehittämä standardi tai tuote. Se on vain yleinen nimitys nykyaikaisille verkkosovelluksille, joissa korostetaan käyttäjien tuottamaa sisältöä, yhteisöllisyyttä sekä tiedon uudelleen käytettävyyttä [Gre07].

Wikipedia on hyvä esimerkki Web 2.0 -sovelluksesta. Sillä on suuri joukko käyttäjiä, jotka kokonaisuudessaan muodostavat yhteisön, mutta jossa on myös pienempiä tiettyihin aihealueisiin keskittyneitä tiiviimpiä yhteisöjä. Kaikki Wikipedian sisältö on käyttäjien tuottamaa. Wikipedian sisältämä tieto on myös käytettävissä sen ulkopuolella. Wikipedian artikkelit on saatavilla kokonaisen tietokantakopion lisäksi XML-muodossa sivu kerrallaan. Verrattuna useimpiin muihin Web 2.0 -sovelluksiin Wikipedia on kuitenkin huomattavan laaja. Wikipedia on tietosanakirja, jossa artikkelit voivat käsitellä mitä tahansa. Tyypillisesti muissa Web 2.0 -sovelluksissa keskitytään johonkin suppeampaan kokonaisuuteen, esimerkiksi jaetaan valokuvia tai videoleikkeitä, arvostellaan verkkosivustoja ja esitellään nettikirjanmerkkejä tai julkaistaan lyhyehköjä mielipidekirjoituksia (blogit).

Viime aikoina on esitetty, miten semanttinen web ja Web 2.0 voisivat toimia yh-

dessä sen sijaan, että ne nähtäisiin kilpailevina tekniikoina. Onkin esitetty sellaisia nimityksiä kuin semanttinen web 2.0 [Gre07] ja web 3.0 [LaH07]. Tekniikoiden yhdistämisessä on molemminpuolisia etuja. Semanttisen webin täsmällinen tietomalli tuntuu luonnolliselta jatkeelta Web 2.0:n hieman epämääräisemmille ja kevyemmille folksonomioille, mikroformaateille ja merkkauksille. Yhteinen tietomalli myös tekisi Web 2.0 -sovellusten yhteiskäytön helpommaksi. Toisaalta Web 2.0 saattaa tarjota semanttiselle webille hyvän alustan levitä laajempaan käyttöön. Web 2.0 -sovellukset ovat osoittaneet, että käyttäjät kyllä haluavat lisätä metatietoa sisältöön, mutta työkalujen on oltava riittävän helppokäyttöisiä [SSR06].

Web 2.0:n ja semanttisen webin yhdistäviä sovelluksia onkin jo kehitetty, esimerkiksi juuri wikijärjestelmiä, joissa artikkelien sekaan lisätään semanttista metatietoa. Erityisesti Wikipedia on tässä yhteydessä mielenkiintoinen tapaus. Sen lisäksi, että semantiikan lisääminen Wikipediaan parantaisi Wikipedian käytettävyyttä, avaisi se valtavan avoimen semanttisen tietovaraston muiden sovellusten käytettäväksi. Semantiikan lisäämisessä on kuitenkin suuri työ. Wikipedian käyttäjäkunta tarjoaa valmiin yhteisön, jolla on intressi tehdä tämä työ, mikäli se todella parantaa Wikipedian laatua ja käytettävyyttä huomattavasti.

Tässä tutkielmassa tutkitaan nykyisten semanttisten wikijärjestelmien toimivuutta ja sitä, miten ne soveltuisivat käytettäväksi Wikipediassa. Käytännössä vain yksi tutkituista wikijärjestelmistä, Semantic MediaWiki [Völ06], on realistisesti suoraan käytettävissä Wikipediassa. Se on alusta alkaen suunniteltu laajennokseksi MediaWikiin, järjestelmään, jota Wikipedia käyttää. Muut semanttiset wikijärjestelmät kuitenkin saattavat tarjota ideoita ja ratkaisuja ongelmiin, joita voidaan soveltaa Semantic MediaWikissä tai muissa wikijärjestelmissä.

Tutkielmassa testataan järjestelmiä syöttämällä niihin todellista Wikipediasta kopioitua sisältöä ja lisäämällä siihen semanttista metatietoa, eli annotoimalla, käyttäen järjestelmän tarjoamia työkaluja. Testauksessa käytetään sekä englannin- että suomenkielistä sisältöä. Testauksessa kiinnitetään huomiota järjestelmien tarjoamiin lisäominaisuuksiin nykyiseen Wikipediaan verrattuna, työkalujen helppokäyttöisyyteen, semantiikan rikkauteen sekä menetelmien soveltuvuuteen Wikipediassa. Tutkielman lopussa analysoidaan järjestelmistä löydetty ongelmat ja esitetään mahdollisia ratkaisuja.

Tutkielmassa testataan lisäksi järjestelmiä, jotka tulkitsevat Wikipedian nykyistä sisältöä ja automaattisesti lisäävät siihen semanttista metatietoa. Tällaisilla järjestelmillä Wikipediasta olisi mahdollista saada nopeasti semanttinen versio, jota

voitaisiin hyödyntää muiden sovellusten kanssa. Toisaalta vastaavilla menetelmillä voitaisiin lisätä semantiikkaa nopeasti siirtymävaiheessa, kun Wikipediassa otetaan käyttöön jokin semanttinen wikijärjestelmä. Koska järjestelmät toimivat automaattisesti, arvioidaan näitä järjestelmiä lähinnä vain annotoinnin laadun suhteen.

Luvussa 2 esitellään tarkemmin käsite wiki sekä Wikipedia ja perehdytään Wikipedian käyttäjien toimintamalleihin. Luvussa 3 esitellään semanttinen web sekä sen yhteydessä yleisesti käytetty RDF-tietomalli ja semanttiset ontologiat. Luvussa 4 esitellään, mitä tarkoitetaan semanttisella wikillä tai semanttisella Wikipedialla. Tämä luku antaa yleiskuvan myöhemmin testattavista järjestelmistä sekä esittelee joitakin selkeitä semantiikan tuomia lisäetuja Wikipedian kannalta, mikä toimii motivaationa ylipäänsä tuottaa tällaisia järjestelmiä. Luku 5 esittelee sellaiset testattavat sovellukset, jotka automaattisesti louhivat tietoa Wikipediasta ja muodostavat siitä semanttisen version. Luku 6 esittelee sellaiset testattavat sovellukset joilla wikin, tai Wikipedian, artikkeleita annotoidaan manuaalisesti. Luku 7 käsittelee manuaalisesti annotoitavien wikien testauksen. Luvussa esitellään testausmenetelmät sekä testitulokset ja käsitellään löydettyjä ongelmia. Luku 8 käsittelee automaattisesti tietoa louhivien järjestelmien testausta. Lopuksi luvussa 9 esitellään johtopäätökset.

2 Wikisovellukset ja Wikipedia

Sana wiki tulee ensimmäisestä wikisivustosta nimeltä WikiWikiWeb ¹. Tämän sivuston käynnisti Ward Cunningham vuonna 1995. WikiWikiWebissä kuka tahansa pystyy helposti muokkaamaan sivuja sekä lisäämään ja poistamaan tietoa. Järjestelmän avoimuus luonnollisesti tekee mahdolliseksi sen, että kuka tahansa voi myös tuhota tietoa. Järjestelmä on kuitenkin tehty siten, että töhritty tai poistettu sivu on helppoa palauttaa alkuperäiseen muotoonsa [WiS05].

Nykyisin on saatavilla useita erilaisia wikisovelluksia eli palvelinohjelmistoja, jotka tarjoavat wikijärjestelmän. Perusajatus on kaikissa edelleen sama, eli kuka tahansa voi muokata ja tuottaa sisältöä. Eräs yleisesti käytetty wikisovellus on MediaWiki. Tämä, kuten useimmat muutkin nykyiset wikit, tarjoaa joitakin lisäominaisuuksia verrattuna ensimmäiseen WikiWikiWebiin. Esimerkiksi jokaiseen sivuun liittyy erillinen keskustelusivu, jossa käyttäjät voivat käydä keskustelua sivun kehittämisestä. Muita lisäyksiä ovat muun muassa uudelleenohjaussivut, sivujen luokittelu kategorioihin, erilaisten sivupohjien käyttäminen, paremmat mahdollisuudet muotoilla

¹<http://www.c2.com/cgi/wiki?WelcomeVisitors>

'''Helsinki''' ({{k-sv|Helsingfors}}) on [[Suomi|Suomen]] [[pääkaupunki]]. Se sijaitsee [[Etelä-Suomen lääni|Etelä-Suomessa]], [[Suomenlahti|Suomenlahden]] rannalla. Helsingin rajanaapureita ovat pohjoisessa ja idässä [[Vantaa]] ja lännessä [[Espoo]], joiden lisäksi Helsingillä on merirajaa [[Sipoo|Sipoon]] kanssa. Helsingissä on noin 570 000 asukasta ja se on noin 1 000 000 asukkaan [[pääkaupunkiseutu|pääkaupunkiseudun]] keskuskaupunki.

Kuva 1: Osa Helsinkiä käsittelevän artikkelin lähdekoodista suomenkielisessä Wikipediassa.

tekstiä.

Internetissä on lukuisia erilaisia wikisivustoja. Yleensä sivustot keskittyvät johonkin tiettyyn aihealueeseen. Alkuperäinen WikiWikiWeb keskittyy toimintamalleihin ohjelmistokehityksessä, muina esimerkkeinä mainittakoon Tähtien sota -wiki ², Go-lautapeli aiheinen wiki ³ ja Raamattu-wiki ⁴. Kiistatta tunnetuin wiki on kuitenkin Wikipedia ⁵. Itseasiassa se on yleisestikin eräs käytetyimmistä sivustoista Internetissä. Alexa.com-seurantasivuston mukaan Wikipedia on vuoden 2007 aikana noussut 10 käytetyimmän internetsivuston joukkoon.

Wikisivusto, tai pelkästään wiki, koostuu joukosta sivuja tai artikkeleita. Kullakin sivulla on otsikko, joka myös toimii sivun nimenä ja tunnisteena. Sivujen nimet siis ovat yksilöllisiä, ja tarvittaessa niissä on käytettävä tarkenteita erottamaan muuten samannimiset käsitteet toisistaan. Sivun sisältö on esitetty wikimerkkauksielellä. Tämän syntaksi vaihtelee hieman wikijärjestelmien välillä. Kuvassa 1 on esitetty esimerkki Helsinkiä käsittelevästä Wikipedian artikkelista.

Wikipedian käyttämässä MediaWiki-järjestelmässä linkit toisille sivuille merkataan asettamalla kahdet hakasulkeet artikkelin nimen ympärille. Hakasulkeiden väliin jäävä teksti, eli artikkelin nimi, näkyy käyttäjälle. Mikäli käyttäjälle halutaan näkyvän jotakin muuta kuin artikkelin nimi, voidaan haluttu teksti antaa artikkelin nimen jälkeen erottamalla se pystyviivalla. Tätä käytetään erityisesti suomenkielisessä Wikipediassa kun perusmuotoiset artikkelin nimet on taivutettava sopivaan muotoon lauseessa. Merkkauksielellä on myös mahdollista tuottaa taulukoita, listoja ja muut-

²<http://starwars.wikia.com>

³<http://senseis.xmp.net>

⁴<http://bible.tmtm.com>

⁵<http://www.wikipedia.org>

taa tekstin fonttia. Esimerkiksi kuvassa 1, kolme puolilainausmerkkiä sanan Helsinki ympärillä muuttaa fontin lihavoiduksi.

Wikipedian itsensä mukaan [Wik07a] se on “monikielinen, verkkotietosanakirjaprojekti, jota hallinnoi epäkaupallinen Wikimedia säätiö.” Nimi Wikipedia kuvastaa hyvin sivuston perusideaa: wiki ensyklopedia eli wiki tietosanakirja. Sivusto siis toimii wikijärjestelmällä, nimenomaan juuri Wikipediaa varten kehitetyllä MediaWikillä. Monikielisyys tarkoittaa sitä, että tietosanakirjasta on useita eri versioita eri kielillä. Englanninkielinen versio on selvästi suurin ja käsittää yli kaksi miljoonaa artikkelia [Wik07b]. Suomenkielinen versio puolestaan käsittää noin 140 000 artikkelia [Wik07c].

Wikipediaa edelsi projekti nimeltä Nupedia, jonka tavoitteena oli olla avoin verkkotietosanakirja. Erona Wikipediaan Nupediassa sisällön tuotanto oli formaalimpaa ja perustui akateemisten julkaisujen toimintamalliin [Lih04]. Projekti ei kuitenkaan lähtenyt kunnolla käyntiin. Nupedian pohjalta kehitettiin avoimempi nykyinen Wikipedia. Epäilemättä juuri avoimuus on yksi oleellisista ominaisuuksista, joka mahdollisti Wikipedian kasvun.

Wikipedian avoimuus on kuitenkin myös yksi sen suurimpia ongelmia. Jos kuka tahansa voi muokata sivuja, miten sivujen sisältöön voidaan luottaa? Wikipediaa kritisoidaan usein juuri tämän takia. Esimerkiksi niin sanotussa Seigenthaler-tapauksessa yhdysvaltalaisen journalistin John Seigenthalerin elämäkertaa Wikipediassa muutettiin pilailu mielessä niin, että Seigenthaler olisi ollut epäilytynä Kennedyn murhaan [Sei05]. Tämä tieto oli Wikipediassa neljä kuukautta ennen kuin Seigenthalerin ystävä havaitsi virheen. Uutisissa on myös raportoitu, miten poliittiset puolueet, firmat ja esimerkiksi Yhdysvaltain tiedustelupalvelu CIA ovat muokanneet Wikipedian sivuja parantaakseen omaa julkisuuskuvansa [Käh07]. Wikipediassa tällaista perusteetonta sivujen muokkausta tai tuhoamista kutsutaan yleisesti vandalismiksi.

Wikipedian tiedon luotettavuuteen liittyen on myös tehty joitakin tutkimuksia. Ehkä tunnetuin tapaus on Nature-lehden tutkimus, joka hieman yllättäen totesi, että Wikipedia sisältää vain hieman enemmän virheitä kuin Encyclopedia Britannican verkkoversio [Gil05]. Britannican vastine tosin kyseenalaisti Naturen tutkimusmenetelmät [Bri06]. Toisaalta Wikipedian artikkelien on todettu olevan sanamuodoiltaan varsin samankaltaisia perinteisten tietosanakirjojen kanssa, mikä viittaa asialliseen sisältöön [WiS05]:stä. Tutkimuksessa ei kuitenkaan kiinnitetty mitään huomiota varsinaiseen tietosisältöön.

Tiedon mahdollisesta vääristelystä huolimatta Wikipedia sisältää huomattavan mää-

rän hyödyllistä tietoa ja sillä on selkeä paikka tiedon esilletuomisessa. Lih kuvaa artikkelissaan [Lih04] kuinka Wikipedia täyttää historiallisen “tietokuilun.” Tietokuilu on se aikaväli, joka on uutisten ja yleistiedon välissä. Jonkin tapahtuman siirtyminen uutisesta historiankirjoihin kestää vuosia ja tällä välillä tapahtumaan ei viitata juuri missään. Perinteiset tietosanakirjat ovat täyttäneet tätä kuilua hieman, mutta jopa niiden verkkoversioista julkaistaan päivityksiä tyypillisesti vain vuoden tai puolen vuoden välein. Wikipedia sen sijaan dokumentoi tapahtumat välittömästi ja yleisö pääsee lukemaan niistä milloin vain. Useat sanomalehtien verkkosivut ovatkin alkaneet viittaamaan Wikipediaan uutistensa yhteydessä [Lih04].

Yksinkertaisesta perusideastaan huolimatta Wikipedia on nykyisin varsin organisoinut. Tämä näkyy esimerkiksi eri nimiavaruuksien suhteellisissa kasvuissa [Vie07]. Wikipedian artikkelit jaetaan eri nimiavaruuksiin sen mukaan, mitä ne käsittelevät. Normaalit tietosanakirja-artikkelit ovat main-nimiavaruudessa. Näihin liittyvät keskustelusivut talk-nimiavaruudessa. Käyttäjien kotisivut ovat nimiavaruudessa user ja näiden keskustelusivut nimiavaruudessa user talk. Wikipediaa koskevat ohjeistukset ja käytännöt ovat wikipedia-nimiavaruudessa ja myös näille on omat keskustelusivunsa nimiavaruudessa wikipedia talk. Kuvat ja muut multimediatiedostot ovat image-nimiavaruudessa. Lisäksi on joitakin pienempiä nimiavaruuksia.

Taulukossa 1 on esitetty eri nimiavaruuksien sisältämien artikkelien lukumäärät toukokuussa 2003 ja lokakuussa 2005 sekä lukumäärien suhteellinen kasvu. Huomionarvoista on, että varsinaiset tietosanakirja-artikkelit ovat suhteessa kasvaneet kaikkein vähiten. Talk-nimiavaruuden artikkelien sisältöä on tutkittu tarkemminkin [Vie07]. Kukin tämän nimiavaruuden keskustelusivu liittyy kiinteästi johonkin varsinaiseen main-nimiavaruuden artikkeliin, ja niiden sisältö koostuu käyttäjien jättämistä kommenteista. Suuri osa kommenteista on keskustelua siitä, miten vastata tietosanakirja-artikkelia voitaisiin parantaa. Keskustelusuilla myös ratkotaan erimielisyyksiä artikkelien sisällöstä. Tässä yhteydessä usein viitataan Wikipedian virallisiin ohjeistuksiin ja käytäntöihin wikipedia-nimiavaruudessa.

Käyttäjien keskustelusivut ovat samankaltaisia kuin artikkelien keskustelusivut, kommentit ovat vain henkilökohtaisemmin osoitettu juuri kyseiselle käyttäjälle. Kommentit koskevat usein käyttäjän tekemiä yksittäisiä muokkauksia, ovat ohjeistuksia Wikipedian käytännöistä käyttäjälle, tai esimerkiksi huomautuksia johonkin artikkeliin suunnitelluista suurista muutoksista. Nimiavaruuksien suhteelliset kasvut selvästi kertovat, miten käyttäjät panostavat keskinäiseen koordinointiin artikkelien ylläpidossa. Lisäksi wikipedia-nimiavaruuden suuri kasvu kertoo koko Wikipediaa

Nimiavaruus	2003	2005	Kasvukerroin
Main	170 369	1 531 095	9
Talk	20 067	229 999	11
User	3 324	76 491	23
User talk	2 564	199 264	78
Wikipedia	1 211	81 738	68
Wikipedia talk	441	7 267	16
Image	6 970	292 451	42

Taulukko 1: Englanninkielisen Wikipedian eri nimiavaruuksien artikkelien lukumäärät toukokuussa 2003 ja lokakuussa 2005 [Vie07].

koskevien suurien muutosten koordinoinnista. On siis perusteltua olettaa, että Wikipedian käyttäjät haluavat kehittää Wikipediaa ja mahdollisesti ottaa käyttöön semanttisen laajennoksen ja että heillä on kyky organisoidusti ratkoa tähän liittyviä ongelmia.

Myös Wikipedian käyttäjien toimintatapoja on tutkittu [BFB05]. Haastattelututkimuksessa selvisi, että usein käyttäjät ovat alun perin löytäneet Wikipedian normaalilla internethakukoneella, joka on antanut tuloksena jonkin Wikipedian artikkelin. Muuttuminen pelkästä tiedonkuluttajasta tuottajaksi tapahtuu tekemällä aluksi pieniä korjauksia olemassa oleviin artikkeleihin. Vähitellen ruvetaan tuottamaan enemmän omaa sisältöä ja kokonaan uusia artikkeleita. Alussa editoinnit keskittyvät käyttäjää itseään kiinnostaviin artikkeleihin, eikä tällöin yleensä olla tietoisia Wikipedian erilaisista virallisista ohjeistuksista tai keskustelusivuista. Aloittelijat eivät myöskään täysin tunne merkkauskieltä, jota käytetään artikkelien editoinnissa. Vähitellen käyttäjät löytävät keskustelusivut ja Wikipedian ohjeistukset. Samalla huomio siirtyy yksittäisistä artikkeleista Wikipediaa koskeviin suurempiin linjoihin. Wikipedian äärimmäisen pieni kynnys siirtyä tiedon kuluttajasta tuottajaksi siis näyttäisi houkuttelevan käyttäjiä. Niinpä mikä tahansa Wikipedian laajennos tulisikin suunnitella siten, että tämä prosessi ei häiriinny.

3 Semanttinen web

Tässä luvussa esitellään semanttinen web. Ensimmäisessä aliluvussa kuvaillaan semanttisen webin perusidea. Toisessa aliluvussa esitellään W3C:n suosittelema se-

manttisen webin tietomalli RDF. Kolmannessa aliluvussa esitellään RDF:n laajennokset RDFS ja OWL, joilla voidaan määritellä ontologioita, eräänlaista perustietämystä.

3.1 Perusidea

Semanttinen web on World Wide Web Konsortion (W3C) idea siitä, miten Internetiä tulisi kehittää eteenpäin. Tim Berners-Lee et. al. esittelivät perusidean vuonna 2001 [BHL01]. Perinteisesti sisältö Internetissä on tuotettu ihmisille, jotka lukevat sen ruudulta tai paperilta. Tietokone ei ymmärrä mitään muuta kuin sen, miten teksti ladotaan ruudulle. Semanttisen webin tarkoituksena on muuntaa sisältö sellaiseen muotoon, että tietokone voi ymmärtää sisällön tai ainakin jäsentää sen oleelliset osat.

Tekstisisällön saaminen tietokoneen ymmärtämään muotoon mahdollistaa nykyistä monipuolisemmat haut ja älykkäiden agenttien toimimisen tiedon perusteella. Tim Berners-Lee et. al. [BHL01] menevät jopa niin pitkälle, että kuvaavat tulevaisuuden järjestelmän, jossa itsenäinen agentti, jonkinlainen henkilökohtainen tekoälyavustaja, automaattisesti varaa henkilölle säännölliset lääkärin tapaamiset. Agentti ottaa huomioon henkilöiden kalentereihin merkatut tapaamiset, potilaan vakuutusehdot, potilaan ja tämän perheen asuntojen sijainnit suhteessa lääkärin vastaanottoon jne. Järjestelmä toimii, koska se ymmärtää mitä henkilön kalenterin tiedot tarkoittavat, mitä lääkäreiden kotisivuilla olevat tiedot tarkoittavat, mitä vakuutusyhtiön kotisivulla olevat tiedot tarkoittavat jne. Näiden tietojen perusteella agentti voi valita sopivan lääkärin samaan tapaan kuin ihminen tekisi. Agentti vain tekee sen automaattisesti.

Käytännössä semanttisen webin järjestelmät ovat vielä toistaiseksi olleet hieman vaatimattomampia. Semanttisen webin tekniikoiden tuoma etu näkyy usein parhaimmin parempina hakumahdollisuuksina. Esimerkiksi moninäkömahaku [MHS06], joka vie perinteisen hierarkkisen luokittelun sekä avainsanahaun askeleen eteenpäin. Käyttäjälle esitetään aineisto luokiteltuna useiden eri ominaisuuksien suhteen, siis useita eri näkymiä aineistoon. Käyttäjä voi rajata hakua useiden eri näkymien suhteen yhtä aikaa.

Kalfoglou et. al. [Kal04] esittää hyvää kritiikkiä semanttisesta webistä. Ongelmia ja ratkaisemattomia asioita liittyy lähes kaikkiin semanttisen webin osa-alueisiin. Miten hoidetaan luottamus eri tietolähteisiin ja muihin agentteihin verkossa ja yksityisyyden suoja? Mitä itseasiassa semanttisen webin yhteydessä täsmälleen tarkoi-

tetaan agentilla ja miten nämä toimivat yhdessä? Miten hoidetaan eri tietolähteissä käytettyjen erilaisten ontologioiden yhteensovitus? Skaalautuvatko menetelmät suureen mittakaavaan? Artikkelin johtopäätöksissä esitetään, että ennen kuin yritetään rakennetaan koko Internetin kattava semanttinen web, olisi hyvä muodostaa pienempiä semanttisen webin ideoita hyödyntäviä järjestelmiä. Näistä pienemmistä järjestelmistä yritetään oppia ja myöhemmin ne voidaan mahdollisesti yhdistää isommaksi kokonaisuudeksi. Tässäkin mielessä semanttisen webin testaaminen Wikipediassa on mielenkiintoista ja voi mahdollisesti tarjota hyödyllisiä kokemuksia semanttisen webin jatkokehitystä varten.

3.2 RDF-tietomalli

Semanttisen webin tarkoituksena on esittää tieto sellaisessa muodossa, että tietokone voi ymmärtää sen tai ainakin jäsentää tiedon. Tätä varten on kehitetty erilaisia tietomalleja. W3C:n suositus on Resource Description Framework (RDF) -metakieli [W3C04a]. Toinen standardoitu tietomalli on aihekartat [GaG06], joka suurelta osin vastaa RDF:ää hyvin läheisesti. Tässä tutkielmassa käytetään RDF-mallia, mutta menetelmät toimisivat pienin muutoksin myös esimerkiksi aihekartoilla. RDF:llä on erilaisia esitysmuotoja, yleisin lienee XML-esitysmuoto. Esitysmuotoa tärkeämpi asia kielessä on kuitenkin sen tietomalli, joka on pääpiirteiltään varsin yksinkertainen.

RDF-tietämys koostuu joukosta järjestettyjä *kolmikoita*. Kolmikon jäseniä kutsutaan nimillä *subjekti*, *predikaatti* ja *objekti*. Jäsenet voivat olla joko *URI-viittauksia* tai objektin tapauksessa myös *literaaleja*. URI-viittauksella tarkoitetaan jotakin *resurssia*, joka identifioidaan URI-tunnisteella (Uniform Resource Identifier). Literaali on vapaamuotoinen teksti, jolle voidaan halutessa määritellä kieli ja tietotyyppi. Kolmikoita kutsutaan myös *RDF-lauseiksi* tai pelkästään lauseiksi.

Jokaista URI-viittausta vastaa jokin käsite. Tietomallin ulkopuolella on määriteltävä, mitä eri URI:t eli resurssit tarkoittavat. Esimerkiksi voidaan käyttää URI-viittausta "http://www.esimerkki.fi#koira", joka sovitaan tarkoittavan käsitettä koira. Tässä tutkielmassa RDF-kolmikot esitetään muodossa (koira, on, eläin). Tässä kolme resurssia ovat järjestyksessä subjekti, objekti ja predikaatti. Tutkielmassa resurssit esitetään yksinkertaisuuden vuoksi kuvaavilla yleiskielisillä ilmauksilla. Todellisuudessa jokainen on URI-viittaus, jolle on sovittu jokin semantiikka. Esimerkissä predikaattina oleva URI-viittaus tarkoittaa kuulumista johonkin luokkaan ja objekti tarkoittaa käsitettä eläin. RDF-kolmikon semanttinen merkitys selviää yleensä

luontevasti ajattelemalla se yksinkertaiseksi luonnollisen kielen lauseeksi, tässä tapauksessa “koira on eläin” tai tarkemmin “koira kuuluu luokkaan eläin.”

Literaalit esitetään tässä tutkielmassa lainausmerkeissä, esimerkiksi (Ressu, nimi on, “Ressu”). Tässä lauseessa kannattaa kiinnittää huomiota muutamaan seikkaan. Lause saattaa aluksi tuntua oudolta ja tarpeettomalta: “Ressun nimi on Ressu.” Subjektina toimii käsite Ressu, joka on URI-viittaus. Tämä käsite tarkoittaa tiettyä tunnettua Charles M. Schulzin sarjakuvissa esiintyvää koira. Tämä käsite ei millään tavalla riipu käytettävästä kielestä tai siitä, mikä koiran nimi on kyseisessä kielessä. Lisäksi URI-viittauksesta ei voida päätellä käsitteen yleiskielistä nimeä. URI voisi olla vaikka “<http://www.esimerkki.fi#01234567-890A-BCDE-F012-34567890ABCD>”, jossa loppuosa on vain jokin yksilöllinen tunniste. Tässä tutkielmassa vain yksinkertaisuuden vuoksi on samaistettu URI ja käsite, jota URI tarkoittaa.

Edellisen esimerkin objekti on tekstiliteraali, joka kuvaa subjektin nimeä. Nimi on suomeksi Ressu, mutta englanniksi Snoopy. Niinpä lause kannattaisi esittää muodossa (Ressu, nimi on, fi:“Ressu”). Tässä objektin eteen on lisätty määre fi kuvamaan sitä, että kyseessä on suomenkielinen tekstiliteraali. Näin on helppo lisätä toinen lause (Ressu, nimi on, en:“Snoopy”), joka antaa nimen englanniksi. Literaaliin voidaan myös lisätä tyyppi, joka tässä tutkielmassa ilmaistaan samalla tavalla, esimerkiksi (Ressu, syntymäpäivä, päiväys:“28.8”). Literaalin tyyppi kertoo jäsentäville tietokoneohjelmalle, minkälaista tekstiä se voi odottaa ja miten teksti tulisi tulkita. Tutkielmassa jätetään pääsääntöisesti tyyppi ja kieli ilmaisematta yksinkertaisuuden vuoksi, ellei nämä ole juuri erityisen oleellisia.

RDF-tietämys koostuu joukosta yksinkertaisia RDF-lauseita. Esimerkki pienestä tällaisesta joukosta on esitetty taulukossa 2. Joukkoa RDF-lauseita voidaan myös ajatella verkkona. Verkon solmuina ovat käytetyt resurssit ja lauseista saadaan suunnatut, nimetyt kaaret. Kaarien niminä on lauseiden predikaatit ja suunta on aina subjektista objektiin. Literaalien tapauksessa kaaren nimen osaksi tulee myös mahdollisesti annettu kieli ja tietotyyppi.

RDF-tietokantaan voidaan tehdä kyselyjä kuten relaatiotietokantoihinkin. Tätä varten on kehitetty erilaisia kyselykieliä. W3C:n suositus on SPARQL-kieli [W3C07]. Esimerkiksi voidaan hakea sellaiset resurssit x, joille on olemassa RDF-lause (x, on, sarjakuva), eli kaikki sarjakuvat. Myös monimutkaisemmat rakenteet onnistuvat. Esimerkiksi voidaan hakea kaikki sarjakuvissa esiintyvät koirat: (x, on, koira) ja (y, sisältää hahmon, x) ja (y, on, sarjakuva). Edellinen kysely olettaa, että tietokanta ymmärtää käsitteen yläluokka semantiikan, mitä käsitellään seuraavassa luvussa.

(Ressu,	nimi on,	fi:"Ressu")
(Ressu,	nimi on,	en:"Snoopy")
(Ressu,	syntymäpäivä,	päiväys:"28.8")
(Ressu,	on,	beagle)
(beagle,	yläluokka,	koira)
(koira,	yläluokka,	eläin)
(Ressu,	omistaja,	Jaska Jokunen)
(Jaska Jokunen,	nimi,	fi:"Jaska Jokunen")
(Jaska Jokunen,	nimi,	en:"Charlie Brown")
(Jaska Jokunen,	on,	ihminen)
(Tenavat,	on,	sarjakuva)
(Tenavat,	piirtäjä,	Charlez M. Schulz)
(Tenavat,	sisältää hahmon,	Jaska Jokunen)
(Tenavat,	sisältää hahmon,	Ressu)
(Charlez M. Schulz,	on,	ihminen)

Taulukko 2: Esimerkki joukosta toisiinsa liittyvistä RDF-lauseista

3.3 Ontologiat ja RDF:n laajennokset

Taulukossa 2 annetun RDF-tietämyksen semantiikka on selvä ihmiselle vaikka lukija ei tietäisi mitään Tenavat-sarjakuvasta. Ihminen voi myös tehdä erilaisia johtopäätöksiä annettujen lauseiden perusteella, esimerkiksi voidaan päätellä, että (Ressu, on, koira) vaikka tällaista lausetta ei ole taulukossa. Tämä johtuu siitä, että RDF-resurssit on taulukossa annettu yleiskielellä ja lukijalla on käsitys siitä, mitä eri resurssit tarkoittavat. Tietokoneella ei vastaavaa perustietoa ole. Vastaava perustieto voidaan kuitenkin määritellä ja liittää osaksi RDF-tietämystä. Koska RDF-tietomallin on tarkoitus pystyä mallintamaan mitä tahansa tietämystä, voidaan sitä käyttää myös tämän perustietämyksen määrittelemiseen. Tällaista perustietämystä kutsutaan semanttisen webin yhteydessä *ontologiaksi*.

W3C on antanut kaksi suositusta, joilla voidaan määritellä ontologioita. Näistä suppeampi on RDF Schema (RDFS) [W3C04c] ja tämän päälle rakentuva monipuolisempi Web Ontology Language (OWL) [W3C04b]. OWL on itsessään vielä jaettu kolmeen eri versioon, joilla on erilaiset ilmaisuvoimat ja vastaavasti erilainen teoreettinen laskennallinen vaativuus päättelyiden suorittamisessa. Sekä RDFS että OWL käyttävät RDF-tietomallia ja standardit oleellisesti vain määrittelevät tarkan

semantiikan tietyille URI-viittauksille. Jos RDF-tietämystä käsittelevä tietokoneohjelma on ohjelmoitu siten, että se ymmärtää tämän semantiikan, se kykenee päätelemään erilaisia asioita joukosta RDF-lauseita. Erityisesti sellaisia asioita, joita RDF-tietämyksessä ei ole suoraan sanottu, mutta jotka esimerkiksi ihminen suoraan havaitsee. Ontologiaa voidaan myös käyttää tietämyksen eheyden tarkistamiseen. Esimerkiksi voidaan sanoa, että jokaisella ihmisellä on enintään kaksi vanhempaa. Jos jollakin RDF-resurssilla on enemmän, voidaan todeta, että tietämys ei ole ontologian mukainen.

RDFS määrittelee esimerkiksi sellaiset perusluokat kuin resurssi, luokka, datatyypipi ja ominaisuus. Nämä perusluokat kertovat, minkä tyyppisiä eri RDF-resurssit ovat. Jokainen RDF-resurssi on resurssi-luokan ilmentymä ja jokainen muu luokka on resurssi-luokan aliluokka. Luokalla luokka voidaan ilmaista, että jokin RDF-resurssi on luokka; luokka onkin itsensä ilmentymä. Vastaavasti datatyyppi-luokan ilmentymät ovat datatyyppisiä ja niin edelleen.

Lisäksi RDFS määrittelee perusominaisuuksia, jotka kaikki luonnollisesti ovat itsessään resursseja ja ominaisuus-luokan ilmentymiä. Ominaisuudella tarkoitetaan resurssia, jota voidaan käyttää predikaattina kolmikossa. Taulukon 2 esittämässä tietämyksessä käytetään esimerkiksi sellaisia ominaisuuksia kuin ”nimi on”, ”syntymäpäivä”, ”on”, ”yläluokka” ja niin edelleen. RDFS:n määrittelemiä ominaisuuksia ovat esimerkiksi lähtö- ja maaliavaruus, tyyppi, yläluokka, yläominaisuus ja nimi.

Lähtö- ja maaliavaruuksilla voidaan määritellä, minkälaisia resursseja ominaisuus käsittelee. Esimerkiksi ominaisuuden piirtäjä lähtöavaruus voisi olla sarjakuva ja maaliavaruus henkilö. Tämä tarkoittaa, että vain sarjakuvilla on ominaisuus piirtäjä ja sarjakuvan piirtäjä on aina henkilö. Itseasiassa RDFS-määrittelee myös, että lähtöavaruus-ominaisuuden lähtöavaruus on ominaisuus ja maaliavaruus on luokka. Tämä on luonnollista, sillä ei ole mielekäästä puhua esimerkiksi henkilön lähtöavaruudesta tai sanoa, että maaliavaruus on Charlz M. Schulz.

Tyyppi-ominaisuus kertoo jonkin resurssin tyyppin eli sen, minkä luokan ilmentymä resurssi on. Yläluokka-ominaisuudella voidaan muodostaa hierarkkisia luokkarakenteita. Yläluokka- ja tyyppi-ominaisuuksille on määritetty semantiikka, jonka mukaan jonkin luokan ilmentymät ovat myös kaikkien luokan yläluokkien ilmentymiä. Jos koira on beaglen yläluokka niin tämä tarkoittaa, että jokainen beaglen ilmentymä on myös koiran ilmentymä vaikka sitä ei erikseen sanottaisi. Yläominaisuus on vastaava asia ominaisuuksille. Esimerkiksi tekijä-ominaisuus voisi olla ominaisuuden piirtäjä yläominaisuus. Tämä tarkoittaa sitä, että koska Schulz on Tenavien piirtäjä,

hän on myös sen tekijä. Nimi-ominaisuudella annetaan yleiskielinen nimi resurssille. Edellä on esitelty vain RDFS:n määrittelemiä URI-viittauksia. Pelkästään näillä perusluokilla ja -ominaisuuksilla voidaan päätellä lukuisia asioita taulukon 2 esimerkissä. Yläluokka-ominaisuuden semantiikan perusteella tiedetään, että Ressu on koira ja eläin. Jos taas tietovarastosta etsitään sarjakuvien tekijöitä, löydetään myös Schulz vaikka häntä ei ole suoraan mainittu Tenavien tekijänä vaan piirtäjänä. Kun tietokone hakee Ressulle yleiskielistä nimeä niin se tietää, mikä ominaisuus tämän kertoo.

Web Ontology Language (OWL) rakentuu RDFS:n päälle ja määrittelee lisää erilaisia URI-viitteitä. OWL:n avulla ominaisuuksien käyttäytymistä voidaan tarkentaa esimerkiksi määrittelemällä ominaisuuksia transitiivisiksi, symmetrisiksi, funktionaalisiksi tai kahden ominaisuuden voidaan sanoa olevan toistensa käänteisominaisuuksia. Esimerkiksi yläluokka-ominaisuus on transitiivinen, jos A on B:n yläluokka ja B on C:n yläluokka niin A on C:n yläluokka, ja lapsi ja vanhempi ovat toistensa käänteisominaisuuksia

OWL tarjoaa myös keinon rajoittaa ominaisuuksia paikallisesti. RDFS:llä voidaan sanoa, että ominaisuuden vanhempi lähtö- ja maaliavaruudet ovat ihminen. Tästä seuraa selvästi ongelmia, jos yritetään kertoa Ressun vanhemmat, sillä näiden tulisi olla koiria. OWL:lla voidaan määritellä, että maaliavaruus on ihminen, jos ominaisuus liittyy ihmiseen ja koira jos ominaisuus liittyy koiraan. Lisäksi ominaisuuksien kardinaalisuutta voidaan rajoittaa, esimerkiksi vanhempia on aina kaksi. OWL:ssa luokkia voidaan myös määritellä toisten luokkien leikkauksilla, unioneilla ja komplementeilla tai luokan jäsenet voidaan luetella suoraan.

OWL:sta on määritelty kolme eri versiota: OWL lite, OWL DL ja OWL full. Näistä lite on ilmaisuvoimaltaan heikoin ja full voimakkain. Ilmaisuvoiman kasvaessa päätelyongelmat tulevat hankalammiksi. Edellä esitettyjen ominaisuuksien käytössä on rajoituksia jos halutaan pidättäytyä lite- tai DL-versiossa. OWL DL on suunniteltu siten, että siinä kaikki päätelyongelmat ovat vielä ratkeavia, mutta voivat olla erittäin hankalia.

4 Semanttinen wiki

Perusajatus semanttisessa wikissä on sama kuin semanttisessa webissä. Wikin sivut vastaavat hyvin läheisesti normaalin webin sivuja. Kukin sivu on tekstimassaa,

josta on linkkejä muihin sivuihin. Niinpä semanttisen webin ideaa voidaan soveltaa sellaisenaan wikeihin. Lisätään sivujen sisältöön koneluettavaa tietoa siitä, mitä sivu sisältää. Myös saatava hyöty on pääpiirteiltään sama wikeissä. Kun järjestelmä pystyy jäsentämään tiedon, voidaan sen avulla suorittaa yksinkertaisia päätelmiä ja koostaa tietoa useilta sivuilta yhteen. Samoin hakuominaisuuksia voidaan kehittää. Aliluvuissa esitellään ensin tarkemmin, miten semantiikkaa tarkkaan ottaen lisätään wikeihin ja sen jälkeen, mitä hyötyä tästä olisi wikien, erityisesti Wikipedian, käyttäjille.

4.1 Toteutus ja artikkelien editointi

Yksinkertainen tapa toteuttaa semanttinen wiki on käyttää wikien perusideaa pelkän RDF-tietämyksen muokkaamiseen. Jokaista RDF-resurssia vastaa yksi sivu, ja jokaisella sivulla voidaan muokata sellaisia RDF-lauseita, joissa subjektina on sivua vastaava resurssi. Jokaisella sivulla siis sanotaan yksinkertaisia asioita sivua vastaavasta resurssista. Tällaista menetelmää käytetään OntoWiki-järjestelmässä [ADR06]. Sivulla ei siis ole vapaamuotoista tekstiä lainkaan lukuun ottamatta RDF-lauseiden literaaliarvoja. Esimerkiksi Ressua vastaavalla sivulla voidaan syöttää RDF-lauseet (Ressu, nimi on, "Ressu") ja (Ressu, syntymäpäivä, päiväys:"28.8"), mutta ei (Tennavat, sisältää hahmon, Ressu) sillä tässä Ressu ei ole subjekti.

Wikipediassa suurin osa tiedosta on esitetty vapaana tekstinä luonnollisella kielellä, ja artikkelit on tarkoitettu ihmisen luettavaksi. Vaikka RDF-lauseiden merkitys onkin yleensä melko selvä sellaisenaan luettuna, RDF-kieltä ei ole tarkoitettukaan korvaamaan luonnollista kieltä. Koska tarkoituksena on pitää Wikipedia pääasiassa entisellään, mutta lisätä siihen joitakin semanttisia ominaisuuksia, ei voida olettaa, että koko sisältö käännetään RDF:ksi.

Tässä tutkielmassa keskitytäänkin pääasiassa järjestelmiin, joissa vapaamuotoinen teksti ja RDF-tietämys ovat yhdessä. Artikkeleja siis muokataan nykyisen Wikipedian tapaan, ja artikkelit on pääasiassa esitetty luonnollisella kielellä ja tekstin muotoilun osalta wikimerkkauksielellä. RDF-tietämys voidaan esittää laajentamalla käytettävää merkkauksielellä, jolloin RDF-lauseet voidaan kirjoitetaan normaalin tekstin sekaan [Völ06, AuA05]. Vaihtoehtoisesti RDF-tietämystä ja varsinaista artikkelia voidaan muokata toisistaan erillään [Joc06, TCC04]. Kummassakin menetelmässä on etunsa ja haittansa. Varsinaisen tekstisisällön ja RDF-lauseiden pitäminen yhdessä varmistaa, että tekstin ja RDF:n sisältämä tieto pysyy yhdenmukaisena, sillä molempia on aina muokattava yhtä aikaa. Lisäksi annotointi lienee helpom-

paa näin, koska RDF-lauseet liittyvät konkreettisesti itse tekstiin. Toisaalta tekstin seasta voi olla myöhemmin hankala löytää jotakin tiettyä RDF-lausetta, ja RDF-lauseiden kiinteä yhteys tekstisisältöön saattaa myös rajoittaa annotointia. Lisäksi Wikipedian merkkaukielen muuttaminen saattaa aluksi hämmentää editoreja.

Kummassakin vaihtoehdossa annotointi, eli semanttisen tiedon lisääminen artikkeleihin, on vapaaehtoista. Järjestelmiin on siis mahdollista tehdä sivuja, joita ei ole annotoitu lainkaan ja jotka sisältävät vain luonnollista tekstiä. Tällöin kynnys lukijasta editoriksi siirtymiseen pysyy matalana, sillä annotoinnin syntaksia tai merkitystä ei heti tarvitse ymmärtää. Kaikki nykyiset Wikipedian sivut myös voidaan sellaisenaan ottaa käyttöön semanttisessa wikissä.

4.2 Hyöty Wikipedian kannalta

Wikipedia sisältää lukuisia sivuja, joiden ainoa tarkoitus on koota yhteen tietoa ja siten parantaa tietosanakirjan käytettävyyttä [Völ06]. Tällaisia sivuja ovat esimerkiksi lista amerikanitalialaisista elokuvanäyttelijöistä tai lista valtioiden päämiehistä. Sivuja ylläpidetään, kuten kaikkia muitakin sivuja, käsin editoimalla sisältöä. Sivujen tekeminen ja ylläpitäminen on siis työlästä. Jos henkilön tiedot muuttuvat, esimerkiksi kun henkilö kuolee, on tämän kuolinaika päivitettävä lukuisille eri sivuille. Wikipedian pienemmissä kieliversioissa ei yksinkertaisesti ole riittävästi editoreja, jotta olisi resursseja ylläpitää lukuisia tällaisia listoja [Völ06].

Osana semanttista Wikipediaa on RDF-tietovarasto, johon voi suorittaa kyselyjä. Sivujen lähdekoodiin voidaan upottaa tietokantakyselyjä, jotka suoritetaan kun sivu esitetään käyttäjälle. Kyselyn tulos esitetään listana tai taulukkona sivulla. Wikipedian nykyiset listat ja taulukot voidaan siis yksinkertaisesti tuottaa kyselyn palauttamasta tuloksesta [Völ06, Joc06]. Tällöin listojen tekeminen on huomattavasti helpompaa ja kaikki tiedot pysyvät automaattisesti ajan tasalla keskenään, sillä tiedot esimerkiksi syntymä- ja kuolinajoista on syötetty vain yhteen paikkaan.

Tiedonhakua voidaan myös oleellisesti parantaa nykyisestä tekstihausta. Aivan kuten erilaisia listoja voidaan tuottaa hauilla RDF-tietokantaan, käyttäjät voisivat suoraan hakea tietoa Wikipediasta suorittamalla RDF-hakuja tietokantaan [Völ06, Joc06]. Jos käyttäjä etsii suomalaisia 60-luvun elokuvatahtiä niin tietokantaan voidaan suorittaa juuri tämä haku ja saadaan täsmällinen hakutulos. Nykyisin käyttäjä saa hakutermeistä riippuen joukon sivuja, jotka hyvällä tuurilla saattavat olla sitä mitä haettiin.

Hakulauseiden muodostaminen vaatii luonnollisesti hakukielen tuntemista, eivätkä peruskäyttäjät välttämättä pysty tekemään monimutkaisempia hakuja. Tämä voidaan osittain ratkaista hakupohjilla, jotka ovat valmiiksi tehtyjä hakuja, joissa jotkin muuttujat on jätetty käyttäjän täytettäväksi [Joc06]. Voidaan myös käyttää muita semanttisen webin interaktiivisia hakumenetelmiä kuten moninäkömähakua [MHS06]. Erilaiset Wikipedian nykyiset listasivut voidaankin ajatella ennalta määritellyiksi hauiksi. Pelkästään näiden huomattavasti helpompi tekeminen ja ylläpitäminen parantaa hakumahdollisuuksia.

Nykyisessä Wikipediassa on eräs ominaisuus, joka mahdollistaa tiedon uudelleenkäytön. Wikipediaan on mahdollista luoda sivupohjia (engl. template), joita voidaan sisällyttää toisten sivujen sisään. Englanninkielisessä Wikipediassa on esimerkiksi sivupohja `Template:Politics_of_Finland`, joka sisältää perustietoa Suomen hallituksesta ja eduskuntapuolueista. Tämä sivupohja on sisällytetty useisiin muihin Suomen politiikkaa käsitteleviin sivuihin. Kun eduskunnan puhemies muuttuu, se voidaan muuttaa vain tähän sivupohjaan ja sama tieto siirtyy sivuihin, joissa sivupohjaa käytetään. Tähän tarkoitukseen sivupohjat on kuitenkin selvästi puutteellinen menetelmä. Tieto päivittyy vain sivupohjan tuottamaan osaan sivua, esimerkiksi sivun laidassa olevaan tietolaatikkoon. Kaikki maininnat puhemiehestä itse tekstissä on edelleen hoidettava käsin. Ei myöskään ole tarkoituksenmukaista tehdä omaa sivupohjaa jokaiselle yksittäiselle faktalle, sillä tämä kasvattaisi niiden määrän suhteettoman suureksi. Sen sijaan sivupohjia käytetään isompien kokonaisuuksien yhteydessä kuten useiden sivujen laidassa olevien tietolaatikoiden tuottamiseen. Semantiikan lisääminen toisi vastaavan edun kaikkeen tietoon wikipediassa ja huomattavasti hienojakoisemmin aina yksittäisten henkilöiden syntymäaajoista lähtien. Sivupohjilla on kuitenkin muitakin käyttöjä, joihin ne soveltuvat hyvin, joten ei ole tarkoitus korvata niitä kokonaan.

RDF-tietomallissa käsitteet ovat täysin kieliriippumattomia. Käsite koira tarkoittaa samaa asiaa kaikilla mahdollisilla kielillä. Ainoastaan käsitteen nimi, siis tässä tapauksessa kolmikko (Koiru, nimi on, fi:“Koiru”), riippuu kielestä. Täten myös lähes kaikki RDF-kolmikot ja siis RDF-tietämys on kieliriippumaton. Wikipedian lukuisat kieliversiot voisivat toimia yhdellä kaikki kielet käsittävällä RDF-tietovarastolla. Verrattuna nykyiseen Wikipediaan tiedon pitäminen yhdenmukaisena kieliversioiden välillä olisi huomattavasti helpompaa, sillä tieto olisi aina vain yhdessä paikassa. Lisäksi pienemmät kieliversiot voisivat hyötyä englanninkielisen Wikipedian huomattavasti suuremmasta tietovarastosta. Pelkän RDF-tiedon avulla ei voida tuottaa täydellisiä artikkeleja, mutta tietolaatikoita siitä voidaan tuottaa helposti.

5 Tiedon louhinta Wikipediasta

Semanttisen webin menetelmien soveltaminen Wikipediassa ei välttämättä vaadi sitä, että Wikipedian tietämys esitetään RDF-tietomallilla. Oleellisin asia on se, että tietämys on jossakin rakenteellisessa muodossa, jota voidaan käsitellä tietokoneella. Kun tietämystä voidaan käsitellä koneellisesti, on melko triviaalia muuntaa tieto muodosta toiseen, esimerkiksi RDF-tietomallin mukaiseksi, jolloin sitä voidaan käyttää muissa sovelluksissa. Osoittautuu, että Wikipedia sisältää jo nyt jonkin verran rakenteellista tietoa, jota voidaan tulkita koneellisesti.

Edellisessä luvussa mainittiin sivupohjat Wikipediassa ja miten niiden avulla voidaan käyttää uudelleen tietoa usealla sivulla. Toinen tapa käyttää sivupohjia Wikipediassa on käyttää niitä tuottamaan yhdenmukaisia käyttöliittymäelementtejä samantyyppisille sivuille, mutta eri tietosisällöllä. Esimerkiksi käytännössä kaikkien elokuva-artikkelien oikeassa reunassa on tietolaatikko, jossa kerrotaan elokuvan ohjaaja, tuottaja, käsikirjoittaja, pääosien näyttelijät ja niin edelleen. Vastaavia tietolaatikoita käytetään hyvin monen tyyppisissä sivuissa. Tietolaatikon sisältö annetaan sivupohjalle varsinaisessa artikkelissa nimettyinä parametreina. Esimerkki tietolaatikosta ja sen tuottamasta lähdekoodista on esitetty kuvassa 2.

Sivupohjien parametrit on luettavissa melko vaivattomasti sivun lähdekoodista. DBPedia-projektissa [AuL07] muodostetaan RDF-tietämyskanta tulkitsemalla Wikipedian artikkelien tietolaatikoiden parametrit RDF-lauseina. Parametrin nimi antaa RDF-lauseen predikaatin ja arvo objektin. Käsiteltävä sivu toimii lauseen subjektina. Tietämyskantaan voi suorittaa kyselyitä SPARQL-kielellä, jolla saadaan täsmällisiä hakutuloksia. RDF-tietämys voidaan linkittää takaisin Wikipedian artikkeleihin, jolloin RDF-hakutuloksesta päästään takaisin Wikipediaan. Näin saadaan siis rakennettua semanttinen hakuliittymä Wikipediaan.

Auer et. al. mainitsevat artikkelissaan [AuL07] joitakin ongelmia sivupohjien parametrien tulkitsemisessa. Toisinaan parametreissa määritellään sivun esitykseen ja taittoon liittyviä asioita kuten esimerkiksi tekstin väri. Parametrien arvot sisältävät usein myös redundanttia informaatiota, esimerkiksi Nokian päätoimiston sijainti on Espoo, Finland. Suomen mainitseminen on semanttisesti tietenkin turhaa, mutta käyttäjän kannalta hyvinkin olennaista, sillä suurin osa Wikipedian käyttäjistä ei varmasti tiedä, missä Espoo sijaitsee. Koneellisesti tätä tilannetta on vaikea erottaa tapauksesta, jossa parametrin arvona on lista asioita, esimerkiksi Nokian tuotteina on matkapuhelimet, tietoliikenneverkot, langattomat järjestelmät. Lista voidaan

```

{{Tietolaatikko:Paavi|
Paavinimi=Johannes XXIII <br />
Ioannes XXIII|
Kuva= [[Image:Ioannes XXIII.jpg|250px]]|
Oikea nimi=Angelo Giuseppe Roncalli|
Paavius alkoi=[[28. lokakuuta]] [[1958]]|
Paavius päättyi=[[3. kesäkuuta]] [[1963]]|
Edeltäjä=[[Pius XII]]|
Seuraaja=[[Paavali VI]]|
Syntymäaika=[[25. marraskuuta]] [[1881]]|
Syntymäpaikka=[[Sotto il Monte]], [[Italia]]|
Kuolinaika=[[3. kesäkuuta]] [[1963]]|
Kuolinpaikka=[[Apostolinen palatsi]],
[[Vatikaani]]|}}

```

Johannes XXIII Ioannes XXIII	
	
Oikea nimi	Angelo Giuseppe Roncalli
Paavius alkoi	28. lokakuuta 1958
Paavius päättyi	3. kesäkuuta 1963
Edeltäjä	Pius XII
Seuraaja	Paavali VI
Syntynyt	25. marraskuuta 1881
Syntymäpaikka	Sotto il Monte, Italia
Kuollut	3. kesäkuuta 1963
Kuolinpaikka	Apostolinen palatsi, Vatikaani

Kuva 2: Oikealla tietolaatikko suomenkielisestä Wikipediasta artikkelista Johannes XXIII. Vasemmalla tietolaatikon tuottava osa artikkelin lähdekoodista. Tietolaatikon tuottamiseen käytetään sivupohjaa Tietolaatikko:Paavi.

purkaa muodostamalla jokaiselle listan jäsenelle erillinen RDF-lause. Lauseissa käytetään samaa subjektia ja predikaattia, mutta niihin saadaan eri objektit listan jäsenistä. Hankalampi tapaus on sellainen, jossa parametrissa pitäisi tulkita useampia lauseita, joiden predikaattien pitäisi myös vaihdella. Esimerkiksi Nokia artikkelissa on perustamista kuvaavan parametrin arvona on "Nokia, Finland (1865)". Tässä pitäisi muodostaa erilliset lauseet perustamispaikalle sekä -ajalle.

Sivupohjien parametrien tulkintaan liittyvät ongelmat ovat epäilemättä seurausta siitä, että sivupohjia ei ole suunniteltu käytettävän näin. Wikipedian oman ohjeistuksen mukaan sivupohjien tarkoitus on juuri toistuvien viestien helppo lisääminen sivulle sekä navigointielementtien ja tietolaatikoiden tuottaminen. Koska parametrien tulkinta toimii kuitenkin melko hyvin jo nykyisin, ei varmaankaan olisi mahdollonta parantaa sivupohjien käyttöä niin, että niiden sisältämä tieto saataisiin tulkittua täsmällisesti koneellisesti.

Toinen tapa louhia tietoa automaattisesti Wikipediasta on hyödyntää kategoriamäärittäjiä. Kukin artikkeli voidaan sijoittaa useisiin kategorioihin, jotka kuvaavat artikkelin aihetta tai artikkelia itseään. Kategoriat voidaan myös asettaa toisten kategorioiden alakategorioiksi, jolloin muodostuu hierarkkinen rakenne. Wikipediassa käytettävät kategoriat voidaan karkeasti jakaa neljään eri ryhmään [SKW07a]:

1. Kattegoria kuvaa artikkelin aihetta ja vastaa luokkaa, johon aihe kuuluu, esimerkiksi Yhdysvaltain kansalainen.
2. Kattegoria liittyy Wikipedian hallinnointiin, esimerkiksi artikkeli tarvitsee viitteitä.
3. Kattegoria sisältää täsmällistä muuta tietoa aiheesta, esimerkiksi syntynyt 1897.
4. Kattegoria kuvaa aiheen yleistä teemaa, esimerkiksi fysiikka.

Kategorioita tulkitsemalla voidaan muodostaa tietämyskanta samaan tapaan kuin sivupohjia tulkitsemalla. YAGO-ontologia käyttää tätä menetelmää [SKW07a]. YAGO käyttää omaa tietomalliaan, joka kuitenkin muistuttaa hyvin läheisesti RDF:ää ja joka on mahdollista muuntaa RDF:ksi tai vastaavilla menetelmillä voitaisiin suoraan muodostaa RDF-tietämyskanta. YAGO-ontologia on muodostettu yhdistämällä tietoa Wikipediasta ja WordNet-thesauuruksesta ⁶. WordNet itsessään sisältää käsitteiden hyperonyymi-hyponyymi -rakenteen, joka muistuttaa Wikipedian kategorioita.

⁶<http://wordnet.princeton.edu>

Wikipedian kategoriahierarkia on kuitenkin epämääräisempi kuin thesauruksen täsmällinen hierarkia. Tästä syystä YAGO:ssa on päätetty käyttää pääasiassa WordNetin hierarkiaa, mutta lisätä luokkahierarkian lehtitason luokat sekä ilmentymät näihin luokkiin Wikipediasta. Ajatuksena on siis yhdistää kummankin hyvät puolet, WordNet sisältää täsmällisen luokkahierarkian, mutta ei juurikaan luokkien ilmentymiä, kun taas Wikipedia sisältää ilmentymät, mutta epämääräisemmän luokkarakenteen.

Ontologian muodostaminen YAGO:ssa toimii seuraavasti [SKW07a]. Kunkin artikkelin jokainen luokka käsitellään erikseen. Luokista erotellaan ja käsitellään vain ne, jotka kuvaavat luokkaa, johon artikkelin aihe kuuluu (edellisen listan tyyppi 1). Nämä luokat erotetaan muista pääasiassa analysoimalla kategorian nimeä luonnollisen kielen menetelmin. Kategorialle etsitään yläluokka WordNetin hierarkiasta jälleen analysoimalla kategorian nimeä. Artikkelin asetetaan ontologiassa kategoriansa ilmentymäksi ja kategoria löydetyn WordNetin käsitteen aliluokaksi.

YAGO:a on lähiaikoina myös kehitetty eteenpäin siten, että kategorioiden lisäksi se tulkitsee joitakin sivupohjien parametreja DBPedian tapaan [SKW07b]. Merkittävänä erona DBPediaan on se, että jäsennettävät sivupohjat ja parametrit on valittu manuaalisesti sen sijaan, että yritetään jäsentää kaikkien sivupohjien kaikki parametrit. Näin saadaan luotettavampaa tietoa, mutta kattavuus ei ole yhtä hyvä kuin DBPediassa.

YAGO-ontologiaa ei ole tarkoitettukaan käytettävän erityisesti Wikipedian yhteydessä, se on vain yleiskäyttöinen ontologia, joka osittain perustuu Wikipediasta louhittuun tietoon. Se on kuitenkin hyvä esimerkki siitä, miten kategoriain sisältävät osittain koneellisesti tulkittavaa tietoa. Tämän tiedon hyödyntäminen on kuitenkin hankalaa nykyisessä muodossa etenkin monikielisessä ympäristössä sillä YAGO:n käyttämässä menetelmässä oleellinen osa on luonnollisen kielen tulkinta. Myös WordNet on englanninkielinen, joten muiden Wikipedian kieliversioiden tiedon yhdistäminen siihen ei onnistuisi suoraan. Kuten sivupohjien käyttöä Wikipediassa, kategorioitakin olisi mahdollista kehittää eteenpäin niin, että semanttisesti oleelliset kategoriat voitaisiin helposti erottaa koneellisesti. Lisäksi kategoriat voisi yhdistää manuaalisesti WordNetin käsitteisiin tai muiden ontologioiden resursseihin.

Molemmat edellä esitellyt menetelmät louhivat tietoa käyttäen säännöllisesti julkaittavia Wikipedian tietokantakopioita. Ne eivät siis toimi reaaliaikaisesti artikkeleita muokatessa. Menetelmät olisi kuitenkin mahdollista integroida kiinteämmin osaksi Wikipediaa siten, että tiedon louhinta tapahtuu kullekin artikkelille erikseen aina,

kun sitä muokataan. Tällöin tietokanta pysyisi täysin synkronissa Wikipedian sisäl-
lön kanssa.

6 Semantiikan lisääminen Wikipediaan

Edellisessä luvussa esiteltiin järjestelmiä, jotka louhivat automaattisesti tietoa Wi-
kipedian nykyisestä versiosta. Nämä eivät siis välttämättä vaadi mitään muutoksia
Wikipediaan tai artikkelien editorien toimintatapaan. Pienillä muutoksilla louhin-
tamenetelmät saataisiin tosin toimimaan huomattavasti paremmin. Tässä luvussa
esitellään menetelmiä, joissa editorit lisäävät artikkeleihin RDF-lauseita suoraan.
Tällöin editorien on siis omaksuttava tämä uusi ominaisuus Wikipediassa, ja kaikki
nykyiset artikkelit on päivitettävä käyttämään sitä. Tällä menetelmällä voidaan kui-
tenkin tuottaa monipuolisempaa tietämystä, sillä käytännössä mikä tahansa RDF-
lause voidaan syöttää Wikipediaan. Lisäksi RDF-tietämys pysyy synkronissa ar-
tikkelien tekstisisällön kanssa, sillä RDF-lauseet luetaan artikkeleista itsestään sen
sijaan, että ne tuotetaan sarja-ajona ajoittaan julkaistavan tietokantakopion avulla.

Testatuista menetelmistä parhaiten Wikipedian kanssa yhteen sopii Semantic Me-
diaWiki [Völ06]. Ei niinkään semanttisilta ominaisuuksiltaan vaan siksi, että se on
alun perin suunniteltu Wikipedian käyttämän MediaWiki-järjestelmän lisäosaksi ja
suunnittelussa on otettu huomioon Wikipedian nykyiset toimintamallit. Se pyrkii siis
aiheuttamaan mahdollisimman vähän muutoksia editorien nykyiseen toimintaan.

Semantic MediaWikissä semanttinen informaatio syötetään artikkelien normaalin
lähdekoodin sekaan. Merkkauskieltä on laajennettu siten, että sillä voidaan mää-
ritellä RDF-lauseita. Kuvan 1 Helsinkiä käsittelevä Wikipedian lähdekoodi on esi-
tetty kuvassa 3 hyödyntämällä Semantic MediaWikin ominaisuuksia. Lähdekoodiin
on tehty kahdenlaisia muutoksia. Osaan linkeistä on laitettu mukaan linkin tyyppi.
Ensimmäiseen linkkiin `[[pääkaupunki::Suomi|Suomen]]` on asetettu tyyppiksi
pääkaupunki. Tämän ansioista tietokone osaa tulkita linkin merkityksen. Artikke-
li Helsinki ja linkitetty artikkeli Suomi eivät ole vain epämääräisessä yhteydessä
keskenään vaan linkki kertoo täsmällisesti, että Helsinki on Suomen pääkaupunki.
Linkistä tehdään RDF-lause (Helsinki,pääkaupunki,Suomi). Lauseessa subjektiksi
tulee aina artikkeli itse, predikaatiksi linkin tyyppi ja objektiksi linkin kohde.

Toinen lisäys on asukaslukujen merkkkaus. Nämä eivät ole linkkejä, eikä niitä alku-
peräisessä lähdekoodissa ole merkattu millään erityisellä tavalla. Semanttisessa ver-
siossa niissä käytetään linkkejä muistuttavaa rakennetta ja niille annetaan tyyppi.

```
'''Helsinki''' ({{k-sv|Helsingfors}}) on [[pääkaupunki::Suomi|
Suomen]] [[pääkaupunki]]. Se sijaitsee
[[sijaitsee::Etelä-Suomen lääni|Etelä-Suomessa]],
[[sijaitsee-rannalla::Suomenlahti|Suomenlahden]] rannalla.
Helsingin rajanaapureita ovat pohjoisessa ja idässä
[[rajanaapuri::Vantaa]] ja lännessä [[rajanaapuri::Espoo]],
joiden lisäksi Helsingillä on merirajaa
[[merirajanaapuri::Sipoo|Sipoon]] kanssa. Helsingissä
on noin [[asukasluku:=570000]] asukasta ja se on noin
[[asukasluku-seutu:=1000000]] asukkaan
[[pääkaupunkiseutu|pääkaupunkiseudun]] keskuskaupunki.
```

Kuva 3: Suomenkielisen Wikipedian Helsinkiä käsittelevän artikkelin lähdekoodi anotoitu käyttäen Semantic MediaWikin syntaksia.

Näistä muodostetaan RDF-lauseet samalla tavalla kuin tyyplitetyistä linkeistä, mutta objektiksi tulee literaali. Tämän ansiosta tekstistä voidaan koneellisesti löytää myös sellaiset oleelliset faktatiedot, jotka eivät esiinny linkkeinä muille sivuille.

Semantic MediaWikissä jokaiselle linkki- ja literaalityypille on mahdollista tehdä oma artikkelinsa. Näin tyyppien merkitys ja käyttö voidaan dokumentoida. Lisäksi tyyppien sivuille voidaan liittää RDF-lauseita, joiden subjektina toimii tyyppi itse. Tämä mahdollistaa OWL-määritysten käyttämisen, jolloin tietämykseen voidaan soveltaa OWL:a ymmärtäviä päättelykoneita. Semantic MediaWiki itsessään hyödyntää OWL-määrittäjiä vain rajoitetusti.

Semantic MediaWikissä literaalityypit on mahdollista määritellä siten, että järjestelmä ymmärtää, kuvaako literaalin arvo pituutta, lukumäärää, lämpötilaa tai jotakin muuta suuretta. Suureille on mahdollista myös määritellä yksiköitä, jolloin literaalien esitys käyttäjälle voidaan toteuttaa käyttäjän valitsemilla yksiköillä, esimerkiksi kilometrit tai mailit, ja eri yksiköissä annettuja literaaleja voidaan myös vertailla keskenään oikein.

Muut testatut wikijärjestelmät ovat enemmän kehitysasteella olevia eivätkä malliltaan sovellu suoraan Wikipediassa käytettäväksi. Platypus-wikissä [TCC04] artikkelien tekstisisältöä ja RDF-lauseita muokataan toisistaan erillisillä käyttöliittymillä. RDF-lauseet syötetään valitsemalla erikseen pudotusvalikoista subjekti, objekti ja predikaatti. Varsinainen tekstisisältö syötetään erillisellä sivulla. Platypus-wikin

Platypus Wiki The Semantic Wiki Wiki Web

Search Help search

Home What is Logo Download License Contact

Platypus Wiki: the Semantic Wiki Wiki Web

Metadata Wiki Page

subject	predicate	object	actions
wikipedia v	Helsinki v	wikipedia v	Save
wikipedia:Helsinki	wikipedia:rajanaapuri	wikipedia:Espoo	Reify
wikipedia:Helsinki	wikipedia:pääkaupunki	wikipedia:Suomi	Reify

Copyright © 2004 Stefano Campanini, Paolo Castagna, Roberto Tazzoli - Platypus Wiki Source Code is licensed under GNU GPL
Web Standards: XHTML, CSS, WCAG1.0 Level AAA (Bobby Approved!)

Kuva 4: Platypus-wikin RDF-lauseiden syöttöliittymä. Lauseen jokainen jäsen valitaan erikseen pudotusvalikoista. RDF-resurssit voidaan ryhmitellä nimiavaruuksiin. Yllä käytetään vain wikipedia-nimiavaruutta.

RDF-lauseiden syöttöliittymä on esitetty kuvassa 4.

Verrattuna Platypus-wikiin OntoWikissä [ADR06] on kehittyneempi käyttöliittymä sekä RDF-lauseiden syöttämiseen että tiedon selailuun. Toisaalta se käsittelee pelkästään RDF-lauseita, eikä vapaamuotoista tekstiä voi syöttää lainkaan paitsi RDF-lauseiden literaaleissa. OntoWikissä on myös mahdollista käyttää erilaisia käyttöliittymäelementtejä riippuen RDF-lauseiden tyypistä. Esimerkiksi päivämääriä voidaan syöttää ja selata kalenterinäkymän kautta. Kuvassa 5 on esitetty OntoWikin käyttöliittymä.

WikSar [AuA05] puolestaan muistuttaa melko paljon Semantic MediaWikiä. RDF-lauseet syötetään muun tekstin seassa merkkiaamalla linkkien tyypit. Tyypitetyistä linkeistä muodostetaan suoraan RDF-lauseet samaan tapaan kuin Semantic MediaWikissä. Muilta ominaisuuksiltaan WikSar ei ole yhtä kehittynyt kuin Semantic MediaWiki.

Resource

Label: Helsinki

Identifier (URI): http://localhost/ontowiki-0.8/wikipedia#Helsinki

Properties

Paakaupunki Suomi

Rajanaapuri

Resource Literal

Espoo

Resource Literal

Vantaa

Submit Values Add Property

Kuva 5: OntoWiki-järjestelmän peruskäyttöliittymä. RDF-lauseet syötetään kirjoittamalla predikaatti ja objekti näille varattuihin kenttiin. Subjektia ei tarvitse erikseen valita, sillä se on aina muokattava aihe itse.

7 Semanttisten wikijärjestelmien testaus

Tässä luvussa käsitellään semanttisten wikijärjestelmien testaus. Ensimmäisessä aliluvussa esitellään testimenetelmät ja sitä seuraavissa aliluvuissa paneudutaan tarkemmin havaittuihin ongelmiin.

7.1 Testimenetelmät

Testausta varten valittiin yhteensä viisitoista artikkelia, jotka annotoitiin. Testaus tehtiin pääasiassa vain Semantic MediaWiki-järjestelmällä, sillä tämä on nykyisistä järjestelmistä pisimmälle kehitetty ja suunniteltu nimenomaan Wikipediaan soveltuvaksi. Suuri osa havaituista ongelmista myös osoittautui liittyvän tiedonmallinukseen RDF:llä, jolloin samat ongelmat esiintyvät kaikissa RDF pohjaisissa järjestelmissä.

Testausta varten valittiin Wikipediassa suositelluiksi artikkeleiksi merkattuja artikkeleita. Nämä ovat artikkeleita, jotka on todettu laadultaan mallikelpoisiksi. Näitä artikkeleita esitellään englanninkielisessä Wikipediassa etusivulla päivittäin ja suo-

menkielisessä hieman harvemmin. Lähtökohtana oli satunnainen joukko artikkeleita. Esitellyissä artikkeleissa ei ole mitään ajallista teemaa, joten riittävän satunnainen otos saatiin valitsemalla joukko peräkkäin esiteltyjä artikkeleita. Valitut artikkelit on esitelty etusivulla 2007-2008 vuodenvaihteessa. Testaukseen valittiin tällä tavalla viisi englanninkielistä ja viisi suomenkielistä artikkelia. Näin saatiin joukko hyvin erilaisia aiheita käsitteleviä artikkeleita.

Erittäin laaja aihepiiri ja pieni otos artikkeleita aiheutti sen, että eri artikkeleissa annotoiduilla tiedoilla ei ole mitään tekemistä keskenään. Tämän takia kymmenen satunnaisen artikkelin lisäksi Wikipediasta etsittiin testiotokseen viisi jotakin eläinlajia käsittelevää artikkelia. Myös nämä olivat suositeltuja artikkeleita. Suositeltujen artikkelien käyttäminen saattoi aiheuttaa sen, että eläinlajit olivat hieman erikoisia tai niillä oli erityistä kulttuurista merkitystä, sillä tavanomaisista eläimistä ei välttämättä kirjoiteta yhtä kattavasti. Satunnaisotoksessa oli jo lisäksi kaksi eläinlajia käsittelevää artikkelia. Useampi samaa aihepiiriä käsittelevä artikkeli mahdollisti monipuolisemman testauksen.

Valitut artikkelit on listattu taulukossa 3. Artikkelit 1-10 on valittu satunnaisotoksella ja artikkelit 11-15 ovat erikseen valitut eläimiä käsittelevät artikkelit.

Testatut artikkelit annotoitiin lisäämällä linkkeihin tyyppejä ja merkkamalla literaaliarvoja. Tyyppejä ei lisätty kuin osaan linkeistä, sillä useissa tapauksissa ei pystytty sanomaan, mikä olisi sopiva tyyppi tai sellaista ei yksinkertaisesti ollut. Artikkeleihin ei myöskään lisätty kokonaan uusia linkkejä eikä niitä muokattu millään muulla tavoin.

Systemaattisen testauksen lisäksi selattiin satunnaisesti muita Wikipedian artikkeleita ja niistä etsittiin muita mahdollisia ongelmakohtia. Suurin osa tällä tavalla löydettyistä ongelmista kuitenkin esiintyi jo valitussa testiaineistossa.

7.2 Yleiset havainnot

Suurimpaan osaan linkeistä ei lisätty mitään tyyppiä, eikä kaikkia potentiaalisia literaaliarvoja merkattu. Ongelmat olivat pääasiassa samankaltaisia sekä linkkien että literaaliarvojen kanssa. Tapaukset, joissa linkkityyppiä ei laitettu tai literaaliala ei merkattu, on karkeasti jaettavissa seuraaviin ryhmiin.

1. Linkki oli lähinnä vain lukijaa varten, eikä linkin artikkeli suoraan liittynyt lainkaan käsiteltävään artikkeliin. Linkille ei yksinkertaisesti ollut mitään jär-

	Artikkeli	Kieli	Aihe
1	Ulysses (poem)	en	Alfred Tennysonin runo
2	Gregory of Nazianzus	en	300-luvun kirkkoisä
3	Fin Whale	en	Sillivalas
4	William Cooley	en	1800-luvun Amerikan uudisasuttaja
5	Queluz National Palace	en	Portugalilainen 1700-luvun palatsi
6	Banaanikärpänen	fi	Banaanikärpänen
7	Johannes XXIII	fi	1900-luvun paavi
8	Liikennevalo-ohjaus	fi	Liikennevalo-ohjaus
9	Gaugamelan taistelu	fi	Aleksanteri Suuren taistelu
10	Maasait	fi	Itäafrikkalainen kansa
11	Californian condor	en	Kaliforniankondori
12	Krill	en	Krillit, eläinlahko, ei eläinlaji
13	Guinea pig	en	Marsu
14	Olm	en	Olmi, salamannerilaji
15	Bald eagle	en	Valkopääamerikotka

Taulukko 3: Testauksessa käytetyt Wikipedian artikkelit

kevää tyyppiä.

Esimerkki liikennevaloja käsittelevässä artikkelissa: “1910-luvulla Yhdysvalloissa syntyi liikennevaloista useita versioita, joista monille ei kylläkään haettu [[patenttia]].”

- Linkkiin on laitettavissa järkevä tyyppi, mutta subjektina ei ole artikkelin aihe vaan jokin muu. Usein tällaiset tapaukset on annotoitavissa subjektina toimivaan artikkeliin itseensä. Paitsi sellaisissa tapauksissa, missä subjekti itsessään ei ole kovin merkittävä ja Wikipediassa ei ole sille artikeelia.

Esimerkiksi Queluz-palatsia käsittelevässä artikkelissa: “[[Frederik Suuri]] käytti Euroopan toista kuuluisaa rokokoopalatsia, [[Sanssouci]]-palatsia, vastaavalla tavalla.”

- Linkki liittyy aiheeseen ja on aiheen kannalta oleellinenkin, mutta asia ei ole ilmaistavissa yksinkertaisella yhden RDF-lauseen subjekti, predikaatti, objekti -rakenteella. Usein tällaisia tapauksia voidaan yksinkertaistaa, jolloin annotointi periaatteessa on järkevä, mutta ei niin merkittävä kuin varsinaisen tekstin asia.

Esimerkki: “Basil vihki Gregoryyn [[Sasiman]] piispaksi vuonna 372.” Linkin tyyppiä voisi tässä tapauksessa laittaa työpaikka, mutta tällöin vuosi ja ammatti jäisi epäselviksi.

Gaugamelan taistelua käsittelevässä artikkelissa sanotaan: “Dareioksen jalkaväkeen kuului miljoona sotilasta. Ratsuväkeäkin oli 40 000.” Tässä tapauksessa puolestaan sotajoukkojen suuruus pitäisi ilmoittaa erikseen taistelun kummallekin osapuolelle, joten yksinkertainen tyyppi, jalkaväen lukumäärä, ei toimi. Tämä voidaan myös nähdä edellisen kohdan ongelmatyyppinä. Subjektiksi voisi ajatella Dareioksen armeijan Gaugamelan taistelussa, mutta tällaista artikkelia ei Wikipediassa ole.

4. Linkkiin on laitettavissa tyyppi, mutta on kyseenalaista käytettäisiinkö samaa tyyppiä missään muussa artikkelissa, jolloin annotointi on turhaa.

“Esimerkiksi alkoholin vaikutusten ja [[alkoholismin]] tutkimuksessa banaanikärpystä on käytetty erittäin paljon.” On mielenkiintoista, että banaanikärpäsän avulla voidaan tutkia alkoholismia, mutta tehdäänkö vastaavia annotointeja missään muussa yhteydessä. Tätäkin olisi mahdollista yksinkertaistaa ja lisätä vain temaattinen linkki alkoholismin ja banaanikärpäsän välille, mutta semanttisesti suhde jää tällöin epäselväksi.

“Kromosomien koko on erittäin suuri tavallisiin interfaasikromosomeihin verrattuna. Ne voivat olla paksuudeltaan jopa 15-25 μm ja pituudeltaan 100-250 μm .” Harvassa eläimiä tai hyönteisiä käsittelevässä artikkelissa kromosomeja käsitellään näin tarkasti. Onko siis mielekästä annotoida tässäkään tapauksessa.

Artikkelien alussa oleva tiivistelmä, kuten kuvissa 1 ja 3, oli yleisesti ottaen helpommin annotoitavissa ja pääosa lisätyistä annotoinneista lisättiin juuri tähän osioon. Niissä artikkeleissa, joissa oli sivupohjalla tuotettu tietolaatikko, oli yleensä suuri osa annotoiduista asioista myös tietolaatikossa. Tämä on ymmärrettävää, sillä tietolaatikkoon on valittu näytettäväksi juuri oleelliset yksinkertaiset faktatiedot, jotka on myös helppo annotoida tekstiin. Tämän pohjalta voi myös sanoa, että käyttämällä pelkästään tietolaatikkoa annotointiin, kuten luvun 5 automaattiset menetelmät tekevät, saadaan annotoitua suurin osa oleellisista asioista, mikäli tietolaatikkoon on ylipäänsä käytetty artikkelissa. Asian voi myös tulkita toisinpäin: Tietolaatikko voitaisiin koostaa itse artikkelissa annotoidusta tiedosta RDF-kyselyjen avulla.

Useissa artikkeleissa jäi annotoimatta oleellisia temaattisia suhteita. Esimerkiksi banaanikärpäsän käyttö geenitutkimuksessa, William Cooleyn rooli Amerikan asutuksessa ja Gregorios Nazianzilaisten suhde kristinuskon alkuvaiheisiin. Tämä johtui siitä, että usein artikkeleissa ei ole yksittäistä selkeää linkkiä, jossa tällainen asia voitaisiin ilmaista semanttisesti oikein. Sen sijaan tällaiset suhteet lienee parasta annotoida jollakin hieman epämääräisellä predikaatilla, joka yleisesti kuvaa käsitteiden läheisyyttä. Tällaiset suhteet on jo nykyisin usein ilmaistu Wikipedian omilla kategoriamäärittelyillä. Esimerkiksi liikennevalo-ohjaus kuuluu kategorioihin tieliikenne ja automaatio ja banaanikärpäsän kategoriaan perinnöllisyystiede.

Edellisen listan kohta 2 liittyy myös Wikipedian tiedon eheyteen. Kun lauseen subjektina ei ole artikkeli itse vaan jokin muu artikkeli, on mahdollista, kenties jopa todennäköistä, että sama asia on ilmaistu subjektin artikkelissa. Vastaava tilanne saattaa esiintyä myös muiden RDF-lauseiden kohdalla siten, että lauseen sisältö on ilmaistu kääntäen objektin artikkelissa. Yksi luvussa 4 mainituista semanttisen wikin eduista oli tiedon pysyminen yhdenmukaisena eri artikkelien välillä, sillä yksi asia on ilmaistu vain yhdessä paikassa. Semantic MediaWikissä on mahdollisuus suorittaa kyselyitä sivun sisällä siten, että vastauksena saadaan esimerkiksi henkilön synnyinvuosi. Täten on mahdollista ilmaista synnyinvuosi vain henkilön itsensä sivulla ja hakea tämä kyselyllä muilla sivuilla. Tällaisten pienten kyselyiden tekeminen Semantic MediaWikissä on kuitenkin melko kömpelöä. Lisäksi yleisesti ottaen artikkelit sisältävät huomattavan määrän päällekkäistä tietoa. Kaiken tämän esittäminen erilaisilla RDF-kyselyillä olisi käytännössä huomattavan hankalaa, vaikka wikijärjestelmässä olisikin kevyempi syntaksi tällaisten RDF-kyselyiden suorittamiseen.

Wikipediassa on mahdollisuus lukita joitakin sivuja. Lukittuja sivuja voi muokata vain pitkäaikaiset rekisteröityneet käyttäjät tai ylläpitäjät. Lukitsemista käytetään lähinnä vandaalitoiminnan estämiseen erityisesti voimakkaita mielipiteitä herättävillä sivuilla. Tyypillinen esimerkki lukitusta sivusta on Yhdysvaltain presidentti George W. Bush. Semanttisessa Wikipediassa eräs merkittävä hyöty on mahdollisuus koostaa ja muilla tavoin hyödyntää muilla sivuilla määriteltä tietoa. Lukittujen sivujen tapauksessa on kuitenkin oltava varovainen, sillä helposti voidaan saada aikaan tilanne, jossa vandaalikäyttäjä voi muokata lukitun sivun sisältöä jonkin toisen sivun kautta. Vaikkei lukittua sivua voisikaan suoraan muokata, voi silti olla mahdollista lisätä Wikipediaan väärää tietoa aiheesta esimerkiksi käyttämällä aihetta lauseen objektina. Yksi ainoa RDF-lause saattaa tulla esille lukuisissa paikoissa, mikäli tietoa koostetaan sivuille RDF-hauilla. Tämä saattaa aiheuttaa uusia

vandalismiin liittyviä ongelmia Wikipediassa.

Erityisen hankala artikkeli annotoida oli liikennevalo-ohjausta käsittelevä artikkeli, sillä sen aihe on varsin laaja. Edes johdanto luvussa ei esiintynyt selkeitä annotoitavia linkkejä, eikä artikkelissa ole myöskään käytetty tietolaatikkoo. Artikkelin kunnollinen annotointi edellyttäisi ensiksi määrittelemään, mitä tarkalleen ollaan annotoimassa. Jos aihe ymmärretään teknologiana tai keksintönä, voidaan heti luetella mielenkiintoisia faktatietoja kuten, keksijä ja käyttöönottovuosi. Tosin tässä tapauksessa nämäkin ovat hankalasti annotoitavissa, sillä liikenteenohjausta on kehitelty useampi eri henkilö samaan aikaan.

Semantic MediaWiki osaa tulkita literaalien numeroarvot ja muuntaa ne eri yksiköille. Se kuitenkin käyttää pistettä desimaalierottimena pilkun sijaan, mikä poikkeaa suomalaisesta käytännöstä. Muita huomattavia eroja englannin- ja suomenkielisten artikkelien välillä ei havaittu.

7.3 Monipaikkaiset relaatiot ja taulukot

Artikkelien yksityiskohtaisemmat osuudet sisälsivät lähes pelkästään linkkejä, joiden annotointi oli hankalaa. Henkilöiden elämäkerrat, taisteluiden taustat, eläinten elintavat ja rakennusten historiat sisältävät monimutkaista tietoa, jonka kääntäminen yksinkertaisiksi RDF-lauseiksi ei ole suoraviivaista. Tarkastellaan esimerkiksi lausetta: ”Roncalli vihittiin papiksi Roomassa vuonna 1904.” Jo tällaisen lauseen kääntäminen RDF:ksi vaatii useamman RDF-lauseen, sillä oleelliset tiedot ovat: henkilö eli Roncalli, tapahtuma eli papiksi vihkiminen, paikka eli Rooma ja aika eli 1904. Kokonaisuus voidaankin ajatella neljäpaikkaisena relaationa. W3-konsortiolla on suositus tällaisten monipaikkaisten relaatioiden koodaamisesta RDF:ksi [W3C06]. Suosituksen mukaan itse relaatiota varten tehdään oma resurssi, johon muut resurssit liitetään omilla lauseilla. Tässä tapauksessa siis saataisiin esimerkiksi seuraavat neljä RDF-lausetta:

(X, henkilö,	Roncalli)
(X, tapahtuman tyyppi,	papiksi vihkiminen)
(X, paikka,	Rooma)
(X, aika,	1904)

Tässä X on tapahtumaa itseään vastaava resurssi. Tällaista tapahtumiin perustuvaa annotointia on tutkittu ja myöskin käytetty muissa semanttisissa järjestelmissä [RuE07]. W3-konsortion suosituksessa mainitaan myös vaihtoehtoinen hieman erilai-

nen tapa koodata monipaikkaiset relaatiot. Tässä menetelmässä yksi RDF-lauseista esitetään toisinpäin. Edellisessä esimerkissä tarkoituksena on kertoa milloin ja missä Roncalli vihittiin papiksi. Papiksi vihkimisen voi siis ajatella predikaatiksi ja Roncallin subjektiksi. Objektiksi tehdään uusi resurssi, johon liitetään muut tiedot. Siis esimerkiksi:

```
(Roncalli, vihitty papiksi, X)
(X, paikka, Rooma)
(X, aika, 1904)
```

Semantic MediaWikissä on mahdollista määritellä tämänkaltaisia relaatioita. Se kuitenkin tarjoaa mahdollisuuden syöttää vain yksi predikaatti, edellisessä esimerkissä ”vihitty papiksi”. Kaksi muuta predikaattia tulevat määritellyiksi automaattisesti juoksevilla numeroinnilla. Relaatiot syötetään kirjoittamalla kukin relaation jäsen erikseen puolipisteellä erotettuna. Jäsenten järjestys on erittäin oleellinen, sillä kunkin jäsenen predikaatti määräytyy tämän perusteella. Siis esimerkiksi: “[**vihitty papiksi:=Rooma;1904**]”. Tällaisten relaatioiden käyttäminen vaatii, että predikaattia varten on tehty oma sivu, johon on merkattu, kuinka monta jäsentä relaatioon kuuluu.

Monipaikkaisten relaatioiden tuottaminen Semantic MediaWikissä on huomattavasti hankalampaa kuin linkkien yksinkertainen tyypittäminen. Voidaankin siis pohtia, täyttyykö enää Semantic MediaWikin tavoite muuttaa editorien työskentelyä niin vähän, että linkkien lisääminen syntyy lähes luonnostaan. Relaatioiden määrittelemisen erikseen omilla sivuillaan ei välttämättä ole ongelma, sillä jokaiselle predikaatille olisi joka tapauksessa hyvä tehdä oma sivu, jotta predikaatin oikea käyttö tulisi dokumentoitua. Kaikkien käyttäjien ei myöskään tarvitse osata tehdä tällaisia ohjesivuja. Itse annotointi saattaa kuitenkin olla hankalampaa normaaleille Wikipedian editoreille.

Vaikka tietolaatikoiden sisältöä ei testauksessa erityisesti annotoitu, havaittiin, että näidenkin sisällössä esiintyi aikaisemman listan ongelma numero 3. Gaugamelan taistelua käsittelevässä artikkelissa tietolaatikossa on ilmaistu erikseen oleelliset tiedot kummastakin taistelun osapuolesta. Sivupohjassa tämä on tehty käyttämällä parametreja: taistelija1, taistelija2, vahvuus1, vahvuus2, tappiot1, tappiot2 jne. Nämä eivät kuitenkaan ole hyviä predikaatteja RDF-lauseille, sillä valinta, kumpi on taistelija 1 ja kumpi taistelija 2, on mielivaltainen. Tietoa haettaessa ei voida tietää, onko Aleksanteri Suuri taistelija 1 vai 2 ja siten esimerkiksi kaikkien Aleksanteri Suuren taisteluiden hakeminen hankaloituu.

Lopputulokset

Sija	Nro	Kuljettaja	Talli	Kierrokset	Aika/keskeytys	Aika-ajo	Pisteet
1.	1	 Fernando Alonso	McLaren	78	1.40.29,329	1	10
2.	2	 Lewis Hamilton	McLaren	78	+ 4,095	2	8
3.	5	 Felipe Massa	Ferrari	78	+1.09,114	3	6
4.	3	 Giancarlo Fisichella	Renault	78	+1 kierros	4	5
5.	10	 Robert Kubica	BMW Sauber	78	+1 kierros	8	4

Kuva 6: Esimerkki Wikipediassa olevasta taulukosta. Artikkelista Monacon Grand Prix 2007.

Gaugamelan taistelun tietolaatikko on eräänlainen pieni taulukko, jossa jokaiselle taistelun osapuolelle on oma sarakkeensa. Testiin valituksi tulleet artikkelit eivät sisältäneet muita merkittäviä taulukoita, mutta lukuisat muut sivut Wikipediassa sisältävät taulukoita tai sen kaltaisia rakenteita. Esimerkiksi urheilukilpailujen tulokset, elokuvien näyttelijät ja toisessa maailmansodassa kuolleiden lukumäärät maittain. Eräs tällainen taulukko on esitetty kuvassa 6.

Kuvan 6 taulukko olisi mahdollista annotoida lisäämällä linkki- ja literaalityypit kuljettaja1, talli1, kuljettaja2, talli2 jne. Tämä tekee tiedonhausta kuitenkin huomattavan hankalaa. Esimerkiksi jos halutaan selvittää, missä talleissa Fernando Alonso on ajanut. Tallin ilmaiseva predikaatti riippuu siitä, monenneksiko Alonso sijoitui kilpailussa. Parempi tapa on lukea jokaiselta riviltä yksi monipaikkainen relaatio. Relaatioon kuuluvat, sijoitus, auton numero, kuljettaja, talli, ajatut kierrokset, loppuaika, aika-ajo sijoitus ja pisteet. Ensimmäiselle riville saadaan siis seuraavat RDF-lauseet.

```
(X, kuljettaja, Fernando Alonso)
(X, talli, McLaren)
(X, sijoitus, "1")
(X, aika, "1.40.29,329")
(X, tapahtuma, Monacon Grand Prix 2007)
(X, tyyppi, lopputulokset)
```

Tämä oltaisiin voitu koodata myös siten, että jokin relaation jäsenistä olisi asetettu subjektiksi. Ei kuitenkaan ole mitenkään selvää, mikä olisi oikea valinta. Tieto on esitetty Monacon Grand Prix'n sivulla, joten se olisi yksi vaihtoehto. Toisaalta

voitaisiin yhtä hyvin käyttää Alonsoa subjektina ja ilmaista hänen sijoitus eräässä kilpailussa.

Kun kaikista kilpailuista on tällaiset lauseet jokaiselle osanottajalle, voidaan helposti hakea kaikki Alonson tallit Grand Prix -kilpailuissa. RDF-kyselyllä vain haetaan kaikki sellaisten resurssien tallit, joilla kuljettajana on Alonso, joiden tyyppi on lopputulokset, ja joiden tapahtumana on Grand Prix tapahtuma. Myös Gaugamelan taistelun tietolaatikon tiedot voidaan annotoida näin, relaatiot on vain luettava sarakeittain.

Suurin puute Semantic MediaWikissä monipaikkaisiin relaatioihin liittyen on, että kaikkia predikaatteja ei voida määritellä. Numeroidut predikaatit tekevät sekä tiedon syötöstä että hausta hankalampaa, sillä jäsenet on aina oltava täsmälleen oikeassa järjestyksessä. Parempi menetelmä olisi sivupohjien tapaan käyttää nimettyjä parametreja. Predikaatin ja objektin voisi yksinkertaisesti erottaa toisistaan yhtäsuuruusmerkillä.

RDF-lauseiden syöttämiseen voitaisiin myös oikeasti käyttää sivupohjia, eikä vain ottaa niistä mallia. Editorit käyttäisivät niitä nykyiseen tapaan ja sivupohjat tuottaisivat tarvittavat RDF-lauseet. Jos järjestelmässä olisi mahdollisuus syöttää mitä tahansa RDF-lauseita, myös sellaisia, joissa subjektina on jokin muu kuin nykyinen artikkeli, niin sivupohjassa monipaikkaiset relaatiot voitaisiin koodata kulloinkin sopivalla mallilla. RDF-syntaksi yleisestikin jäisi suurimmalta osalta Wikipedian editoreja kokonaan piiloon. Käyttäen sivupohjien syntaksia aikaisempi esimerkki näyttäisi tältä:

```
{{Formula 1 lopputulokset
|kuljettaja=Fernando Alonso
|talli=McLaren
|sijoitus=1
|aika=1.40.29,239}}
```

Sivupohja voisi tuottaa kaikki tarvittavat RDF-lauseet nimettyjen parametrien avulla. Sivupohjan tekeminen luonnollisesti vaatii järjestelmän parempaa tuntemista, mutta tämä on tilanne jo nykyisinkin, sillä sivupohjat ovat usein varsin monimutkaisia.

Semantic MediaWikissä, ja useammassa muussa vastaavassa wikijärjestelmässä, voidaan sivuihin liittää RDF-kyselyitä, joiden tulokset näytetään taulukkona käyttäjälle. Niinpä varsinaisessa artikkelissa oleva taulukko on mahdollista luoda automaati-

tisesti, kunhan tarvittavat RDF-lauseet on syötetty. Hyöty ei ehkä ole ilmeinen, kun nykyinen tulostaulukko koodataan RDF:ksi, josta automaattisesti tehdään alkuperäinen tulostaulukko. Automaattisesti tuotettuja taulukkoja voidaan kuitenkin käyttää muillakin sivuilla. Koodaamalla tuloslistat RDF:ksi voidaan monille muille sivuille lisätä uusia kyselyiden tuottamia taulukoita. Kuljettajien sivuilla voidaan listata kaikki tallit, joissa he ovat ajaneet, sijoitukset eri osakilpailuissa, voitettujen kilpailujen lukumäärä ja niin edelleen.

7.4 Subjektin paikallisuus

Artikkelit käsittelivät toisinaan useampaa selkeää eri käsitettä. Esimerkiksi Queluz-palatsin artikkelissa käsitellään erikseen rakennuksen eri siipiä ja toisinaan jopa eri huoneita yksitellen. Eri siivet on rakennettu eri aikoina, ja ne ovat eri henkilöiden suunnittelema. Sillivalaalla puolestaan on pohjoinen ja eteläinen alalaji, joille on molemmille annettu hieman eri keskipituus ja -paino. Kummallakaan alalajilla tai Queluz-palatsin eri siivillä ei kuitenkaan ole omaa artikkelia Wikipediassa. Näiden asioiden annotointi tarkasti edellyttäisi joko uusien sivujen tekemistä Wikipediaan tai mahdollisuutta määritellä RDF-lauseita, joiden subjektina ei toimi lauseen sisältävä sivu. Edellisessä aliluvussa nähtiin myös, miten monipaikkaisten relaatioiden annotointi edellyttää yhtä ylimääräistä resurssia ja miten tässä yhteydessä olisi hyödyllistä määritellä RDF-lauseen subjektiksi jokin muu kuin nykyinen sivu.

Wikipediassa jotkin hyvin laajat artikkelit on jaettu osiin editoinnin ja lukemisen helpottamiseksi. Esimerkiksi lähes kaikki valtioita käsittelevät artikkelit ovat useammassa osassa. Englanninkielisessä Wikipediassa Suomen artikkelissa on linkit muun muassa artikkeleihin, joissa käsitellään tarkemmin Suomen sisällissotaa, Suomen maantiede, Suomen politiikka, uskonto Suomessa jne. Näillä sivuilla RDF-lauseiden subjektin pitäisi usein olla nimenomaan Suomi eikä esimerkiksi Suomen maantiede. Esimerkiksi Suomen allekirjoittamat ilmastositoumukset on mainittu vain Suomen maantiede sivulla. On selvää, että nämä sivut sisältävät oleellista annotoitavaa tietoa Suomesta.

On selvä tarve poistaa vaatimus, että subjektina on aina sivu itse. Tämä RDF-lauseiden paikallisuusvaatimus on kuitenkin oleellinen nykyisessä Semantic MediaWikissä [Völ06]. Sen ansiosta RDF-tietokannan ylläpitäminen sivuja muokattaessa on tehokasta, sillä sivuun tehdyt muutokset voivat vaikuttaa vain tiettyyn hyvin määritettyyn joukkoon RDF-lauseita. Sivua muokatessa RDF-tietokannasta poistetaan vanhat RDF-lauseet eli kaikki ne lauseet, joissa sivu on subjektina. Tilalle laite-

taan muokatusta sivusta luetut RDF-lauseet. Sivun alalaitaan tuotetaan myös automaattisesti lista aiheeseen liittyvistä annotoinneista. Paikallisuuden takia tämä lista voidaan tuottaa lukemalla pelkästään kyseisen artikkelin sisältö. Se ei ikinä sisällä lauseita, jotka on määritelty jossakin muussa artikkelissa.

Mikäli paikallisuusvaatimus ei olisi käytössä, olisi vanhojen RDF-lauseiden poistaminen hankalampaa. Nämä pitäisi joko lukea sivun edellisen version sisällöstä, tai tietokannassa täytyisi pitää kirjaa, millä sivulla kukin lause on määritelty. Olisi myös mahdollista, että sama RDF-lause on määritelty useammalla sivulla, jolloin sen poistaminen yhdeltä sivulta ei saa poistaa sitä kokonaan tietokannasta. Myös annotointilistan tuottaminen itsessään olisi työläämpää. Siihen pitäisi sisällyttää RDF-lauseita, jotka on voitu syöttää millä tahansa sivulla. Lauseet täytyisi siis hakea koko RDF-tietokannasta sen sijaan, että ne voidaan lukea suoraan artikkelista itsestään artikkelia talletettaessa.

Paikallisuusvaatimuksen poistaminen voi vaikuttaa myös käytettävyyteen. Editoreille ei välttämättä olisi aina selvää, millä sivulla jokin RDF-lause on määritelty. Asiallisessa käytössä lauseet varmasti pyrittäisiin sijoittamaan loogisiin paikkoihin, mutta oikea paikka lauseelle ei aina välttämättä ole selvä. Vandaalikäyttäjät saattaisivat myös sijoittaa RDF-lauseita satunnaisesti epämääräisiin paikkoihin. Järjestelmä siis tarvitsisi menetelmän etsiä, millä sivulla jokin RDF-lause on määritelty. On myös huolehdittava siitä, että lukituille sivuille ei voida lisätä RDF-lauseita muiden sivujen kautta.

Paikallisuusvaatimuksen täydellisen poistamisen sijaan sitä voitaisiin myös vain lieventää. Tässä tutkielmassa on havaittu kolme tapausta, joissa paikallisuus hankaloihtaa annotointia:

1. Monipaikkaisten relaatioiden annotointi. Monipaikkaisen relaation osapuolet esiintyvät RDF-lauseissa objekteina, subjektina on kaikissa nimetön kokoa-va RDF-resurssi. Tätä resurssia ei vastaa mikään artikkeli Wikipediassa eikä myöskään ole mielekästä tehdä sille artikkelia. Tämä tapaus voidaan hoitaa käyttämällä omaa syntaksia monipaikkaisille relaatioille, kuten Semantic MediaWiki osittain jo tekee.
2. Sivun aiheen osien tai alakäsitteiden annotointi. Alakäsitteille tai osille voisi periaatteessa tehdä Wikipediaan oman artikkelinsa, mutta niitä ei ole katsottu riittävän merkittäviksi itsenään. Alakäsitteet voidaan nimetä järkevästi.
3. Pääartikkelin annotointi alasivulla. Laaja aihe on jaettu useampaan artikke-

liin. Alasivuilla RDF-lauseiden subjektina pitäisi toimia pääartikkeli.

Ei siis välttämättä ole tarvetta määritellä RDF-lauseita, joissa subjektina voi olla mikä tahansa RDF-resurssi. Riittää, että järjestelmä pystyy käsittelemään tarpeelliset tapaukset.

7.5 Wikipedian kategoriat

Wikipedian annotointi RDF:llä aiheuttaisi useita muutoksia nykyiseen kategoriakäytäntöön. Kuten aikaisemmin todettiin, Wikipedian kategoriahierarkia toimii lähinnä navigoinnin helpottamiseksi ei niinkään täsmällisenä ontologiana [SKW07a]. Vain pieni osa kategorioista kuvaa aiheen tyyppiä. Esimerkiksi Ulysses-runon artikkelilla on kolme kategoriata: 1833 runot, 1842 runot ja Alfred Tennysonin runot. Kaksi ensimmäistä kategoriata viittaa runon kirjoittamis- ja julkaisuvuosiin, viimeinen runoilijaan. Nämä kaikki tulisi annotoiduiksi omiksi RDF-lauseiksi, täsmällisellä predikaatilla, joten kategoriat ovat turhia. Toisaalta pitäisi lisätä yksinkertainen kategoriata runot korvaamaan kaksi ensimmäistä kategoriata. Vastaava tilanne oli lähes kaikissa muissakin testatuissa artikkeleissa. Useissa artikkeleissa esiintyi lisäksi kategorioita, jotka on tarkoitettu Wikipedian hallintaan.

Lisäksi Semantic MediaWiki tulkitsee olemassa olevat kategoriat resurssin tyyppiä, jolla on selkeä semantiikka RDF:ssä sekä RDFS- ja OWL-laajennoksissa. Vain osa nykyisistä kategoriamäärittämisistä on sellaisia, joissa on soveliaa käyttää RDF:n omaa tyyppimäärittäystä. Olisi selvästi väärin sanoa, että banaaniäpärpän on perinnöllisyystiede, joka on yksi banaaniäpärpän nykyisistä kategorioista. Tämänkin takia nykyistä kategoriajärjestelmää täytyisi muuttaa.

Luonnollinen tapa korvata poistetut kategoriat on käyttää RDF-hakuja. Kaikki vuonna 1833 kirjoitetut runot on helppo hakea RDF-tietokannasta, kun annotoinnit on tehty. RDF-haut tarjoavat myös huomattavasti monipuolisemmat haut kuin kategoriat, sillä hakuehtoja voidaan yhdistellä mielivaltaisesti. Lisäksi nykyinen kategoriajärjestelmä on rajoittunut vain tapauksiin, joille on ennalta päätetty tehdä kategoriat.

Peruskäyttäjää varten tarvitaan kuitenkin menetelmä suorittaa ainakin nykyisiä kategorioita vastaavat haut ilman, että täytyy tuntea hakukieltä. Semantic MediaWiki tekee joitakin valmiita hakulinkkejä sivun alalaitaan tuotettavaan annotaatiolistaukseen. Linkeillä voidaan hakea muita artikkeleita, joissa on jokin yksittäinen vastaava RDF-lause kuin nykyisellä sivulla. Vastaavalla RDF-lauseella tarkoitetaan lausetta,

jossa on sama predikaatti ja objekti kuin nykyisen sivun lauseessa. Ulyssesin tapauksessa voidaan siis esimerkiksi hakea kaikki muut runot tai kaikki muut resurssi, joissa on kirjoitusvuosi 1833, tämä saattaa sisältää paljon muitakin sivuja kuin vain runoja. Tällä menetelmällä ei siis saavuteta nykyistä toiminnallisuutta. Yhdistämällä kaksi ehtoa saadaan kaikki vuoden 1833 runot eli nykyistä kategorialla vastaavat artikkelit. Tämä vastaa joissakin semanttisen webin sovelluksissa käytettyä moninäkökymähakua [MHS06].

7.6 Ontologioihin liittyvät havainnot

Semantic MediaWiki tukee RDFS tai OWL-päättelyä erittäin rajoittuneesti. Se tuntee ainoastaan alaominaisuuden (`subPropertyOf`) ja alaluokan (`subClassOf`) käsitteet. Pelkästään näistä on jo hyötyä Wikipedian annotoinnissa. Eläimiä käsittelevissä artikkeleissa mainittiin lukuisia eri pituuksia, painoja ja elinikiä. Artikkeleissa annettiin tyypillisesti eri lukuarvot eri sukupuolille, maksimi ja keskimääräiset lukuarvot sekä erikseen maksimi luonnossa ja vankeudessa eläneille yksilöille. Nämä voidaan kaikki annotoida eri ominaisuuksiksi eri predikaateilla. Jokainen erilainen pituus määritellään yleisen pituus ominaisuuden alaominaisuudeksi. Täten eläinlajit löytyvät hakemalla pituus-ominaisuudella sen sijaan, että täytyy hakea esimerkiksi keskimääräisellä pituudella. Järjestelmässä kuitenkin säilyy mahdollisuus hakea nimenomaan maksimipituudella, mikäli ollaan kiinnostuneita juuri tästä ominaisuudesta.

Edellisessä aliluvussa mainittiin, miten Wikipedian nykyiset kategoriamääritykset eivät sovellu täsmälliseksi ontologiaksi. Tämä aiheuttaa ongelmia, sillä Semantic MediaWiki tulkitsee alakategoriat RDFS:n mukaisesti: Jokainen tiettyyn kategoriin kuuluva artikkeli kuuluu myös kaikkiin kyseisen kategorian yläkategorioihin. Tästä aiheutuu se, että esimerkiksi RDF-haku, joka hakee kaikki nisäkkäät, antaa vastauksena muun muassa silmälasit. Tämä johtuu siitä, että nisäkkäiden alakategoriana on nisäkkäiden anatomia, jonka alta löytyy kategoriat ihmisen anatomia, silmä ja optometria, joka on syötetty silmälasien luokaksi.

Vaikkei Semantic MediaWiki tuekaan OWL-päättelyä, voidaan OWL-lauseita kuitenkin määritellä sivuille muiden RDF-lauseiden tapaan. Wikipedian RDF-tietokanta voidaan ottaa käyttöön jossakin muussa järjestelmässä, joka osaa hyödyntää näitä lauseita. Olisi siis toivottavaa, että Wikipediassakin käytettäisiin tarkkaan mietittyä ja määritettyä ontologiaa riippumatta siitä, otetaanko Wikipediassa itsessään käyttöön OWL:a hyödyntäviä ominaisuuksia vai ei. Yhdenmukaiset annotaatiot koko

Wikipediassa myös parantavat hakuominaisuuksia. Hyvän ontologian kehittäminen ei kuitenkaan aina ole mitenkään suoraviivaista. Etenkin Wikipediassa tämä saattaa osoittautua ongelmalliseksi, sillä aihepiiriä ei ole lainkaan rajoitettu.

Testatessa etukäteen määritellyn ontologian puute ilmeni siten, että aina ei ollut selvää kannattaisiko jokin tietty asia annotoida vai ei. Predikaatin tarkka nimi ei myöskään ole mitenkään yksiselitteinen. Olemassa oleva ontologia määritteli, mitä predikaatteja artikkeleissa tulisi ja voisi käyttää. Testissä annotoitiin useampia eläinlajeja koskevia artikkeleita. Tästä muodostui implisiittisesti pieni eläinontologia. Annotoitaessa oikean predikaatin valitseminen onnistui useimmiten lähinnä siksi, että sama henkilö annotoi jokaisen artikkelin ja kaikki artikkelit käsiteltiin lyhyessä ajanjaksossa. Tällöin on helppo muistaa, mitä predikaattia käytettiin aikaisemmissa artikkeleissa. Tästä huolimatta testissä piti toisinaan tarkistaa aikaisemmin annotoiduilta sivuilta oikea predikaatin nimi.

Tietolaatikoiden laaja käyttö useilla aihealueilla viittaa siihen, että ainakin kevyehköjä yhden aihepiirin kattavia ontologioita olisi mahdollista määritellä. Sivupohjien ja tietolaatikoiden hyödyntäminen annotoinnissa olisi hyödyllistä myös tiedon yhteisyyden kannalta. Annotoitu tieto tulisi määritellyn ontologian mukaiseksi ilman, että editorien tarvitsee edes tietää mitään ontologioista. Editorien täytyisi vain käyttää tietolaatikoita ja muita sivupohjia nykyiseen tapaan, ja sivupohjat tuottaisivat ontologian mukaiset RDF-lauseet. Toisaalta joissakin tapauksissa tämä edellyttäisi nykyisten tietolaatikoiden miettimistä uudelleen. Esimerkiksi juuri eläinten tietolaatikat ovat melko vaatimattomia, eivätkä sisällä juuri mitään muuta tietoa kuin eläimen tieteellisen nimen ja taksonomian, uhanalaisuuden, kuvan ja levinneisyyden karttana.

Sivupohjat määrittellään omissa artikkeleissaan, jotka sisältävät myös käyttöohjeet. Käyttöohjeissa kerrotaan, mitä parametreja sivupohjassa voidaan käyttää, mitä parametreja on käytettävä ja parametrien mahdolliset arvot. Tämä vastaa oleellisesti RDFS- ja OWL-määrittelyä, mutta tekstimuodossa. Tämän tiedon annotointi muiden normaalien artikkelin tapaan siis määritteli Wikipedian pieniä sivupohjan aihepiirin kattavia ontologioita. Kun tämä ontologia on koneellisesti luettavissa, olisi artikkelin tallennuksen yhteydessä mahdollista tarkastaa, että kaikkia parametreja käytetään oikein.

Formaali ontologia edellyttäisi, että Wikipedian sisältämää tietoa ainakin pyrittäisiin pitämään ontologian mukaisena, muuten ontologia käytännössä menettää merkityksensä. Jochen et. al. [Joc06] mainitsevat joitakin ongelmia ontologian käsitte-

lemisestä wikiympäristössä. Tarvitaan jonkinlainen lukitsemismenettely tai tarkastus, että useamman henkilön yhtäaikaiset muokkaukset ovat keskenään yhteensopivia. Lisäksi Wikipedian nykyinen mahdollisuus palauttaa sivut aikaisempaan versioon yksitellen ei välttämättä riittäisi. Aikaisempi versio ei nimittäin välttämättä ole yhteensopiva muiden nykyisten sivujen kanssa vaikka se aikoinaan olisikin ollut yhteensopiva.

Niiltä osin kuin Semantic MediaWiki tukee ontologiamäärittelyä se käyttää vapaampaa menettelyä. Ontologian vastaisista muutoksista varoitetaan käyttäjää tallennuksen yhteydessä, mutta muutokset tallennetaan silti sellaisenaan. Kuten todettiin aikaisemmin, jotkin sivupohjat vastaavat nykyisin ontologiamäärittelyä. Näiden versiointia tai lukitsemista ei kuitenkaan hallinnoida mitenkään eri tavalla kuin muidenkaan sivujen. Monet sivupohjat on lukittu, mikä tarkoittaa sitä, että kuka tahansa ei voi muokata niitä. Tämä varmasti helpottaa niiden hallintointia. Koska sivupohjat toimivat nykyisin ilman suurempia ongelmia, Semantic MediaWikin käytäntö saattaa hyvinkin olla riittävä, ainakin niin kauan kuin Wikipediassa pysytään melko kevyissä ontologioissa. Lisäksi olisi mahdollista tehdä erilaisia työkaluja tukemaan ontologian hallintaa. Samaan tapaan kuin Wikipediassa on nykyisin mahdollista etsiä sivut, joille ei ole yhtäkään linkkiä, voitaisiin tehdä työkalu, joka etsii sivut, jotka ovat ristiriidassa ontologian kanssa.

8 Tiedonlouhintamenetelmien testaus

Tässä luvussa käsitellään tiedonlouhintamenetelmien testaus. Aluksi esitellään testimenetelmät ja tämän jälkeen havainnot erikseen kahdesta testatusta järjestelmästä, YAGO:sta ja DBPediasta.

8.1 Testimenetelmät

Tiedonlouhinnan menetelmiä testattiin vastaavalla tavalla verrattuna edellisen luvun menetelmiin. Testiin valittiin yhteensä viisitoista suositeltua sivua. Tässä tapauksessa kaikki annotoinnit oli tehty automaattisesti testatuilla järjestelmillä ja testauksessa tarkasteltiin annotoinnin tulosta. Testaus tehtiin erikseen DBPedialle [Aue07, AuL07] ja YAGO:lle [SKW07a, SKW07b]. DBPedia jäsentää artikkelien tietolaatikoiden ja muiden sivupohjien parametreja ja muodostaa näiden perusteella RDF-tietokannan. Yago tarkastelee artikkelien kategorioita ja sijoittaa ne WordNet-

	Artikkeli	Kieli	Aihe
1	Gregory of Nazianzus	en	300-luvun kirkkoisä
2	Fin Whale	en	Sillivalas
3	William Cooley	en	1800-luvun Amerikan uudisasuttaja
4	Oregon State Capitol	en	(TODO)
5	Boshin War	en	Japanin sisällissota 1800-luvulla
6	Johannes XXIII	fi	1900-luvun paavi
7	Gaugamelan taistelu	fi	Aleksanteri Suuren taistelu
8	Maasait	fi	Itäafrikkalainen kansa
9	Paavali VI	fi	1900-luvun paavi
10	Oulu	fi	Kaupunki Suomessa
11	G. Ledyard Stebbins	en	Amerikkalainen biologi
12	Ælle of Sussex	en	Saksien kuningas 400-luvulla
13	Chicxulub Crater	en	Kraateri Keski-Amerikassa
14	Swedish emigration to the United States	en	Ruotsalaiset siirtolaiset Yhdysvalloissa
15	Trembling Before G-d	en	Dokumenttielokuva

Taulukko 4: Tiedonlouhinnan testauksessa käytetyt Wikipedian artikkelit.

thesauruksen ontologiaan. Lisäksi YAGO jäsentää tietoa joistakin sivupohjista DBPedian tapaan.

Testiin valittiin osittain samoja artikkeleita kuin edellisessä luvussa. Edellisen luvun testiartikkeleista 1-10 valittiin sellaiset, joissa oli käytetty tietolaatikkoa. Sellaiset artikkelit, joissa ei ollut tietolaatikkoa vaihdettiin sellaisiin suositeltuihin sivuihin, joissa tietolaatikkoa on käytetty. DBPedia perustuu pääosin tietolaatikoiden jäsentämiseen, joten sitä ei ole mielekästä testata aineistolla, joissa tietolaatikoita ei ole käytetty. On kuitenkin syytä muistaa, että aineiston valinnasta johtuen, kaikkia Wikipedian sivuja ei tällä menetelmällä välttämättä saada annotoitua yhtä hyvin. Lisäksi valittiin viisi muuta satunnaista englanninkielistä suositeltua artikkelia, sillä YAGO kykenee tulkitsemaan vain englanninkielisiä artikkeleita. Kaikki testiartikkelit on esitelty Wikipedian etusivulla 2007-2008 vuodenvaihteessa. Taulukossa 4 on esitetty testiaineisto.

Annotointien tarkastelussa kiinnitettiin huomiota siihen, kuinka hyvin artikkelin oleelliset asiat on annotoitu, mitä muita asioita on annotoitu, onko virheellisiä annotointeja sekä ovatko annotaatiot käyttökelpoisia.

(Johannes XXIII,	paavi nimi,	fi:"Johannes XXIII \n Ioannes XXIII")
(Johannes XXIII,	kuva,	Ioannes XXIII.jpg)
(Johannes XXIII,	oikeaNimi,	fi:"Angelo Giuseppe Roncalli")
(Johannes XXIII,	paaviusAlkoi,	28. lokakuuta)
(Johannes XXIII,	paaviusPäättyi,	3. kesäkuuta)
(Johannes XXIII,	edeltäjä,	Pius XII)
(Johannes XXIII,	seuraaaja,	Paavali VI)
(Johannes XXIII,	syntymäAika,	25. marraskuuta)
(Johannes XXIII,	syntymäPaikka,	Sotto il Monte)
(Johannes XXIII,	syntymäPaikka,	Italia)
(Johannes XXIII,	kuolinAika,	3. kesäkuuta)
(Johannes XXIII,	kuolinPaikka,	Apostolinen palatsi)
(Johannes XXIII,	kuolinPaikka,	Vatikaani)

Taulukko 5: DBPedian tietolaatikon perusteella tuottamat RDF-lauseet suomenkielisen Wikipedian artikkelista Johannes XXIII.

Edellisessä luvussa käsiteltiin useita ongelmia, jotka liittyivät tiedon esittämiseen RDF:llä eikä suoranaisesti Semantic MediaWikiin tai mihinkään muuhun semanttiseen wikijärjestelmään. Nämä ongelmat säilyvät siis myös tiedonlouhintaan perustuvissa menetelmissä. Näitä ei käsitellä tässä luvussa uudelleen.

8.2 Havainnot DBPediasta

DBPedian tekemissä annotoinneissa havaittiin huomattavan paljon erilaisia virheitä tai puutteita. Taulukossa 5 on esitetty DBPedian tuottamat RDF-lauseet suomenkielisen Wikipedian artikkelissa Johannes XXIII. Taulukossa on vain lauseet, jotka on tuotettu tietolaatikon perusteella. Saman tietolaatikon lähdekoodi ja graafinen esitys on kuvassa 2.

Artikkeli Johannes XXIII edustaa yhtä parhaiten annotoiduista testatuista artikkeleista. Siinäkin on kuitenkin puutteita ja virheitä: Paavin nimessä on kaksi eri versiota yhdessä literaali arvossa, päivämääriin ei ole tullut vuosia mukaan lainkaan, ja syntymä- ja kuolinpaikat on jaettu kahteen eri lauseeseen, joista toinen on turha.

DBPedia on erittäin riippuvainen sivupohjista ja niiden parametrien muotoilusta. Järjestelmä yrittää jäsentää parametreja ja toisinaan saakin niistä muodostettua

listoja. Esimerkki jossa tämä toimii on Gregorios Nazianzilainen, jonka tittelit on jokainen annotoitu siististi omaksi RDF-lauseekseen. Samassa artikkelissa on myös hyvä esimerkki, jossa menetelmä ei toimi. Pyhimyksien tietolaatikoissa on kuvaus, miten pyhimys yleensä esitetään ikonografiassa. Tässä tapauksessa tietolaatikossa tämän parametrin arvo on seuraava:

```
“[[Vestments|Vested]] as a [[bishop]], wearing [[omophorion]]; holding a [[Gospel Book]] or [[scroll]]. [[Iconography|Iconographically]], he is depicted as balding with a bushy white beard.”
```

Tämän perusteella DBPedia jäsentää kuusi erillistä RDF-lausetta, joilla on objektit vestments, bishop, omophorion, gospel book, scroll ja iconography. Toisin sanoen DBPedia käyttää tässä tapauksessa vain linkkejä jättäen muun kuvailun täysin pois. Lisäksi ensimmäinen ja viimeinen linkki on lähinnä vain käyttäjiä varten, eivätkä ne varsinaisesti kuvaa pyhimyksen ikonografista esitystä.

Asioita on tietolaatikoissa myös listattu monilla eri tavoilla. Toisinaan asiat on erotettu pilkuilla, rivinvaihdolla, puolipisteillä tai kuten yllä, ne on vapaassa tekstissä. DBPedialla on toisinaan vaikeuksia jäsentää listat oikein. Joskus listan jäsenissä on käytetty linkkejä ja toisinaan ei. Wikipediassa on tapana, että yhdessä artikkelissa samaan artikkeliin ei tehdä linkkiä kahta kertaa lähekkäin tekstissä. Esimerkiksi Gregorios Nazianzilaisen synnyinpaikkaa kuvaavan parametrin arvo on “[[Arianzum]], [[Cappadocia]]” ja heti perään kuolinpaikan arvo on “Arianzum, Cappadocia.” Synnyinpaikan DBPedia on jäsentänyt listana ja siitä tehdään kaksi RDF-lausetta, mutta kuolinpaikan DBPedia jäsentää yhtenä literaaliarvona. Tämä on myös esimerkki, jossa ei oikeastaan ole edes kyse listasta useita paikkoja. Nazianzilaisen synnyinpaikka on Arianzum eikä mitään muuta. Jälkimmäinen linkki on lähinnä lukijaa varten. Sen annotoiminen RDF:ksi on kuitenkin turhaa, sillä Arianzum määrittää synnyinpaikan jo täsmällisesti, huolimatta siitä, että samannimisiä paikkoja saattaa olla useita. Wikipedian artikkelien nimet ovat yksilöllisiä ja niissä on jo tarvittaessa käytetty tarkentimia erottamaan samannimiset käsitteet toisistaan.

Literaaliarvoja DBPedia ei jäsennä lainkaan, mikä aiheuttaa lukuisia ongelmia. William Cooleyn lapset annetaan yhdessä parametrissa rivinvaihdolla erotettuna. Yhdellekään lapselle ei ole omaa artikkelia Wikipediassa, joten parametrin arvo ei sisällä linkkejä. Kaikki lapset tulevatkin saman RDF-lauseen literaali arvoksi sen sijaan, että jokaiselle tehtäisiin oma RDF-lause. Lisäksi jokaisen lapsen syntymä- ja kuolin-aika annetaan nimien perässä sulkeissa. Yksi RDF-lause siis sisältää tietoa kolmesta eri henkilöstä ja jokaisesta nimen sekä synnyin- ja kuolinajan. Ihminen kyllä ym-

märtää, mistä on kyse, mutta tietokoneen kannalta tällainen RDF-lause on lähes käyttökelpoton.

Literaalien vaihteleva muotoilu aiheuttaa myös ongelmia. Esimerkiksi Maasaiden lukumääräksi annetaan: “noin 883 000.” Tämä tulee sellaisenaan RDF-lauseen literaaliohjelmaksi. Noin sana arvon alussa hankaloittaa tiedonhakua merkittävästi. Vastavia tapauksia oli useissa parametreissa. Päivämäärien esitysmuoto myös vaihtelee, mikä tekee tietyn päiväyksen etsimisestä hankalaa.

Wikipediassa on mahdollista käyttää sivupohjien parametrien arvoissa muita sivupohjia. Gregorius Nazianzilaisten artikkelissa kuolinajassa on käytetty omaa sivupohjaa. Kyseinen sivupohja muotoilee päiväyksen ja tekee linkit vuotta ja päivämäärää vastaaviin artikkeleihin. Lisäksi sivupohjan ohjeistus mainitsee, että sivupohjalla voidaan lisätä metatietoa artikkeliin, jonka käyttäjän selain voi lukea. DBPedian tapauksessa erillinen sivupohja kuitenkin hankaloittaa tiedon jäsentämistä, eikä kuolinaikaa tule missään muodossa annotoiduksi. Sivupohjia on käytetty parametrien arvona myös muissa artikkeleissa esimerkiksi maantieteellisten koordinaattien esittämiseen sekä maiden lippujen ja nimien esittämiseen. Näissäkin tapauksissa DBPedia ei saa jäsenettyä tietoa kunnolla, eikä se siten tule annotoiduksi.

DBPedian automaattinen annotointi tuotti lisäksi lukuisia turhia RDF-lauseita. Nämä ovat seurausta artikkeleissa käytetyistä sivupohjista, jonka parametreja DBPedia yrittää jäsentää. Tämä on kuitenkin täysin turhaa useissa tapauksissa. Esimerkiksi suomenkielisissä artikkeleissa kaikkien lähteiden tiedot tulivat annotoiduksi itse artikkeliin. Englanninkielisissä artikkeleissa näin ei käynyt, ilmeisesti DBPedia osaa jättää joitakin sivupohjia huomiotta. Lähteiden esittämisessä käytetyn sivupohjan lisäksi muita yleisiä tapauksia olivat muun muassa mittayksiköiden esittämisessä käytetyt sivupohjat sekä ulkomaisten sanojen translitteraatioissa ja kääntämisessä käytetyt sivupohjat.

Joissakin artikkeleissa DBPedia on lisännyt annotaatioita, jotka sisältävät suorastaan väärää tietoa artikkelista. Sotien tietolaatikoissa kerrotaan sodan seuraukset. Boshin-sodan artikkelissa annetaan seuraukset parametrille arvoksi “[Late Tokugawa shogunate|End of the Shogunate]”. Käyttäjä siis näkee seurauksina, shogunaatin päättymisen. Wikipedia kuitenkin jättää huomiotta linkin muotoilun ja ottaa RDF-lauseen objektiksi vain kyseisen artikkelin eli Tokugawa shogunaatin. DBPedian RDF:ää tutkimalla saa siis käsityksen, että Tokugawa shogunaatti oli Boshin sodan seurauksena shogunaatti päättyi.

Toinen tapaus väärästä informaatiosta on sillivalaan artikkelissa. Eläinten taksono-

mia esitetään tietolaatikossa, jonka yhtenä parametrina on tietolaatikossa käytetty väri. DBPedia kuitenkin asettaa RDF-lauseiden subjektiksi artikkelin itsensä, kun tässä tapauksessa oikea subjekti olisi tietolaatikko. Tämän seurauksena DBPedian mukaan sillivalas on väriltään pinkki. Vastaavia tilanteita esiintyi joidenkin muiden sivupohjien kanssa, mutta usein lauseen predikaatti ei ole mielekäs artikkelin kanssa, joten RDF-lauseen sivuuttaa roskana. Lähteiden esittämisessä käytetty sivupohja tuottaa usein lauseita, joiden predikaatti on esimerkiksi vuosi, julkaisija tai osoite. Nämäkin voivat aiheuttaa sekaannuksia joidenkin artikkelien yhteydessä.

Edellisessä luvussa mainittiin Gaugamelan taistelun tietolaatikko esimerkkinä monipaikkaisista assosiaatioista. Tämä aiheuttaa ongelmia myös DBPediassa. Artikkeleissa käytetään parametreja `taistelija1`, `taistelija2`, `komentaja1`, `komentaja2` ja niin edelleen. Käyttämällä näitä parametreja tietosisältö saataisiin koodattua RDF:ksi, ainoa ongelma on, että valinta kumpi osapuoli on taistelija 1 ja kumpi taistelija 2 on mielivaltaisen. DBPedia kuitenkin jättää parametrien lopusta numeron pois. Ilmeisesti syynä on se, että tällaisia parametrinimiä käytetään muissa artikkeleissa listaamaan asioita. Tämä aiheuttaa sen, että taisteluun liittyy kaksi RDF-lausetta, joiden predikaatti on `taistelija` ja kaksi lausetta, joiden predikaatti on `komentaja`. RDF-lauseista ei mitenkään käy ilmi, kumpi komentajista komensi kumpaa armeijaa. Kuten edellisessä luvussa esitettiin, tämä pitäisi mallintaa monipaikkaisena relaationa.

Yleisesti ottaen jokaisessa testatussa artikkelissa oli jotakin huomautettavaa annotoinnista. Kaksi paavia käsittelevää artikkelia, `Paavali VI` ja `Johannes XXIII`, olivat laadultaan parhaimmat. Osa havaituista ongelmista oli samoja, jotka järjestelmän kehittäjät olivat myös havainneet [AuL07]. DBPedian tuottamassa annotoinnissa on siis väärää tietoa, ylimääräisiä RDF-lauseita ja siitä puuttuu tietoa. Virheitä tai puutteita on lisäksi huomattavan paljon. Täten nykyisessä muodossaan DBPedian tietokannasta ei voi luotettavasti hakea täsmällistä tietoa.

Suuri osa virheistä aiheutuu sivupohjissa esiintyvistä vaihtelevista käytännöistä. Tämä taas puolestaan epäilemättä johtuu siitä, että sivupohjia ei ole tarkoitettu automaattiseen tiedon jäsentämiseen vaan editoinnin helpottamiseen ja sivujen ulkoasujen yhtenäistämiseen. DBPedian kehittäjät ehdottivatkin joitakin muutoksia sivupohjien käyttötapoihin ja itse MediaWiki järjestelmään, jotka parantaisivat automaattisen tiedon jäsennyksen toimivuutta [AuL07]. Tiedon esitys ja asiasisältö pitäisi paremmin erottaa toisistaan. Esitys tulisi olla kokonaan sivupohjassa ja asiasisältö parametreissa. Parametrien käytössä tulisi kiinnittää huomiota täsmällisyy-

teen. Samaa parametria käytetään aina samalla tavalla. MediaWikin tulisi myös tarjota mahdollisuudet määritellä sivupohjien parametrien tyypit tarkemmin, jolloin tiedettäisiin, miten parametrin arvo tulisi jäsentää. Myös kunnollinen tuki eri mittayksiköille helpottaisi jäsentämistä.

8.3 Havainnot YAGO:sta

Yago eroaa DBPediasta merkittävästi kahdella tavalla. Se muodostaa perusluokarakenteen käyttäen WordNetin relaatioita ja jättää Wikipedian kategoriahierarkian käytännössä huomiotta. Wikipedian kategorioiden avulla artikkelit sijoitellaan WordNetistä saatuun hierarkiaan. Toinen merkittävä ero on, että tiedon louhinnassa Wikipedian tietoa hyödynnetään vain sellaisissa paikoissa, joissa sen on havaittu olevan helposti tulkittavissa koneellisesti. Ei siis yritetäkään jäsentää kaikkea mahdollista vaan pelkästään pieni osa luotettavaa tietoa. Näin saadaan RDF-tietokanta, joka sisältää huomattavasti vähemmän virheellisiä tai turhia lauseita, mutta kattavuus on myös selvästi pienempi. Nämä erot DBPediaan verrattuna ovat selvästi nähtävissä automaattisen annotoinnin lopputuloksessa.

YAGO:n RDF-lauseiden havaittiin sisältävän erittäin vähän virheitä tai ylimääräisiä lauseita. Testatuista artikkeleista ainoa, jossa YAGO tuottaa selvästi väärää tietoa on Gregorius Nazianzilainen. Hänet on Wikipediassa luokiteltu kategorioihin “church fathers” (kirkkoisät) ja “doctors of the church” (kirkonopettajat). Näillä molemmilla on tarkka teologinen merkitys ja itseasiassa myös WordNet sisältää kummankin käsitteen. YAGO on kuitenkin tulkinut luokat väärin ja saanut niille yläluokat isä ja lääketieteen tohtori.

YAGO:n kehittäjät ovat tehneet kattavammat tutkimukset YAGO:n sisältämien RDF-lauseiden oikeellisuudesta [SKW07a]. He ovat tutkineet RDF-lauseita eri predikaateilla, kutakin predikaattia kohden noin sadasta kolmeensataan lausetta, ja saaneet yleiseksi luotettavuudeksi noin 97%. Luotettavuus vaihteli eri predikaattien välillä noin 90 prosentista lähes 99 prosenttiin. Tämän tutkielman yhteydessä huomattavasti pienemmällä otoksella päädyttiin samankaltaisiin tuloksiin.

YAGO:ssa tiedon kattavuus on huomattavasti pienempi kuin DBPediassa. Testatut artikkelit sisälsivät annotoinnit pääasiassa vain Wikipedian kategorioista ja WordNetistä saaduista aiheen tyypeistä. Henkilöille oli saatu synnyin- ja kuolinaika ja -paikka. Osa henkilöiden nimistä oli saatu jäsennettyä etu ja sukunimeksi ja William Cooleyille oli merkattu puolue. Juuri muita RDF-lauseita ei esiintynyt paitsi

Trembling before G-d -dokumenttielokuvassa, joka sisälsi RDF-lauseet elokuvan ohjaajasta, tuottajasta, pituudesta, kielestä sekä linkin elokuvan kotisivuille ja IMDB-sivustolla ⁷ käytetyn elokuvan tunnistenumeron. Tämän artikkelin selkeästi kattavampi annotointi johtuu siitä, että elokuvissa käytetty tietolaatikko on yksi niistä sivupohjista, joita YAGO hyödyntää.

9 Johtopäätökset

Tutkielmassa testattiin kahta eri lähestymistapaa hyödyntää semanttisen webin tekniikoita Wikipediassa. Wikipedian artikkeleita voidaan annotoida manuaalisesti tai Wikipedian nykyistä sisältöä voidaan yrittää jäsentää ja muodostaa siitä tietämyskanta. Testeissä käytettiin saatavilla olevia valmiita ohjelmistoja ja Wikipedian todellisia suomen- ja englanninkielisiä artikkeleita.

Kumpikin testattu lähestymistapa sisälsi lukuisia ongelmia tai jatkokehitystä vaativia asioita. Tästä huolimatta molempia tapoja voidaan jo nyt hyödyntää ja ne tarjoavat hyödyllisiä uusia ominaisuuksia Wikipediaan nähden. Automaattisesti tietoa louhivat DBPedia ja YAGO molemmat tarjoavat verkkokäyttöliittymän semanttiseen tietoon. Näillä liittymillä voidaan suorittaa RDF-kyselyitä louhittuun tietoon. Hakuliittymät antavat maistiaisen siitä, mitä RDF-hauilla voisi tehdä semanttisessa Wikipediassa. Vaikkei Semantic MediaWikiä ainakaan vielä käytetä Wikipediassa, on se kuitenkin käytössä joissakin muissa wikisivustoissa.

DBPedian ja YAGO:n tapauksessa ei kuitenkaan voida puhua varsinaisesta semanttisesta wikistä. Näissä itse Wikipedia pysyy muuttumattomana, siitä vain louhitaan automaattisesti tietoa mahdollisuuksien mukaan. Nämä tiedonlouhintamenetelmät kuitenkin tuovat hyvin esille joitakin Wikipedian nykyisiä ongelmia ja niistä voidaan myös sovelletusti ottaa mallia kehitettäessä varsinaista semanttista wikiä.

Tutkielmassa keskityttiin semanttisten wikijärjestelmien osalta pääasiassa Semantic MediaWikiin. Tämä järjestelmä lähinnä vastaa sitä, mitä semanttinen Wikipedia nykyisillä tekniikoilla voisi olla. Tutkielmassa mainitut muut järjestelmät ovat pääpiirteiltään sen verran samankaltaisia, että suurin osa Semantic MediaWikissä havaituista ongelmista esiintyy myös näissä järjestelmissä. Tutkielmassa ei testattu järjestelmien suorituskykyä, missä saattaa hyvinkin olla eroja eri järjestelmien välillä, kun aineiston ja käyttäjien määrä kasvaa Wikipedian mittoihin.

⁷Internet Movie Database <http://www.imdb.com>

Ennen Semantic MediaWikin, tai minkä tahansa muun wikijärjestelmän, integroimista Wikipediaan tulisi selvittää, kuinka tässä tutkielmassa havaitut ongelmat ratkaistaan. Semanttisten ominaisuuksien hyödyntäminen aiheuttaa lukuisia muutoksia Wikipediassa, osa näistä melko suuria. Seuraavassa on esitetty merkittäviä tutkielmassa havaittuja Wikipedian käyttöön vaikuttavia tekijöitä:

1. Tarvitaan yleiset käytännöt semanttisen tiedon syöttämisestä. Minkälaista tietoa halutaan annotoida sekä missä ja miten annotoinnit tehdään?
2. Miten hallitaan eri aihealueita koskevat ontologiat? Vaikkei ontologioita varsinaisesti määriteltäisi Wikipediaan RDF-kielen laajennoksilla, annotointien tulee silti olla jonkin ontologian mukaisia. Muuten annotointien hyödyntäminen on hankalaa eikä niistä ole merkittävää hyötyä.
3. Kategorioihin liittyvät uudet käytännöt. Suuri osa nykyisistä kategorioista tulisi korvatuiksi RDF-lauseilla. Toisinaan artikkeleille pitäisi lisätä nykyistä yleisempiä kategorioita niiden kategorioiden tilalle, jotka on korvattu RDF-lauseilla. Ontologian järkevyyden kannalta kategoriahierarkian tulisi kuvata vain asioiden kuulumista luokkaan ja luokkien alaluokka-yläluokka -suhdetta. Kategorioilla ei pitäisi ilmaista mitään muita asiayhteyksiä. Tämä vaatisi hierarkian laajaa uudelleen organisointia.
4. Käytännöt tiedon koostamisessa ja uudelleen käytössä. Tämä on yksi merkittävimmistä semanttisuuden tuomista eduista. Sen hyödyntämistä on syytä pohtia tarkemmin. Milloin artikkeleihin on sopivaa liittää RDF-kyselyn tuottama lista tai taulukko? Tarvitaanko erityisiä käytäntöjä lukittujen sivujen suhteen? Erilaisten listauksien tekeminen ja ylläpitäminen helpottuu huomattavasti. Halutaanko Wikipediaan huomattava määrä erilaisia listasivuja? Aiheuttaako lukuisat semanttiset haut sivuilla suorituskyöngelmia?

Tämän tutkielman perusteella RDF-lauseiden syöttäminen pääasiassa tai pelkästään tietolaatikoiden ja muiden sivupohjien kautta vaikuttaa järkevältä ratkaisulta. Tällöin kaikkien editorien ei välttämättä tarvitsisi lainkaan tuntea uutta RDF-lauseisiin liittyvää syntaksia. RDF-lauseiden käyttö tulisi automaattisesti yhdenmukaiseksi samaa sivupohjaa käyttävien artikkelien välillä. Ontologian määrittelemineen olisi myös luontevaa sivulla, joka määrittelee tietolaatikon tai muun sivupohjan. DBPedian luokisten sivupohjien jäsentämiseen liittyvien ongelmien perusteella voidaan kuitenkin

sanoa, että tällaisten semanttisten sivupohjien käytössä on oltava hieman tarkempi kuin Wikipedian nykyisten sivupohjien kanssa.

Wikipediassa on nykyisin joitakin sivupohjia, jotka annotoivat tietoa niin sanottujen mikroformaattien avulla. Sivupohjat lisäävät tuotettuun HTML-koodiin merkintöjä, jotka käyttäjän selain voi lukea, mikäli siihen on asennettu sopiva lisäosa. Eräs tällainen sivupohja englanninkielisessä Wikipediassa on persondata, jolla annotoidaan henkilötietoja. Tällainen sivupohjien käyttö muistuttaa hyvin paljon ehdotettua menetelmää, jossa sivupohjilla tuotetaan RDF-lauseet. Ainoana erona on tietomalli ja Semantic MediaWikin sisäinen tuki sivupohjan tuottamille annotaatioille.

Semantic MediaWikissä, ja kaikissa muissakin tutkielmassa mainituissa semanttisissa wikijärjestelmissä, jätetään hyödyntämättä monia annotointien mahdollistamia ominaisuuksia. Tällaisia ovat esimerkiksi kunnollinen tuki RDFS- ja OWL-päätelylle ja tiedon jakaminen eri kieliversioiden välillä. Näiden lisääminen Semantic MediaWikiin varmasti toisi joukon uusia ongelmia, joita ei tässä tutkielmassa ole käsitelty lainkaan. Esimerkiksi OWL-päätelyn seurauksena artikkeleihin saataisiin pääteltyä automaattisesti RDF-lauseita, joita ei suoraan ole mainittu missään. Tästä voi aiheutua ongelmia, sillä toisinaan voi olla hankala sanoa, miksi jokin RDF-lause esiintyy artikkelissa tai millä tavalla jokin pienehkö muutos vaikuttaa joihinkin muihin artikkeleihin. Wikipedian kieliversiot eivät myöskään tällä hetkellä ole keskenään yhdenmukaisia, joten RDF-lauseiden jako eri versioiden välillä ei onnistu aivan helposti. Useissa tapauksissa olisi varmasti toivottavaa, että eri kieliversio olisivat yhdenmukaisia, mutta onko näin välttämättä aina?

Semantic MediaWiki olisi otettavissa käyttöön Wikipediassa jo nykyisessä muodossaan. Huolimatta havaituista puutteista sillä olisi mahdollista lisätä huomattava määrä hyödyllisiä annotaatioita artikkeleihin. Tämä mahdollistaisi käyttäjien tekemät semanttiset haut ja automaattisesti päivittyvät listaukset. Wikipediassa tarvittavat uudet käytännöt ja muutokset kuitenkin aiheuttavat sen, että järjestelmän käyttöönotto ei ole aivan suoraviivaista. Koska kyseessä on melko suuri muutos Wikipedian kannalta, voi olla järkevää odottaa, että semanttisen wikin loputkin ongelmat saadaan ratkaistua ennen kuin Wikipediaa ruvetaan kunnolla annotoimaan.

Lähteet

- AuA05 Aumüller, D. ja Auer, S., Towards a Semantic Wiki Experience – Desktop Integration and Interactivity in WikSAR. *Proceedings of 1st Workshop on The Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure*, marraskuu 2005, URL http://www.semanticdesktop.org/SemanticDesktopWS2005/final/22_aumueller_semanticwikiexperience_final.pdf.
- Aue07 Auer, S., Bizer, C., Lehmann, J. et al., DBpedia: A Nucleus for a Web of Open Data. *Proceedings of the 2007 International Semantic Web Conference*, marraskuu 2007.
- ADR06 Auer, S., Dietzold, S. ja Riechert, T., OntoWiki – A Tool for Social, Semantic Collaboration. *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, marraskuu 2006, sivut 736–749.
- AuL07 Auer, S. ja Lehmann, J., What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. *Proceedings of European Semantic Web Conference (ESWC07)*, 2007, sivut 503–517.
- BFB05 Bryant, S., Forte, A. ja Bruckman, A., Becoming Wikipedian: Transformation of participation in a Collaborative Online Encyclopedia. *Proceedings of the International Conference on Supporting Group 2005*, marraskuu 2005.
- BuG06 Buffa, M. ja Gandon, F., SweetWiki: semantic web enabled technologies in Wiki. *Proceedings of the 2006 International Symposium on Wikis*, 2006, sivut 69–78.
- BHL01 Berners-Lee, T., Hendler, J. ja Lassila, O., The Semantic Web. *Scientific American*, 284,5(2001), sivut 34–43.
- WiS05 Emigh, W. ja Herring, S. C., Collaborative authoring on the web: A genre analysis of online encyclopedias. *HICSS 2005. 38th Annual Hawaii International Conference on System Sciences*, 2005, sivu 99b.
- Bri06 Encyclopaedia Britannica, Fatally flawed, maaliskuu 2006. http://corporate.britannica.com/britannica_nature_response.pdf

- Gil05 Giles, J., Internet encyclopaedias go head to head. *Nature*, 438, sivut 900–901.
- GaG06 Garshol, L. ja Moore, G., *Topic Maps – Data Model (ISO 13250)*, kesäkuu 2006. URL <http://www.isotopicmaps.org/sam/sam-model/>.
- Gre07 Greaves, M., Semantic web 2.0. *Intelligent Systems*, 22,2(2007), sivut 94–96.
- Joc06 Jochen, F. et al., Ideas and Improvements for Semantic Wikis. *The Semantic Web: Research and Applications. 3rd European Semantic Web Conference*, kesäkuu 2006, sivut 650–663.
- Kal04 Kalfoglou, Y., Alani, H., Schorlemmer, M. ja Walton, C., On the emergent semantic web and overlooked issues. *Proceedings of the Third International Semantic Web Conference*, marraskuu 2004, sivut 576–590.
- Käh07 Kähkönen, V., CIA ja Vatikaani jäivät kiinni Wikipedian käpälöinnistä. *Helsingin Sanomat*, 16.8.2007.
- KVV05 Krötzsch, M., Vrandečić, D. ja Völkel, M., Wikipedia and the Semantic Web – The missing links. *Proceedings of the 1st International Wikimedia Conference, Wikimania*, elokuu 2005.
- LaH07 Lassila, O. ja Hendler, J., Embracing "web 3.0". *IEEE Internet Computing*, 1,3(2007), sivut 90–93.
- Lih04 Lih, A., Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. *Proceedings of the Fifth international Symposium on Online Journalism 2004*, huhtikuu 2004.
- MHS06 Mäkelä, E., Hyvönen, E. ja Saarela, S., Ontogator — a semantic view-based search engine service for web applications. *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, marraskuu 2006.
- RuE07 Ruotsalo, T. ja Hyvönen, E., An event-based approach for semantic metadata interoperability. *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*. Springer-Verlag, marraskuu 2007.

- Sei05 Seigenthaler, J., A false Wikipedia 'biography'. *USA Today*, 29.11.2005.
- SKW07a Suchanek, F., Kasneci, G. ja Weikum, G., YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *Proceedings of the 16th International World Wide Web Conference*, toukokuu 2007, sivut 697–706.
- SKW07b Suchanek, F., Kasneci, G. ja Weikum, G., YAGO: A Large Ontology from Wikipedia and Wordnet. Tekninen raportti MPI-I-2007-5-003, Max-Planck-Institute for Computer Science, joulukuu 2007. URL <http://www.mpi-inf.mpg.de/~suchanek/publications/yagotr.pdf>.
- SSR06 Schraefel, M., Smith, D. A., Russel, A. ja Wilson, M., Semantic web meets web 2.0 (and vice versa): The value of the mundane for the semantic web. *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, 2006.
- TCC04 Tazzoli, R., Castagna, P. ja Campanini, S. E., Towards a Semantic Wiki Wiki Web. URL <http://platypuswiki.sourceforge.net/whatis/documents/platypuswiki.pdf>. 2006.
- ToS06 Tolksdorf, R. ja Simperl, E., Towards wikis as semantic hypermedia. *Proceedings of the 2006 International Symposium on Wikis*, 2006, sivut 79–88.
- Völ06 Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H. ja Studer, R., Semantic Wikipedia. *Proceedings of the 15th International Conference on World Wide Web*, 2006.
- Vie07 Viegas, F., Wattenberg, M., Kriss, J. ja van Ham, F., Talk before you type: Coordination in Wikipedia. *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, 2007, sivut 78–88.
- W3C04c World Wide Web Consortium, *RDF Vocabulary Description Language 1.0: RDF Schema*, helmikuu 2004. URL <http://www.w3.org/TR/rdf-schema/>.
- W3C04a World Wide Web Consortium, *Resource Description Framework (RDF)*, helmikuu 2004. URL <http://www.w3.org/RDF/>.

- W3C04b World Wide Web Consortium, *Web Ontology Language (OWL)*, helmikuu 2004. URL <http://www.w3.org/2004/OWL/>.
- W3C06 Defining N-ary Relations on the Semantic Web. Tekninen raportti, World Wide Web Consortium, huhtikuu 2006. URL <http://www.w3.org/TR/swbp-n-aryRelations/>.
- Wik07a Wikipedia. URL <http://en.wikipedia.org/wiki/Wikipedia>.
- Wik07b Size of wikipedia. URL http://en.wikipedia.org/wiki/Size_of_Wikipedia.
- Wik07c Wikipedia statistics. URL <http://stats.wikimedia.org/EN/Sitemap.htm>.
- W3C07 World Wide Web Consortium, *SPARQL Query Language for RDF*, marraskuu 2007. URL <http://www.w3.org/TR/rdf-sparql-query/>.