Semanttinen Finlex

Laki ja oikeus avoimena linkitettynä datana

Eero Hyvönen[1], Jouni Tuominen[1], Matias Frosterus[1], Eetu Mäkelä[1], Aki Hietanen[2]
Aalto-yliopisto, Semanttisen laskennan tutkimusryhmä (SeCo) (1) ja oikeusministeriö (2)
Projektin kotisivu: http://www.seco.tkk.fi/projects/lawlod/

1 MIKSI LAINSÄÄDÄNTÖ JA OIKEUSKÄYTÄNTÖ PITÄÄ JULKAISTA AVOIMENA DATANA?

Oikeusnormiemme keskeisimpiä sisältöjä, kuten lakeja, näitä tarkentavia asetuksia ja niiden soveltamiseen liittyvää oikeuskäytäntöä (oikeustapauksia) on voinut lukea verkossa oikeusministeriön tuottamassa Finlex-palvelussa[1]. Finlexissä julkaistaan lisäksi erilaisia viranomaismääräyksiä, valtiosopimuksia, hallituksen esityksiä ja säädösvalmisteluun liittyviä oppaita. Laki- ja oikeustieto on ollut tältä osin avointa ihmislukijoille verkkosivuina, mutta aineistot eivät ole olleet saatavilla avoimena *datana* siten, että tietotekniset sovellukset ja toiset verkkopalvelut voisivat sitä hyödyntää aineistoja lataamalla tai avoimien rajapintojen kautta.

Monilla tahoilla on kuitenkin selvä tarve lainsäädännöllisten sisältöjen saamiseksi käyttöön datana:

- **Tietoportaalit.** Monien eri alojen verkkopalveluissa on tarvetta viitata mm. lakien ja asetusten eri kohtiin ja näyttää niitä lukijalle. Tämä ei onnistu, jos lainsäädännön kohdat eivät ole viitattavissa ja luettavissa verkon kautta datana.
- **Lainopilliset verkkopalvelut.** Tällaisia palveluita ovat esimerkiksi Suomen Laki ja Edilex, jotka tarjoavat lakitietoa erityisesti oikeusalan ammattilaisille, kuten tuomareille, asianajajille ja yritysjuristeille, mutta myös yksityishenkilöille. Nykyiset järjestelmät perustuvat suuressa määrin käsityöhön, koska dataa ei ole saatavilla koneen ”ymmärtämässä” muodossa, vaan ainoastaan erimuotoisina PDF, Word ym. dokumentteina.
- **Lainvalmistelutyö.** Laadittaessa uusia säädöksiä, joilla muutetaan, täydennetään tai korvataan aiempia säädöksiä, joudutaan perehtymään aiemman lainsäädännön määräyksiin vaikutusten arvioimiseksi ja ristiriitaisuuksien välttämiseksi. Semanttista tietoa lainsäädännön eri versioista ja keskinäisistä riippuvuuksista ei kuitenkaan ole saatavilla kuin tekstimuodossa.
- **Lakiaineistojen toimittaminen ja julkaiseminen**. Lainsäädäntöön liittyvää tietoa tuotetaan nykyisin epäyhtenäisellä tavalla. Käytössä on erilaisia tekstiformaatteja ja eri tapoja ja asiasanastoja sisällön kuvailussa. Jos dokumentit laaditaan jo tuotantovaiheessa rakenteisena datana yhdessä sovittuja standardeja noudattaen, helpottuu niiden jatkokäsittely ja yhdistäminen toisiin dokumentteihin esimerkiksi eduskunnassa tai Finlexin kaltaisessa julkaisujärjestelmässä.
- **Media.** Mm. yritysmaailmaan ja politiikkaan liittyvissä uutisissa viitataan usein lain eri kohtiin, jolloin olisi tarpeen voida johdattaa lukija alkuperäisten lakitekstien äärelle. Tämä ei onnistu, jos lainsäädännön kohdat eivät ole viitattavissa ja käytettävissä datana.
- **Älykkäämmät palvelut.** Juridisiin ongelmatilanteisiin, kuten vaikkapa avioeroon tai perinnönjakoon liittyvä tieto on lainsäädännössä usein pirstaloitunut eri lakeihin, asetuksiin ja oikeuskäytännön esimerkkitapauksiin. Ei auta, vaikka säädöksiä ja oikeustapauksia olisi saatavilla, jos kokonaisuuden hahmottaminen ei lukijalle, kuten tavalliselle kansalaiselle, onnistu. Lainsäädännön dokumenttien esittäminen koneen ”ymmärtämänä” eli semanttisena datana mahdollistaa aiempaa älykkäämpien sovellusten kehittämisen kansalaisille ja muille tahoille. Lakitekstejä voidaan esimerkiksi linkittää niihin liittyviin toisiin teksteihin ja oikeustapauksiin tai lakia selittäviin sanastoihin automaattisesti.
- **Lainsäädännön ja -käytön tutkimus.** Lainsäädäntötyö ja säädösten soveltaminen ovat oikeustieteellisen tutkimustyön kohteena, ja tässä voidaan hyödyntää mm. data-analyysin menetelmiä. Tällainen työ kuitenkin

[1] http://www.finlex.fi/

edellyttää, että säädökset, niiden väliset yhteydet ja tieto soveltamisesta oikeustapauksissa on saatavilla systemaattisesti esitettynä datana.

Suorastaan yllättävää on, että mikään Suomen lainsäädäntöön liittyvä data ei ole ollut avoimesti saatavilla verkosta, vaikka laki on periaatteessa avointa, ja jokaisen kansalaisen oletetaan sitä tuntevan ja kuuliaisesti noudattavan. Toki lakiin on voinut tutustua verkkosivujen ja paperitulosteiden välityksellä.

Semanttinen Finlex on ensimmäinen Suomen lainsäädännön ja –käytön avoimen datan julkaisu. Se perustuu uuteen Linked Data –teknologiaan ja julkaisumalliin, joka mahdollistaa 1) erilaisten oikeustieteellisten aineistojen sisällöllisen rikastamisen toistensa avulla dataa linkittämällä sekä 2) datan kustannustehokkaan hyödyntämisen erilaisissa verkkopalveluissa ja sovelluksissa.

Seuraavassa kuvataan lyhyesti uusi Finlexille kehitetty semanttinen datajulkaisupalvelu sekä esimerkkinä palvelun hyötykäytöstä sen avulla kehitetty Semanttinen selain -sovellus.

2 SEMANTTINEN FINLEX DATAPALVELUNA

Tutkimus ja kehitystyö Finlexiin kuuluvien sisältöjen (Hietanen, 2010) julkaisemiseksi avoimena linkitettynä datana, ”Semanttisena Finlexinä”, käynnistyi v. 2012. Työ oli yksi case-tutkimus Aalto-yliopiston vetämässä Linked Data Finland -projektissa [2] (2012-2014) ja siinä olivat mukana Aalto-yliopiston ja Helsingin yliopiston ohella oikeusministeriö, Edita Publishing Oy ja Talentum Oyj Tekesin toimiessa päärahoittajana. Hanke jatkuu 2015-2016 valtiovarain- ja oikeusministeriön erillisellä rahoituksella liittyen valtionhallinnossa menossa olleeseen laajempaan Avoimen tiedon ohjelmaan[3] ja tuottavuuden kehittämistavoitteisiin.

Semanttinen Finlex -hankkeessa tutkitaan ja kehitetään Suomen lainsäädännön ja oikeuskäytännön avointa julkaisemista semanttisen webin Linked Data -teknologioiden ja julkaisuperiaatteiden (Heath, Bizer, 2011) avulla. Yhtenä esikuvana työlle oli Alankomaissa tehty Metalex Document Server (Hoekstra, 2011) lakidatan julkaisemiseksi linkitetyn datan palveluna. EU:n piirissä oli käynnistynyt standardointityötä tavoista, joilla voitaisiin viitata kansallisiin lakeihin ja oikeustapauksiin. Tämä on tarpeen, kun esimerkiksi jäsenvaltioiden lakitietoa kootaan EU-tasolla erilaisiin palveluihin.

Tutkimustyön tuloksena syntyi ensimmäinen RDF-perustainen avoimen datan julkaisu keskeisimmistä Finlexin sisällöistä (Frosterus et al, 2014), joka koostuu v. 2012 ajantasaisista laeista (2413 kpl) ja valikoimasta niihin liittyviä korkeimman oikeuden ratkaisuja (11 904 kpl vuosilta 1926–2012) ja korkeimman hallinto-oikeuden ratkaisuja (1490 kpl vuosilta 1944–2012). Semanttinen Finlex on käytettävissä Linked Data Finland -palvelussa (Hyvönen et al., 2014) sekä rajapintojen kautta että julkaisun eri datajoukot ja metadataskeemat lataamalla. Hankkeen yhteydessä kerättiin ja jalostettiin myös lainsäädännön kuvailussa eri organisaatioissa käytettyjä asiasanastoja RDF-muotoon (yhteensä n. 6000 termiä 26 eri sanastosta) ajatuksena niiden harmonisointi myöhemmin asteittain yhä yhteentoimivammaksi ontologiaksi. Lisäksi muodostettiin lainsäädäntöön ja –käyttöön liittyvistä uutisista oma datajulkaisu.

Semanttinen Finlex julkaistiin Linked Data Finland -alustalla (Hyvönen et al., 2014) avoimena datapalveluna (service) ”finlex” joka sijaitsee osoitteessa:

http://www.ldf.fi/dataset/finlex

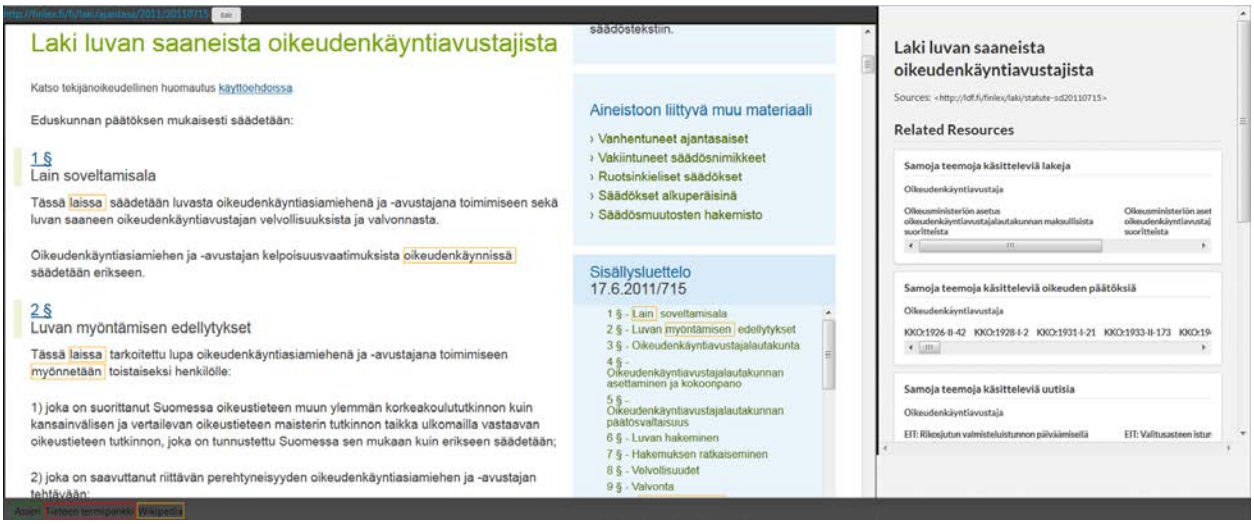
Sivulta löytyvät linkit datan ja sen skeemojen lataamiseen, datan automaattisesti tuotettuun dokumentaatioon, palvelun kuvaukseen, sanastoanalyysiin, Linked Data –selailuun, visualisointiin Googlen grafiikoilla, sekä sovellusten kannata keskeiseen W3C:n standardien ja käytänteiden mukaiseen SPARQL-rajapintaan:

http://ldf.fi/finlex/sparql

Rajapintaa käytetään ohjelmallisesti, mutta jos osoitteen kirjoittaa selaimeen, pääsee sitä käyttämään myös interaktiivisesti LDF.fi-palveluun asennetun SPARQL-editorin YASGUI [4] avulla. Palvelun oletuskysely YASGUI-

[2] http://www.seco.tkk.fi/projects/ldf/
[3] http://vm.fi/avointieto
[4] http://www.semantic-web-journal.net/content/yasgui-family-sparql-clients

editorissa hakee eri graafit, joita on tällä hetkellä yhdeksän, laskee niiden koot kolmikkoina sekä näyttää esimerkkinä kustakin graafista yhden kolmikon. Sen resurssilinkkien kautta voi tutustua dataan selailemalla. Esimerkiksi lait sisältävässä graafissa on n. kolme miljoonaa solmujen välistä yhteyttä (triple).



Kuva 1. Finlexin sivun sisällön rikastaminen Semanttisen selaimen avulla.

3 ESIMERKKI SOVELLUKSESTA: SEMANTTINEN FINLEX-SELAIN

Demonstraationa rajapinnan käytöstä on toteutettu ”Semanttinen Finlex-selain”, joka hyödyntää SeCO-ryhmässä kehitettyä Contextual Reader –ohjelmistoa (Mäkelä et al., 2015). Semanttinen selain linkittää verkossa olevien HTML-sivujen ja PDF-dokumenttien sanoja ja pidempiä ilmauksia reaaliajassa ja automaattisesti niitä taustoittaviin tietoihin. Systeemi perustuu palvelussa oleviin ontologioihin ja SeCo LAS -kieliteknologiseen palveluun[5] (Mäkelä, 2014), jossa sanamuodot ensin analysoidaan ja haetaan niitä vastaavat URI-tunnisteet ontologiapalvelusta. Näiden avulla voidaan sitten muodostaa SPARQL-kyselyt, joiden avulla on mahdollista hakea ilmauksiin liittyvää lisätietoa verkon kautta. Semanttisen Finlexin tapauksessa voidaan näin löytää esimerkiksi ilmauksiin liittyviä lakeja ja oikeustapauksia, hakea sisältöihin liittyviä uutisia tai löytää määrittelyjä Finlexin teksteissä oleville oikeustieteellisille termeille.

Kuvassa 1 on esimerkki tästä käyttötapauksesta, joka löytyy osoitteesta:

http://tinyurl.com/pbaxmh4

Siinä verkossa olevaan Finlexin lakisivuun ”Laki luvan saaneista oikeudenkäyntiavustajista”

http://www.finlex.fi/fi/laki/alkup/2011/20110715

on liitetty termien määritelmiä kolmesta eri datalähteestä: 1) Finlexin Asseri-sanastosta, 2) Helsingin yliopiston Tieteen termipankin oikeustieteellisestä tietokannasta ja 3) Suomenkielisestä Wikipediasta, joista kaikista on luotu Linked Data –perustainen palvelu verkkoon. Linkitetty termi näkyy käyttäjälle värikoodattuna laatikkona tekstissä. Viemällä kursori laatikon päälle, voi lukea siihen liittyvää lisätietoa infoboxin kautta. Laatikon väri ilmaisee linkin tietolähteen. Tässä tapauksessa kaksi ensimmäistä tietolähdettä ovat mukana Semanttisessa Finlexissä ja DBpedia on julkaistu erikseen LDF.fi-palvelussa. Järjestelmässä voidaan käyttää mitä tahansa muutakin verkon SPARQL-palvelua.

Lisäksi kuvan 1 alkuperäisen Finlex-sivun oikealle puolelle on automaattisesti luotu joukko semanttisia suosittelulinkkejä (content-based recommendation) lakiin liittyviin muihin Semanttisen Finlexin aineistoihin. Tässä

[5] http://demo.seco.tkk.fi/las/

tapauksessa linkit johtavat 1) samoja teemoja käsitteleviin toisiin lakeihin, 2) samoja teemoja käsitteleviin oikeustapauksiin ja 3) samoja teemoja käsitteleviin uutisiin.

Demonstraattorin ideana on osoittaa, miten Semanttisen Finlexin avulla voidaan helpottaa lukijalle lakiteksteihin tutustumista tarjoamalla hänelle automaattisesti lakeihin liittyvää hyödyllistä lisätietoa, niiden lukukonteksti.

4 SEMANTTINEN FINLEX 2

Edellä esitellyn työn jatko-osa 2015-2016 on käynnissä Aalto-yliopistossa oikeusministeriön johdolla yhteistyössä Finlexin teknisen ratkaisun toteuttaneen Edita Publishing Oy:n kanssa. Tavoitteena on alkuperäisen Semanttisen Finlexin (SF) prototyypin ja sovellusten kehittäminen eteenpäin seuraavilla tavoilla:

- **IRI-tunnistejärjestelmät.** SF:n eri sisältökohteiden (lait, asetukset, pykälät, oikeustapaukset, organisaatiot jne) IRI-tunnistejärjestelmän (W3C:n standardoima International Resource Indentifier) muuttaminen uusien standardien ECLI[6] (European Case Law Identifier) ja ELI[7] (European Law Identifier) mukaiseksi. IRI-tunnisteiden avulla voidaan 1) indeksoida ja sisällönkuvailla tietosisältöjä yksikäsitteisellä tavalla, 2) muodostaa rajapintakyselyjä koneluettavan datan hakemiseksi ja lukemiseksi SF:sta eri muodoissa (kuten RDF, JSON ja CSV) ja 3) viitata Finlexin vastaaviin ihmisluettaviin sivuihin eri sovelluksissa.
- **Metadatamallit.** Nykyisen SF:n metadatamallit kehitetään helppokäytteisemmäksi niin, että lakitekstejä ei tarvitse koostaa niiden osista, kuten alkuperäisessä SF:ssa, jonka tietomalli myötäili vielä paljolti alkuperäistä, dokumenttien rakennetta kuvaavaa XML-skeemaa, eikä niinkään niiden sisältöä dokumentteina. Uusi tietomalli laaditaan ECLI-standardin suosittaman metadatamallin mukaiseksi ja perustuu paljolti webissä laajassa käytössä olevaan Dublin Coreen.
- **Julkaisun automaattinen päivitys.** Hankkeessa toteutetaan SF:n julkaisun automaattinen päivitys Finlex-julkaisun muutoksien mukaan, jolloin käytössä olisi aina ajantasainen Finlexiä vastaava datajulkaisu. Tällä hetkellä verkossa oleva versio ei ole ajantasainen, vaan perustuu Linked Data Finland –projektin käyttöön v. 2012 luovutettuun versioon.
- **Sanastot ja ontologiat.** Tavoitteisiin kuuluu myös sisällönkuvailussa käytettyjen asiasanastojen ja ontologioiden kehittäminen, mukaan lukien oikeusministeriössä valmistunut Asseri-sanasto. Uutuutena on myös SF:n yhdistäminen Tieteen termipankin oikeustieteellisen tietokannan [8] määritelmien (tällä hetkellä n. 1650 määritelmää/sivua).
- **Uudet datajulkaisut.** Semanttiseen Finlexiin otetaan mukaan uusia tietosisältöjä Finlexistä.
- **Datan validointi.** Järjestelmään kehitetään ja lisätään automattinen datan laadun tarkistin.
- **Sovellusdemonstraatiot.** Semanttisen Finlexin hyödyllisyyttä testataan toteuttamalla semanttinen haku- ja suosittelukone, jonka avulla lakisisältöjä voidaan hakea paitsi perinteisellä tekstihaulla myös tekstihakua täydentävillä ”älykkäämmillä” tavoille kuten ns. fasettihaulla. Hakutuloksia voidaan myös linkittää automaattisesti toisiinsa semanttisten suosituslinkkien avulla. Semanttinen haku ja suosittele perustuvat kieliriippumattomiin käsitteisiin ja mahdollistavat näin myös hakujen suorittamista ja suosittelua yli kielirajojen. Hankkeessa kehitetään myös semanttisen selaimen prototyyppiä ja siinä tarvittavaa kieliteknologiaa eteenpäin kohti tuotantokäyttöä.

5 SEMANTTISEN FINLEXIN INNOVATIIVISUUS

Lainsäädäntöön liittyvän tiedon rakenteistamista on aiemmin tutkittu Suomessa SGML- ja XML-perustaisen teknologian pohjalta. Myös nykyinen Finlex on XML-perustainen. Finlexin data ei kuitenkaan ole ollut avoimesti saatavilla, ja julkaisussa käytetty formaatti on suunniteltu lähinnä järjestelmän toimittajan Editan Publishing Oy:n kehittämää julkaisujärjestelmää varten. W3C:n standardoima XML-soveltuu hyvin dokumenttien rakenteen esittämiseen, mutta metatietojen esittämiseen W3C on sittemmin kehittänyt semanttisen webin alueelle kuuluvan RDF-tietomallin (Resource Description Framework), SKOS- ja OWL-kielet sisällönkuvailuissa tarvittavien sanastojen ja käsitteiden esittämiseksi, SPARQL-kielen aineistojen kyselyjä varten ja erityisen Linked Data -metodiikan

[6] https://en.wikipedia.org/wiki/European_Case_Law_Identifier
[7] https://en.wikipedia.org/wiki/European_Legislation_Identifier
[8] http://tieteentermipankki.fi/wiki/Oikeustiede

datajulkaisujen toteuttamiseksi verkossa[9]. Jo vuonna 2006 nähtiinkin RASK2-hankkeessa semanttinen web ja RDF lupaavaksi uudeksi teknologiaksi lainsäädäntötyön tiedonhallinnassa (Nurmeksela et al., 2006) ja eduskunnan kirjasto oli mukana kansallisessa FinnONTO-hankkeiden sarjassa (2003-2012), jossa kehitettiin suomalaisen semanttisen webin infrastruktuuria ja sovelluksia (Hyvönen, 2014).

Semanttinen Finlex on teknisesti edistyksellinen kansainvälisiin verrokkeihin nähden (Frosterus et al., 2015) ja sovittaa oikeusinformatiikassa vasta kehitteillä olevia uusia ratkaisumalleja Suomen lakeihin ja oikeuskäytäntöön. Poikkeuksellista on, että työssä on vahvasti mukana datajulkaisun ohella käytännön sovellusten kehittämisen näkökulma.

Kiitokset

Kiitos Mika Walhroosille datamuunnoksista projektin alkuvaiheessa. Jari Linhala ja Risto Talo ovat vastanneet Finlexin teknisestä toteutuksesta ja datan tuotannosta projektin käyttöön Edita Publishing Oy:ssä. Talentum Oyj:n Stina Wikberg oli mukana hankkeen ensimmäisen vaiheen johtoryhmässä. Hanketta ovat rahoittaneet Tekes, oikeusministeriö, liikenne- ja viestintäministeriö, valtiovarainministeriö, Edita Publishing Oy, Talentum Oyj. Projektin jatko-osassa Semanttinen Finlex 2 tutkimusryhmään ovat liittyneet Arttu Oksanen ja Minna Tamper.

KIRJALLISUUTTA

Matias Frosterus, Jouni Tuominen, Eero Hyvönen: Facilitating Re-use of Legal Data in Applications--Finnish Law as a Linked Open Data Service. Proceedings of the 27th International Conference on Legal Knowledge and Information Systems (JURIX 2014), IOS Press, Krakow, Poland, December, 2014. http://www.seco.tkk.fi/publications/2014/frosterus-et-al-finnish-law-lod-2014.pdf

Eero Hyvönen: FinnONTO-hanke loi ontologisen perustan kansalliselle webin tietoinfrastruktuurille. Tieteessä tapahtuu, no. 3, 2014. http://ojs.tsv.fi/index.php/tt/article/view/41559

Eero Hyvönen, Jouni Tuominen, J., Miika Alonen, Eetu Mäkelä: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: Proceedings of ESWC 2014 Demo and Poster Papers, Springer-Verlag (May 2014) http://www.seco.tkk.fi/publications/2014/hyvonen-et-al-ldf-2014.pdf

Aki Hietanen, Sähköinen, autenttinen ja ajantasainen – sähköisen säädösjulkaisemisen kehitysvaiheista. Teoksessa Heikki E. S. Mattila; Sari Pajula; Aino Piehl (toim.), Oikeuskieli ja säädöstieto, s. 277-287. Suomalaisen Lakimiesyhdistyksen julkaisuja C 41, Helsinki 2010.

Rinke Hoekstra: The MetaLex Document Server legal documents as versioned linked data. In: Proceedings of the ISWC 2011, Bonn, Germany. pp. 128–143. Springer-Verlag (2011)

Eetu Mäkelä: Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. Proceedings of the ESWC 2014 demonstration track, Springer-Verlag, 2014.

Eetu Mäkelä, Thea Lindquist, Eero Hyvönen: CORE - A Contextual Reader based on Linked Data. Aalto University, Semantic Computing Research Group, Submitted Papers, 2015. http://www.seco.tkk.fi/publications/

Reija Nurmeksela, Maiju Virtanen, Antti Lehtinen, Matti Järvenpää, Airi Salminen: Suomalaisen lainsäädäntötyön tiedonhallinta. Suuntana semanttinen web. Eduskunnan kanslian julkaisu 2/2006. http://docplayer.fi/750837-Suomalaisen-lainsaadantotyon-tiedonhallinta-suuntana-semanttinen-web.html

[9] http://www.w3.org/standards/semanticweb/