



SOTASAMPO

TALVI- JA JATKOSOTA SEMANTTISESSA WEBISSÄ

Työryhmä: Eero Hyvönen (vetäjä), Jouni Tuominen, Eetu Mäkelä, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper

Organisaatiot: Aalto-yliopisto, Semanttisen laskennan tutkimusryhmä (SeCo) ja lukuisat yhteistyökumppanit

Sotasampo julkaisee linkitetynä avoimena datana talvi- ja jatkosotaan liittyviä keskeisiä kansallisia tietoaaineistoja yhteistyössä Kansallisarkiston, puolustusvoimien, Maanmittauslaitoksen, Suomen Sotahistoriallisen Seuran, Yleisradion ja muun yhteistyöverkoston kanssa. Tavoitteena on Suomen historian suurimman tragedian ymmärtämisen edistäminen ja Digital Humanities -tutkimuksen tukeminen uusimman semanttisen webin ja kieliteknologian avulla. Datapalvelun hyödyntämistä pilotoidaan uudessa Sotasampo.fi-verkkopalvelussa, joka koostuu erilaisista näkymäsovelluksista aineistoihin, kuten sotien tapahtumat aikajanalla ja kartalla, henkilöt, joukko-osastot, paikat, veteraanien muistelmät ja sotasurmat. Hankkeen tavoitteena on tarjota semanttisesti rikas kokonaisuus sekä avoimena linkitetynä datana sovelluksille että verkkosivustona ihmisille. Tieteellisenä innovaationa on rikastaa eri datasiilojen aineistoja toisensa avulla linkittämällä dataa yhteisen tapahtumaperustaisen ontologiainfrastruktuurin kautta.

1 SOTASAMMON TAVOITTEET

G. W. F. Hegelin mukaan "historiasta opimme sen, että emme opi historiasta mitään". Toivottavasti näin ei tapahdu Suomessa toisen maailmansodan osalta, johon liittyvät asiat ovat taas olleet esillä mm. Ukrainan kriisin yhteydessä. Itsenäisyytemme ajan sotahistoria, erityisesti talvi- ja jatkosota, ovat 100-vuotisjuhlinsa valmistautuvan Suomen kansallisen identiteetin ja maanpuolustustahdon keskeinen tukipilari ja laajaa kiinnostusta herättävä historian tutkimuskohde niin tutkijoille kuin muillekin kansalaisille. Useimpien suomalaisten vanhemmat tai isovanhemmat osallistuivat tavalla tai toisella sodan eri tapahtumiin, pahimmassa tapauksessa surmansa saaden. Sotasampo-hankkeen tavoitteena on rakentaa uusimman Linked Open Data -teknologian (Heath, Bizer, 2011) avulla suomalaista identiteettiä lisäämällä ymmärrystämme viime sodistamme ja edistää näin rauhaa. Kehitettäviä uusia teknisiä ratkaisuja voidaan soveltaa muihinkin Digital Humanities -alueen aineistoihin.

Vaikka verkossa on runsaasti tietoa sodistamme ja sotahistoriallista kirjallisuutta paljon saatavilla, ei tietoa ole ollut saatavilla koneluettavassa muodossa avoimena datana. Tämä on estänyt tiedonsaantia, datan analysointia ja käyttöä tutkimuksessa sekä sotahistoriallisten sovellusten kehittämistä.

Sotasampo-hankkeen tavoitteena on luoda:

1. **Sotahistoriallinen linkitetyn avoimen datan palvelu.** Hankkeessa avataan Suomen toiseen maailmansotaan liittyvää tietoa datapalveluna sitä tarvitseville tutkijoille, historian harrastajille, sovellusten kehittäjille ja toisille verkkopalveluille (Hyvönen et al., 2015ab).
2. **Semanttinen portaali Sotasampo.fi.** Hankkeessa testataan ja osoitetaan datapalvelun hyödyllisyys ja haasteet pilotoimalla sitä semanttisessa portaalissa Sotasampo.fi sekä yhteistyökumppanien kautta heidän omissa sovelluksissaan.

Hanke on syksyllä 2015 julkaissut ensimmäisen demonstraattorin Sotasampo.fi-portaalista ja siinä olevat datajoukot avoimena linkitetynä datana Linked Data Finland -alustalla¹. Sotasampo on jatkoa Semanttisen laskennan tutkimusryhmän (SeCo) aiemmille semanttisen webin "sammoille" Kulttuurisampo.fi, Matkailusampo.fi ja Kirjasampo.fi ja hyödyntää kansallisissa FinnONTO- ja Linked Data Finland -hankkeissa kehitettyjä avoimia ohjelmistoja tietoinfrastruktuuria (Hyvönen, 2014). Työ on samalla jatkoa SeCo-ryhmän ja Coloradon yliopiston tutkimukselle ensimmäisen maailmansodan tapahtumien ja aineistojen mallintamisesta ja hyödyntämisestä avoimena linkitetynä datana (Hyvönen et al., 2012; Lindquist et al., 2013; Mäkelä et al., 2014).

¹ <http://www.ldf.fi>

2 SOTASAMMON AINEISTOT JA JULKAISIJAYHTEISÖ

Maassamme on viime vuosina julkaistu merkittäviä, laajan suosion saavuttaneita verkkopalveluita toiseen maailmansotaan liittyen, ja arkistojen kätköistä löytyy runsaasti lisää näihin liittyvää tietoa:

- **Sota-ajan valokuvat.** *Sotamuseon* SA-kuva-arkistopalvelu sisältää 160 000 Suomen talvi-, jatko- ja Lapin sodan ajan valokuvaa vuosilta 1939–1945 metatiedoilla varustettuna. Kuvat ovat katseltavissa verkossa. Sotamuseolla on myös arkistoissaan mm. n. 1300 piirrosta kattava jatkosodan ajan TK-piirroskokoelma, sotasaalislippukokoelma, ilmavoimien 60 000 kuvan 1918–1960 kokoelma ja yksityisten henkilöiden kuvakokoelmia. SA-kuva-arkiston verkkojulkaisulla oli ensimmäisen vuoden aikana n. miljoona vierailijaa.
- **Sotasurmatietokannat.** *Kansallisarkisto* on julkaissut verkkopalveluna Suomen sotaurmat 1914–1922², Suomen sodissa 1939–1945 menehtyneet³ ja Suomi, sotavangit ja ihmisluovutukset 1939–1955⁴. Suomen sotien 1939–1945 sotaurma-aineistot ovat olleet kovassa käytössä verkossa: esimerkiksi vuonna 2013 aineistoihin tutustui n. 83 000 kävijää. Uusia kiinnostavia aineistoja on ilmestymässä, kuten 30 000 digitoitua sotapäiväkirjaa, joiden avulla voi mm. tutustua yksittäisen sotilaan sotahistoriaan, kun hänen joukko-osastonsa tunnetaan.
- **Kansa taisteli – miehet kertovat -lehdet (1957–1986).** Lehden artikkelit, yhteensä n. 3360 kpl, ovat paljolti rintamalla palvelleiden veteraanien itsensä kirjoittamia muistoja sota-ajan tapahtumista. Ne on kesällä 2014 julkaistu verkossa Suomen Sotahistoriallisen Seuran ja Bonnierin toimesta.
- **Sota-ajan filmi- ja ääniaineistot.** *Yleisradiolla* ja *KAVA*:lla on runsaasti arkistoissa sota-aikaan liittyvää kuva-, radio-, filmi- ym. materiaalia, josta osa löytyy jo YLE:n Elävästä arkistosta.
- **Sotahistorialliset kokoelmat.** Sotamuseossa ja eräissä muissakin maamme museoissa on runsaasti sota-aikaan liittyvää esineistöä ja dokumentteja, joita on osin julkaistu mm. Kulttuurisampo- (kulttuurisampo.fi) ja Finna-portaaleissa (finna.fi).
- **Kartat ja paikkatieto.** *Maanmittauslaitos* on avannut paikkatietodataansa. Saatavilla on digitoituna avoimena datana mm. luovutetun alueen vanhat Karjalan kartat. Jyrki Tiittasen talkootyönä on syntynyt n. 35 000 paikkanimen geokoodattu paikkarekisteri luovutetun Karjalan alueilta.
- **Sota-ajan muistot.** *Suomalaisen Kirjallisuuden Seuran* arkistoissa on sota-ajalla kerättyä tietoa ja muistoja sota-ajan elämästä, samoin *Svenska Litteratur Sällskapetilla* on aineistoja mm. Ruotsiin lähetettyihin sotalapsiin liittyen.

Sotasampo-hanke on koonnut edellä mainitut datan julkaisijat mukaan yhteisiin talkoisiin ajatuksena luoda eri organisaatioiden erillisten datasiilojen aineistoista suuri, avoin linkitetyn datan tietämysverkko (Google käyttää tästä nimitystä ”knowledge graph”). Tutkimus- ja kehitystyötä tehdään osana Aalto-yliopiston Semanttisen laskennan tutkimusryhmän Linked Open Data Science -hanketta, joka kuuluu opetus- ja kulttuuriministeriön rahoittamaan Avoin tiede ja tutkimus -ohjelmaan⁵. Sotasampo on myös yksi Suomen itsenäisyyden satavuotisjuhlavuoden ohjelmahankkeista⁶.

Sotasampo.fi:n ensimmäiseen demonstraattoriin on muunnettu linkitetyksi dataksi taulukossa 1 luetellut datajoukot.

² <http://vesta.narc.fi/cgi-bin/db2www/sotasurmaetusivu/main?lang=fi>

³ <http://kronos.narc.fi/menehtyneet/index.php>

⁴ <http://kronos.narc.fi/>

⁵ <http://avointiede.fi/>

⁶ <http://suomi100finland.fi/etusivu>

Datan luovuttaja	Aineisto (25.11.2015)	Määrä	Datan lähteet ja tekijänoikeus	Lisenssi linkitettyyn dataan
Puolustusvoimien SA-	TK-kuvat ja videot	160 000	Puolustusvoimat & Aalto/SeCo	CC BY 4.0
Kansallisarkisto	Menehtyneet 39-45	95 000	Kansallisarkisto & Aalto/SeCo	CC BY 4.0
Kansallisarkisto	Organisaatiokortisto	500	Kansallisarkisto & Aalto/SeCo	CC BY 4.0
Kansallisarkisto	Sotapäiväkirjat	26 500	Kansallisarkisto & Aalto/SeCo	CC BY 4.0
Kansallisarkisto	Senaatin kartat	404	Kansallisarkisto & Aalto/SeCo	CC BY 4.0
Suomen Sotahistoriallinen Seura Bonnier Publications Oy	Kansa taisteli -lehtien artikkelit 1957-1986	3 360	Metadata: Timo Hakala & Aalto/SeCo. Artikkelit © Suomen Sotahistoriallinen Seura ja Bonnier, http://sshs.fi .	CC BY 4.0 (metadata)
Maanmittauslaitos	Paikannimirekisteri	800 000	Maanmittauslaitos (MML) & Aalto/SeCo	CC BY 4.0
Maanmittauslaitos	Karjalan topograf. kartat	47	MML & Aalto/SeCo	Ei-kaupallisiin tarkoituksiin
Jyrki Tiittanen	Karjalan paikat geokood.	35 000	Jyrki Tiittanen & Aalto/SeCo	CC BY 4.0
Aalto-yliopisto	Sotien tapahtumat	11 275	Aalto/SeCo. Lähteet: Talvi- ja jatkosodan pikkujättiläiset, organisaatiokortisto ym.	CC BY 4.0
Aalto-yliopisto	Henkilöt (Menehtyneet 39- 45 datan lisäksi)	1 050	Aalto/SeCo. Lähteet: Talvisodan historia, sotapäiväkirjat, organisaatiokortisto, Suomen Rintamamiehet 1939-1945, Maanpuolustuskorkeakoulu/Ohto Manninen, Wikipedia ym.	CC BY 4.0
	Biografiat	6 300	Suomalaisen Kirjallisuuden Seura / Kansallisbiografia & Aalto/Seco	CC BY 4.0
	Joukko-osastot	3 180	Aalto/SeCo. Talvisodan historia, Talvisodan sotapäiväkirjat, organisaatiokortisto	CC BY 4.0
YHTEENSÄ		1 142 616		

Taulukko 1. Sotasammon aineistot, niiden lähteet ja datan määrä

Seuraavassa esitellään lyhyesti ensin Sotasampo.fi-portaalisovellus ja sitten sen perustana oleva datapalvelu.

3 SOTASAMPO.FI-SOVELLUS

Sotasampo-portaali koostuu joukosta toisiinsa linkitettyjä näkymäsovelluksia (perspective), jotka tarjoavat erilaisia näkökulmia sodan aineistoihin. Kuvassa 3.1 näkyvät palvelun etusivulla olevat linkit eri näkymiin, jotka ovat Tapahtumat, Henkilöt, Joukko-osastot, Paikat, Kansa taisteli ja Menehtyneet.

Keskeinen tieteellistekninen perusta Sotasammossa on ns. tapahtumaperustainen mallinnus (event-based modeling): data on harmonisoitu kansainvälisen CIDOC CRM -standardin mukaisen tapahtumaontologian tapahtumiksi, joita karakterisoi tapahtuman paikka, aika, siihen liittyvät toimijat ja muu tieto. Tarkastellaan esimerkkinä näkymästä Tapahtumat-näkymäsovellusta, jonka pääsivu on kuvassa 3.2. Näkymä esittää talvi- ja jatkosodan poliittisia ja sotilaallisia tapahtumia aikajanalla ja viiden päivän aikaikkunaan sisältyvät tapahtumat kartalla. Taustalla näkyvä lämpökartta lasketaan valittuna aikaikkunan aikana menehtyneistä sotilasta sotasurmadatan perusteella. Ikkunaa ajassa siirtämällä muuttuvat tapahtumat ja sotasurmien määrää ilmaisevan lämpökartan väritys (punainen-keltainen-vihreä), mikä antaa yleiskuvan tapahtumien kehittymisestä ajan ja paikan suhteen.

Käyttäjä voi klikkaamalla valita tapahtuman aikajanalta tai kartalta. Kuvassa on valittu tapahtuma ”taistelut Suomussalmen kirkonkylän seudun omistuksesta ...”, jolloin siihen liittyvä data ja semanttiset linkit muihin aineistoihin näytetään oikealla. Tapahtuman kuvauksessa on seuraavat toiminnot: 1) ”Kylänmäki” ja ”Suomussalmi” ovat linkkejä Sotasammon Paikat-näkymän paikkojen tietosivuille, joilla voi katsella mm. paikkoja historiallisilla kartoilla sekä paikkoihin eri tavoin linkittyvää tietoa. Sekä historialliset kartat että muu tieto tulevat jälleen toisista datajoukoista. 2) Linkki ”Näytä tapahtuman tietosivu” johtaa tapahtuman omalle ”kotisivulle”, jossa esitellään eri tavoin tapahtumiin loogisesti liittyviä muita tietoja, esimerkiksi kootusti eri aikoina samalla paikalla käytyjä taisteluita ja muita tapahtumia tai henkilöitä, jotka osallistuivat ko. tapahtumaan. Tällaisia semanttisia suosittelevia voidaan luoda joustavasti SPARQL-rajapinnan kyselyiden avulla. 3) Lisäksi tapahtumatietoon on valmiiksi linkitetty puolustusvoimien SA-kuva-arkiston kuvia, jotka on otettu samassa paikassa ja samaan aikaan kuin valitun tapahtuman tiedetään tapahtuneen.

Ideana on siis rikastaa eri datajoukkojen sisältöjä toistensa avulla, jolloin kaikkien julkaisijoiden dataa voidaan rikastaa yhteistyöllä ”ilmaiseksi” ja yhteisöllisen datajulkaisun luomiselle syntyy hyvä motivaatio. Esimerkiksi Henkilöt-näkymän kautta voidaan sotilaan joukko-osastotietoa yhdistää organisaatiokorteista saatavaan tietoon joukko-osaston tapahtumista sekä joukko-osaston sotapäiväkirjoihin, jolloin voidaan rakentaa ja visualisoida kartalla paitsi joukko-osastojen niin myös yksittäisten sotilaiden liikkeitä ja sotahistoriaa.

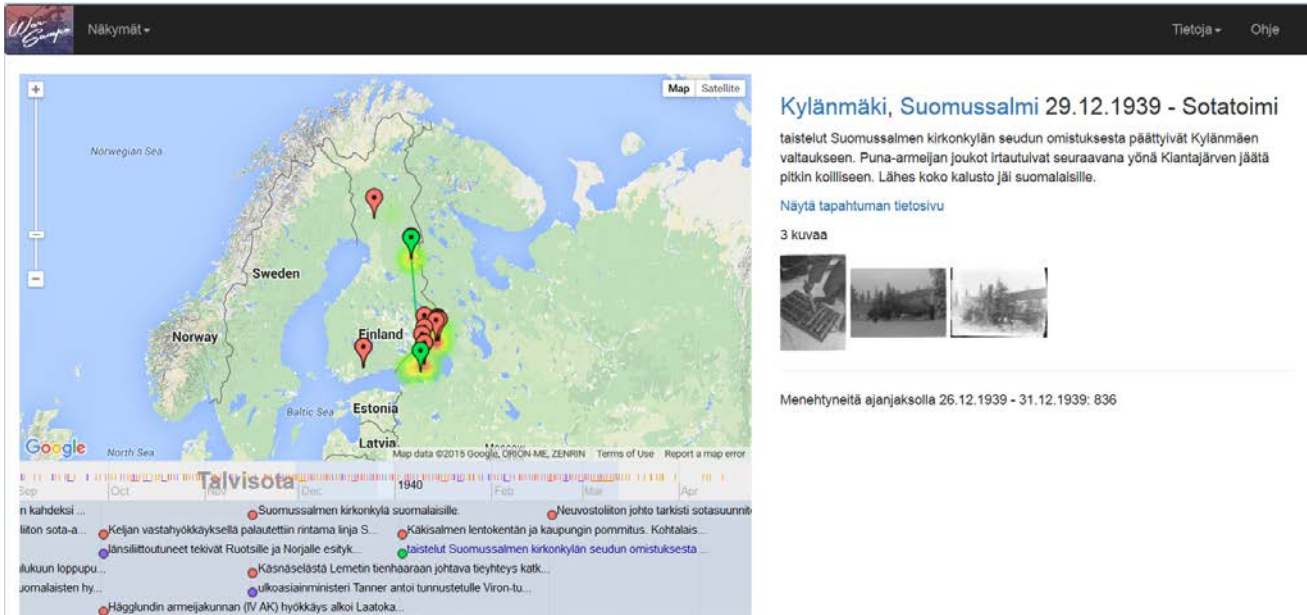
Eri sovellusten aineistojen tietosivuihin voidaan viitata systemaattisesti yhteisesti sovittujen URI-tunnisteiden avulla muodossa <http://www.sotasampo.fi/SOVELLUS/page?uri=URI>. Esimerkiksi Suomussalmen tietosivu Paikat-sovelluksen tuottamana palveluna on osoitteessa:

http://www.sotasampo.fi/places/page?uri=http://ldf.fi/warsa/places/municipalities/m_place_87

Tämä mekanismi mahdollistaa tietosivujen uudelleenkäytön paitsi Sotasampo.fi:ssä niin myös missä tahansa toisissa sovelluksissa HTML-muodossa.

Tapahtumat	Henkilöt	Joukko-osastot	Paikat	Kansa taisteli	Menehtyneet
Talvi- ja jatkosodan tapahtumat kartalla ja aikajanalla muihin aineistoihin linkitettyinä	Aineistoissa esiintyviä henkilöitä ja heihin eri tavoin liittyviä tietoja	Joukko-osastoihin liittyviä tapahtumia ym. tietoa mm. kartoilla visualisoituna	Hae historiallisia paikkoja kartoilla ja niihin liittyviä tapahtumia ja kohteita	Kansa taisteli -lehtien artikkelien haku ja kontekstuaalinen linkitys täydentäviin aineistoihin	Sotasurmatietokannan sankarivainajien etsiminen ja datan tutkiminen fasettihaun avulla

Kuva 3.1. Sotasammon pääsivun näkymäsovellukset ja niiden toiminta



Kuva 3.2. Sotasammon Tapahtumat-sovellusnäkyä

4 SOTASAMMON KÄYTTÖ AVOIMEN DATAN PALVELUNA

Kaikki Sotasammon datajoukot taulukossa 1 ovat käytettävissä avoimella Linked Data Finland -palvelualustalla (LDF.fi), jonka SPARQL-rajapinnan käyttöön myös kaikki Sotasampo.fi-portaalin eri näkymäsovellukset suoraan perustuvat. Datajoukkoja julkaistaan LDF.fi-palvelussa avoimen datan lisensseillä sitä mukaa kuin niiden toimivuus testataan ja data dokumentoidaan.

Sotasammon datajoukkojen julkaisu alkoi Kansallisarkiston kolmen sotasurmatietokannan avaamisella, joista Sotasampo käyttää talvi- ja jatkosodassa menehtyneiden sotilaiden avointa datapalvelua. Se on dokumentoitu LDF.fi:ssä osoitteessa:

<http://www.ldf.fi/dataset/narc-menehtyneet1939-45>

Datajoukkojen SPARQL-rajapintapalvelu löytyy LDF.fi-palvelussa aina osoitteessa, jonka muoto on <http://ldf.fi/SERVICE/sparql>. SERVICE on datajoukon nimi. Kirjoittamalla rajapinnan osoite selaimen, kuten

<http://ldf.fi/narc-menehtyneet1939-45/sparql>

pääsee käyttämään rajapintaa interaktiivisesti LDF.fi-palveluun asennetun SPARQL-editori YASGUI⁷ avulla.

Sotasammon datajulkaisut on koottu omaan uuteen palveluun nimeltä "warsa" (WarSampo), jossa eri datajoukot ovat mukana omina RDF-graafeinaan:

<http://ldf.fi/warsa/sparql>

⁷ <http://www.semantic-web-journal.net/content/yasgui-family-sparql-clients>

Palvelun oletuskysely YASGUI-editorissa hakee eri graafit, laskee niiden koot kolmikkoina sekä näyttää esimerkkinä niistä yhden kolmikon. Sen resurssilinkkien kautta voi tutustua dataan selailemalla.

5 SOTASAMMON UUTUUSARVO

Vaikka toiseen maailmansotaan liittyviä aineistoja on julkaistu verkossa runsaasti, on Sotasampo tiettävästi ensimmäinen datapalvelu maailmassa, jossa aineistojen data on saatavilla myös linkitettyinä avoimena datana. Sotasampo.fi-sovelluksen innovatiivisuus liittyy sen arkkitehtuuriin, jossa eri datasiilojen tietoa tarjotaan käyttäjille erilaisten toisiaan semanttisesti rikastavien näkymien kautta. Sisällöt on harmonisoitu yhteentoimiviksi yhteisten tapahtuma- ja muiden ontologioiden avulla. Linkitetty data perustuu joukkoon W3C:n standardoimia semanttisia teknologioita ja käytäntöjä (best practices)⁸, joiden avulla voidaan julkaista laajoja tietosisältöjä siten, että ne yhdistyvät toisiinsa käsitteellisellä tasolla. Näin muodostuva webin *datapilvi*, englanniksi "Web of Data" tai "Giant Global Graph" (GGG), kuten Wikipediasta louhitun DBpedia⁹ ympärille syntynyt Linked Open Data -pilvi (LOD), on kustannustehokkaasti hyödynnettävissä sovelluksissa sekä hyvin määritellyn standardoidun RDF-perustaisen (Resource Description Framework) rakenteensa että rajapintojensa (SPARQL, REST ym.) kautta¹⁰. Linked Data -lähestymistapa on otettu käyttöön mm. Googlen "Knowledge Graph"- sekä Microsoftin "Satori"-järjestelmissä. Linked Data -teknologia on keskeisessä asemassa monissa Digital Humanities -alueen hankkeissa (Hyvönen, 2012), jollainen Sotasampokin on.

Kiitokset Kiitos Kansallisarkiston Tomi Ahorannalle, puolustusvoimien Pasi Leskiselälle sekä Timo Hakalalle ja Jyrki Tiittaselle datojen kehittämisestä ja luovuttamista Sotasammon käyttöön. Jérémie Dutruit ja Kasper Apajalahti kehittivät eräitä Sotasammon osia. Sotasammon syntymiselle oli ratkaisevaa opetus- ja kulttuuriministeriön rahoitus.

KIRJALLISUUTTA

1. Tom Heath, Christian Bizer: *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, Palo Alto, CA, 2011.
2. Eero Hyvönen: *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Morgan & Claypool, Palo Alto, CA, 2012.
3. Eero Hyvönen, Thea Lindquist, Juha Törnroos and Eetu Mäkelä: *History on the Semantic Web as Linked Data - An Event Gazetteer and Timeline for World War I*. Proceedings of CIDOC 2012 - Enriching Cultural Heritage, Helsinki, Finland, CIDOC, <http://www.cidoc2012.fi/en/cidoc2012/programme>, June, 2012.
4. Eero Hyvönen, Jouni Tuominen, Eetu Mäkelä, Jérémie Dutruit, Kasper Apajalahti, Erkki Heino, Petri Leskinen and Esko Ikkala: *Second World War on the Semantic Web: The WarSampo Project and Semantic Portal*. Proceedings of 14th International Semantic Web Conference 2015 (ISWC 2015), Posters and Demos, CEUR-WS Proceedings, Bethlehem, PA, USA, October, 2015a. <http://www.seco.tkk.fi/publications>
5. Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen and Eetu Mäkelä: *Publishing Second World War History as Linked Data Events on the Semantic Web*. October, 2015b. Submitted. <http://www.seco.tkk.fi/publications>
6. Eero Hyvönen. FinnONTO-hanke loi perustan kansalliselle tietoinfrastruktuurille verkossa. Tieteessä tapahtuu, 3/2014. <http://ojs.tsv.fi/index.php/tt/article/view/41559>
7. Eero Hyvönen, Jouni Tuominen, Miika Alonen, Eetu Mäkelä. *Linked Data Finland – A 7-star Model and Platform for Publishing Linked Data*. Proceedings of ESWC 2014, Demo and Poster Papers. Springer-Verlag, 2014.
8. Thea Lindquist, Michael Dulock, Juha Törnroos, Eero Hyvönen and Eetu Mäkelä: *Using Linked Open Data to Enhance Subject Access in Online Primary Sources*. *Cataloging & Classifying Quarterly*, vol. 51, no. 8, Francis & Taylor, 2013.
9. Eetu Mäkelä, Thea Lindquist and Eero Hyvönen: *World War 1 as Linked Open Data*. Submitted, 2015. <http://www.seco.tkk.fi/publications>

⁸ <http://www.w3.org/standards/semanticweb/>

⁹ <http://dbpedia.org/>

¹⁰ <http://www.w3.org/2001/sw/>