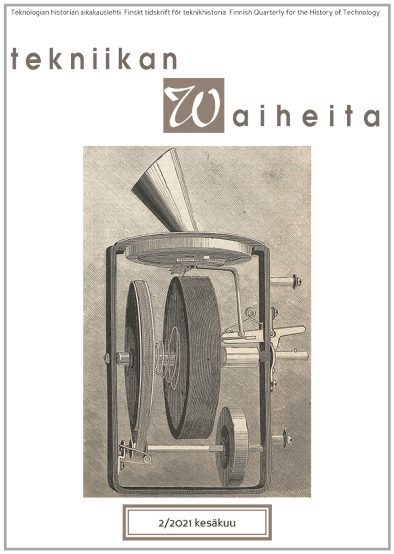


Tekniikan Waiheita
ISSN 2490-0443
Tekniikan Historian Seura ry.
39. vuosikerta: 2
2021
https://journal.fi/tekniikanwaiheita

Sammon taontaa semanttisessa webissä

Eero Hyvönen

To cite this article: Eero Hyvönen, ”Sammon taontaa semanttisessa webissä” Tekniikan Waiheita 39, no. 2 (2021): 87–105. https://doi.org/10.33355/tw.102864

To link to this article: https://doi.org/10.33355/tw.102864

Sammon taontaa semanttisessa webissä

Eero Hyvönen[1]

Internetissä on miljardeja verkkosivuja, joiden sisältöä voidaan hakea Googlen kaltaisilla hakukoneilla ja selailla linkkien avulla. Tämän ihmiselle näkyvän ”sivujen verkon” (Web of Pages) sisälle on rakentunut dataan perustava ”tiedon verkko” (Web of Data), semanttinen web. Se linkittää toisiinsa käsitteitä, tietoa ja dataa tietokoneiden ymmärtämällä tavalla. Semanttisen webin kehittäminen käynnistyi toden teolla 20 vuotta sitten vuonna 2001 webin infrastruktuuria koordinoivan World Wide Web -konsortion (W3C) ja sen johtajan, webin ”isän” Tim Berners-Leen johdolla. Samana vuonna järjestettiin Suomessa konferenssi Semantic Web Kick-off in Finland ja ensimmäiset kotimaiset tutkimushankkeet käynnistyivät. Artikkelissa esitellään semanttisen webin idea, lyhyt kansainvälinen historia ja Suomessa tehtyä tutkimustyötä Aalto-yliopistossa ja Helsingin yliopistossa erityisesti digitaalisten ihmistieteiden saralla.

Akseli Gallen-Kallela: Sammon taonta (yksityiskohta), 1893, Ateneumin taidemuseo (tekijänoikeusvapaa)

[1] Semanttisen laskennan tutkimusryhmä (SeCo) professori, tietotekniikan laitos, Aalto-yliopisto johtaja, HELDIG-keskus, Helsingin yliopisto, http://seco.cs.aalto.fi/u/eahyvone

Tiedon verkko – Web of Data

Älykkäiden verkkopalveluiden edellytyksenä on, että tietokoneet ymmärtävät verkon sisältöjä. Tämä on mahdollista kahdella tavalla. Konetta voidaan opettaa ymmärtämään verkon sisältöjä ihmisen tapaan tekoälyn avulla ja muodostamaan automaattisesti sovellusten tarvitsemia tietorakenteita. Lähestymistavan haasteena on mm. luonnollisen kielen ymmärtämisen vaikeus. Toisaalta koneelle voidaan tarjoilla valmiiksi pureskellussa muodossa dataa, kuten museoiden, kirjastojen ja arkistojen kokoelmatietoja, sosiaalisen median verkostoja tai yritysten tuotekuvauksia. Käytännössä molempia lähestymistapoja ja niiden yhdistämistä tarvitaan.

Semanttisen webin[2] ideana on esittää verkon sisällöt ns. semanttisena verkkona, jonka merkitys (semantiikka) on määritelty logiikan avulla. Logiikka on yleinen oppisuunta ja mekaaninen malli ihmisen ajattelusta. Sen kehitys alkoi Aristoteleen (382−322 eaa.) muotoiltua ensimmäiset päättelysääntönsä eli syllogismit. Logiikasta on tullut myöhemmin mm. klassisen tekoälyn perusta.[3] Loogista päättelyä ei ole sidottu mihinkään tiettyyn sovellusalaan tai luonnolliseen kieleen ja se soveltuu siksi monialaisen ja monikielisen webin semanttiseksi perustaksi.

Semanttinen web kokoaa yhteen ihmiskunnan tietoa muodossa, jota tietokoneet voivat yhdistellä, ”ymmärtää,” ja käsitellä laskennallisesti tekoälyn avulla. Tällaisen tiedon verkkoon kuuluu esimerkiksi se tosiasia, että Väinö Linna oli Urjalassa vuonna 1920 syntynyt ja Kangasalalla 1992 kuollut suomalainen kirjailija, ja että Mars on aurinkokuntamme neljäs planeetta, jonka säde on 3390 km. Maailmanlaajuinen semanttinen web, josta on käytetty myös nimitystä Giant Global Graph (GGG), sisältää tuhansia toisiinsa yhdistettyjä verkkomuotoisia aineistoja, kuten Wikipedioiden sisältöä datana julkaiseva Wikidata[4] ja DBpedia[5], GeoNames-palvelun[6] miljoonat paikkatiedot, yleiseurooppalainen kulttuurikokoelmia yhdistävä Europeana[7] ja monien kansalliskirjastojen viitetietokannat, kuten Suomen Kansallisbibliografia[8]. Dataa verkkosivuilla julkaisemalla yritykset ja organisaatiot kertovat Googlen kaltaisille hakukoneille ja muille verkkopalveluille tarjoamistaan tuotteista, palveluista, aukioloajoistaan ja sijainnistaan. Facebookin kaltaiset sosiaalisen median palvelut hyödyntävät semanttista webiä mm. käyttäjien kiinnostuksen kohteiden ja verkostojen esittämisessä, palveluiden personoinnissa ja mainonnan kohdistamisessa. Jokainen Facebookin käyttäjä on osa jättiläismäistä semanttista verkkoa nimeltä Open Graph[9].

[2] W3G, Semantic Web: https://www.w3.org/standards/semanticweb/ ; Suomeksi aiheesta on ilmestynyt oppikirja Hyvönen 2018.

[3] Sowa 2000.

[4] Wikidata, kotisivu: https://www.wikidata.org/wiki/Wikidata:Main_Page

[5] DBpedia, kotisivu: https://wiki.dbpedia.org/

[6] Geonames, kotisivu: http://www.geonames.org/

[7] Europeana: https://pro.europeana.eu/page/linked-open-data

[8] Kansallisbibliografia: http://urn.fi/URN:NBN:fi:bib:me:W0006048250O

[9] Open Graph Protocol, Facebook: https://ogp.me/

Semanttisen webin historiaa

Semanttisen webin idea syntyi samaan aikaan kuin koko webin idea jo 90-luvun vaihteessa[10]. Ensin kehitettiin kuitenkin ihmisille tarkoitettu hypertekstiin perustuva World Wide Web (WWW). Ratkaiseva käännekohta GGG:n kehittämiselle oli toukokuussa 2001 *Scientific American* -lehdessä ilmestynyt artikkeli[11], jossa kiteytettiin ajatus semanttisesta webistä ja sitä hyödyntävistä verkkopalvelusta ja älykkäistä agenteista (agent). Nämä voisivat esimerkiksi suunnitella automaattisesti vaikkapa tutkijan konferenssimatkan ja tehdä tarvittavat varaukset. Visiota toteuttamaan käynnistettiin W3C:ssa erityinen semanttisen webin ohjelma Semantic Web Activity ja sitä tukeva verkkopalveluiden ohjelma Web Service Activity.

2000-luvun alussa alan keskeiseksi tutkimusteemaksi muodostui aluksi ontologiat[12], joiden avulla voidaan määritellä ja kuvata reaalimaailman liittyvät käsitteet, tietomallit ja data sovelluksia varten[13]. Vuosikymmenen lopulla alan huomio kiinnittyi dataan, tietoyhteiskuntien polttoaineeseen. Silloin lanseerattiin linkitetyn datan (Linked Data) idea ja julkaisuperiaatteet datapalveluina ja WWW-sivuille upotettuna merkkauksina[14]. Tässä vaiheessa aiheesta kiinnostuivat mm. Googlen ja Microsoftin kaltaiset verkkojätit, jotka sopivat vuonna 2015 datan verkkosivuille upottamisessa tarvittavasta Schema.org[15] määrittelystä ja alkoivat kehittää hakukoneiden perustaksi jättiläismäisiä semanttisia verkkoja, Google Knowledge Graphia ja Microsofti Satoria, mikä termi tarkoittaa Japanin zenbuddhalaisuudessa valaistumista.

Linkitettyä dataa kerätään ja julkaistaan eri maissa ja sovellusalueilla verkossa oleviin datapilviin ja -palveluihin. Näistä yksi tunnetuimmista on Linked Open Data Cloud[16]. Sen ytimeen tilastoitiin esimerkiksi vuonna 2018 kuuluvaksi 10 000 toisiinsa linkitettyä datajoukkoa ja 150 miljardia tietojen välistä yhteyttä. Esimerkiksi se tosiasia, että Väinö Linna kirjoitti Tuntemattoman sotilaan muodostaa yhden yhteyden käsitteiden ”Väinö Linna” ja ”Tuntematon sotilas -romaani” välillä, ja ”kirjoittaja”-ominaisuus romaanista sen julkaisuvuoteen 1955 toisen.

Linkitetyn datan perustalle on syntynyt uusia innovatiivisia toimintamalleja ja käytännön sovelluksia, joita on otettu käyttöön tietoyhteiskunnissa eri puolilla maailmaa. Monissa maissa, kuten Iso-Britanniassa, avattiin julkisen sektorin avoimen linkitetyn datan portaaleja. Näiden kautta käyttäjä löytää sekä tietovarantoja että niihin perustuvia fokusoituja sovelluksia, vaikkapa lähellä olevien koulutusmahdollisuuksien tai hoitokotien löytämiseen tai liikennetietojen seuraamiseen. Yksi ensimmäisiä sovelluksia oli BBC:n kotisivujen verkkopalvelu, jossa organisaation eri tahojen mediasisältöjen yhdistämiseen käytettiin Wikipedian semanttisen webin muunnosta DBpediaa.

Linkitetyn datan käsitteisiin liittyvään tietoon voi tutustua selaimella URI/URL-tunnisteiden avulla. Esimerkiksi Väinö Linna -käsitteen data DBpediassa löytyy osoitteesta:

http://dbpedia.org/resource/Väinö_Linna

[10] Berners-Lee & Fischetti 1999.

[11] Berners-Lee et al. 2001.

[12] Staab & Studer 2009.

[13] Hitzler 2021.

[14] Heath & Bizer, 2011.

[15] Schema.org määrittelyt: https://schema.org/

[16] Linked Open Data Cloud: https://lod-cloud.net/

Datassa näkyy mm. linkki syntymäpaikkaan Urjala (http://dbpedia.org/resource/Urjala) ja romaaniin Tuntematon sotilas kirjoittajan ominaisuudessa. Vastaavasti Urjalan ja Tuntemattoman sotilaan käsitteet linkittyvät eteenpäin niitä kuvaileviin tietoihin. Tiedot ovat käytettävissä paitsi HTML-sivuina ihmislukijaa varten myös standardimuotoisena datana (RDF-muodossa) sovelluksissa ohjelmointirajapinnan kautta.

Wikipedian dataa kerätään nykyisin määrätietoisesti Wikipedia-yhteisön itsensä ylläpitämään Wikidata-järjestelmään. Sen ideana on muodostaa erikielisten Wikipedioiden perustaksi yhteinen kieliriippumaton linkitetyn datan ydin, Wikidata, jota sitten hyödynnetään eri kielissä, ihmislukijoille tarkoitetuissa Wikipedioissa. Wikidatassa voidaan esimerkiksi esittää tieto Helsingin asukasluvusta tai Yhdysvaltojen nykyisestä presidentistä, jolloin sitä ei tarvitse erikseen kertoa Wikipedian espanjan kielisessä tai kymmenissä muissa laitoksissa. Kun Helsingin asukasluku muuttuu tai Yhdysvaltoihin valitaan uusi presidentti, voidaan erikieliset Wikipediat päivittää automaattisesti Wikidatasta.

Linkitetyn datan jälkeen semanttisen webin keskeinen teema ja avaintermi on ollut tietämysverkot (knowledge graph), jotka ovat olennaisesti yhdistelmä ontologioita ja linkitettyä dataa käytettynä jonkin yrityksen tai rajatun yhteisön datan hallintaan, hyödyntämiseen ja julkaisemiseen[17]. Tietämysgraafit voivat linkittyä toisiinsa ja näin kasvaa vähitellen myös globaali semanttinen web osiensa summana.

Maailmanlaajuisen tiedon verkon syntyminen on mahdollista vain kieli- ja kulttuurirajat ylittävällä yhteistyöllä, mistä perinteinen WWW standardeineen on eriomainen esimerkki. Semanttisen webin kehittämisessä keskeistä onkin ollut kansainvälinen yhteistyö teknisten standardien ja parhaiden käytäntöjen luomiseksi. Tätä työtä on ohjannut ja koko webin kehitystyötä koordinoinut World Wide Web Consortium (W3C) johtajanaan Tim Berners-Lee, joka on saanut suomalaisen Millenium-palkinnon vuonna 2004 lukuisten muiden tunnusten ohella.

Standardien kehittämisen ohella toinen keskeinen idea semanttisessa webissä on yhteentoimivan datan julkaiseminen verkossa siten, että dataa voidaan löytää, hakea ja rikastaa toisten datajulkaisujen ja tekoälyn avulla, ja että dataa voidaan mahdollisimman helposti käyttää uusissa sovelluksissa. Linkitetyn avoimen datan maailmassa toteutuvat modernit FAIR-periaatteet[18] sille, että tiedon pitää olla löydettävissä (Findable), saavutettavissa (Accessible), yhteentoimivaa (Interoprable) ja uudelleen käytettävää (Re-usable)

Kolmas ja tavallisen verkon käyttäjän kannalta kiinnostavin semanttisen webin kehityssuunta on linkitetyn datan ja datapalvelujen avulla luotavat käytännön sovellukset eri aloilla.

Kehitystyötä Suomessa

Kehitys Suomessa seurasi kansainvälistä kaviouraa ontologioista linkitetyn datan kautta tietämysgraafeihin ja sovelluksiin; työssä on kuitenkin ollut alusta alkaen vahva panostus käytännön sovelluksiin. Lähtölaukaus alan kehittämiselle ammuttiin Semantic Web Kick-off in Finland -konferenssissa[19] lokakuussa 2001 joitain kuukausia Scientific American -lehden artikkelin ilmestymisen jälkeen. Tilaisuus järjestettiin Helsingin yliopiston ja Teknillisen kor-

[17] Noy et al. 2019.

[18] FAIR-periaatteet: https://www.go-fair.org/

[19] Hyvönen 2002.

keakoulun HIIT-tutkimuskeskuksessa yhteistyössä Suomen tekoälyseuran ja W3C:n edustajien kanssa ja siihen osallistui yli 200 henkeä.

Helsingin yliopistossa ja Teknillisessä korkeakoulussa (HIIT) käynnistyivät vuoden 2002 alussa ensimmäiset tutkimushankkeet. Tutkimustyön prototyyppeinä valmistui Promoottori-sovellus[20] Helsingin yliopiston museoon edistämään promootioperinnettä ja vuonna 2004 julkistettiin Kansallismuseossa pidetyssä tilaisuudessa Museosuomi – Suomen museot semanttisessa webissä[21], varhainen esikuva nykyisille museokokoelmia aggregoiville verkkopalveluille kuten Finna.fi ja Europeana.eu.

Visio kansallisesta tietoinfrastruktuurista

Tässä työssä syntyi visio yhteisen kansallisen semanttisen webin avoimen ”sisältöinfrastruktuurin” luomisesta[22], joka täydentäisi W3C:n standardien yleistä loogista viitekehystä joukolla toisiinsa yhdistyviä ja täydentäviä alakohtaisia suomalaisia ontologioita. Nämä perustuisivat maassamme jo käytössä oleviin laajoihin asianastoihin, joiden kruununa olisi Kansalliskirjaston Yleisen suomalaisesta asiasanaston (YSA) pohjalle luotava Yleinen Suomalainen Ontologia (YSO)[23]. Ajatuksena oli, että yhteinen eri alojen linkittynyt ontologioiden pilvi, joka nimettiin sittemmin KOKO-ontologiaksi, voitaisiin ottaa kustannustehokkaasti käyttöön erityisillä ontologiakirjastopalveluilla[24].

Vision toteuttamiseksi käynnistyi FinnONTO-hankkeiden sarja 2003−2012[25], joita rahoitti Tekes ja lopulta lähes 40 eri organisaatiota. Näin syntyi KOKO-ontologioiden pilvi ja ONKI-ontogiapalvelu, joka tuotteistettiin[26] Kansalliskirjastossa vuonna 2014 nykyiseksi Finto.fi-palveluksi − nimi on muistuma FinnONTO-projektista. Vuonna 2019 Fintolla oli 280 000 käyttäjää ja sen rajapintoihin tehtiin 32 miljoonaa kutsua. Ontologiatyö on jatkunut Kansalliskirjastossa. Esimerkiksi Finton uusi KANTO-ontologia kattaa kansallisbibliografian kuvailun yhteydessä tuottamat ohjeelliset nimenmuodot Suomessa julkaistujen aineistojen toimijoista mukaan lukien musiikkiaineistojen tekijät.

FinnONTO-hanketta seurasi Linked Data Finland -hanke[27] (LDF), jossa ONKI-konseptia laajennettiin linkitetyn datan julkaisemiseen. Tätä varten toteutettiin Linked Data Finland alusta LDF.fi[28], jossa on julkaistu kymmenittäin suomalaisia ja kansainvälisiä datajoukkoja toiminnallisina palveluina ns. SPARQL-palvelupisteinä[29]. Sekä FinnONTO- että LDF-hankkeissa kehitettiin infrastruktuurin testaamiseksi ja arvioimiseksi sovelluskohtaisia tietämysgraafeja ja näihin perustuvia portaaleja ja verkkopalveluita. Tämä työ on jatkunut aktiivisena ja suuntautunut yhä enemmän kulttuurialan ja digitaalisten ihmistieteiden linkite-

[20] Hyvönen et al. 2004.

[21] Hyvönen et al. 2005. Demonstraattori on edelleen kokeiltavissa osoitteessa http://museosuomi.fi.

[22] Hyvönen et al. 2008.

[23] Seppälä & Hyvönen 2014.

[24] Viljanen et al. 2008.

[25] FinnONTO-hankkeiden kotisivut: https://seco.cs.aalto.fi/projects/finnonto/

[26] Suominen et al. 2014.

[27] Linked Data Finland -hankkeen kotisivu: https://seco.cs.aalto.fi/projects/ldf/

[28] Linked Data Finland – alusta: https://ldf.fi; Hyvönen et al. 2014.

[29] SPARQL on W3C:n standardoima linkitetyn datan kyselykieli: https://www.w3.org/TR/sparql11-query/

tyn avoimen datan infrastruktuurin kehittämiseen, josta on alettu käyttää nimitystä Linked Open Data Infrastructure for Digital Humanities (LODI4DH)[30]. Infrastruktuuria ylläpidetään nykyään CSC − Tieteen tietotekniikan keskus Oy:n tarjoamilla palvelimilla.

Humanistisilla aloilla infrastruktuurin käsite on abstraktimpi ja epämääräisempi kuin luonnontieteissä tai kieliteknologiassa. Helsingin yliopiston Digitaalisten ihmistieteiden keskuksen HELDIG:n johdolla koordinoitu laaja-alainen kansallinen ehdotus ”Common Language Resources and Technology Infrastructure” (FIN-CLARIAH) on kuitenkin saatu vuonna 2020 mukaan Suomen Akatemian tutkimusinfrastruktuurien uudelle tiekartalle. Mukana FIN-CLARIAH yhteistyössä ovat Helsingin yliopisto, Aalto-yliopisto, CSC – tieteen tietotekniikan keskus Oy, Itä-Suomen yliopisto, Jyväskylän yliopisto, Kansallisarkisto, Kotimaisten kielten keskus, Tampereen yliopisto, Turun yliopisto ja Vaasan yliopisto. Tavoitteena on liittyminen jatkossa täysjäsenenä EU-tason DARIAH-infrastruktuuriohjelmaan CLARIN-ohjelman tapaan, jossa Suomi on ollut mukana. ”FIN-CLARIAH”-nimi tulee ajatuksesta yhdistää Suomessa CLARIN- ja DARIAH-ohjelmat, sillä ne muodostavat synergeettisen kokonaisuuden digitaalisissa ihmistieteissä. Vastaava ajatus on otettu käyttöön jo Alankomaiden CLARIAH-ohjelmassa.

Sampo-malli

Suomessa semanttisen webin teknologiaa ja infrastruktuuria on kehitetty ja sovellettu erityisesti ”Sampo-järjestelmissä”[31], joilla on ollut miljoonia käyttäjiä verkossa. Näitä järjestelmiä yhdistää toisiinsa niiden toteuttamisen tuloksena vähitellen syntynyt ”Sampo-malli”, joka on kehitetty Aalto-yliopiston ja Helsingin yliopiston Semanttisen laskennan tutkimusryhmässä (SeCo). Nimi johtuu Kalevalan Sammon ehkä yleisimmästä tulkinnasta muinaisen edistyneen teknologian metaforana.

Sampo-mallin perustana on linkitetyn datan idea yleisöllisestä julkaisemisesta, jossa kaikki voittavat: Tiedon julkaisijat voivat rikastaa sisältöjään ”ilmaiseksi” toisten julkaisijoiden dataa linkittämällä ja uutta tietoa päättelemällä, ja loppukäyttäjille voidaan tarjota aiempaa runsaampia tietosisältöjä aiempaa älykkäämpien käyttöliittyminen ja työkalujen kautta. Datan rikastaminen perustuu eri aineistoja yhdistävään ontologiainfrastruktuurin, jonka mukaisiksi eri aineistot muunnetaan tai käsitteistöjen välille rakennetaan yhteys (ontology mapping).

Sampo-mallin mukaisessa järjestelmässä on kaksi erillistä osaa: linkitetyn datan SPARQL-palvelupiste ja sitä hyödyntävä käyttöliittymäkerros, semanttinen Sampo-portaali. Datapalvelu pyritään julkaisemaan avoimena datana, jolloin sitä voivat hyödyntää rajapintojen kautta tai dataa lataamalla kaikki halukkaat. Tämä mahdollistaa toisaalta data-analyysien ja visualisointien tekemisen digitaalisten ihmistieteiden tutkimusmenetelmillä ja toisaalta uusien sovellusten kehittämisen, kuten Sampo-portaaleissa. Datapalvelua voi käyttää suoraan rajapintojen kautta millä tahansa ohjelmointikielellä webin HTTP-protokollaa käyttäen.

[30] LODI4DH -hankkeen kotivisu: https://seco.cs.aalto.fi/projects/lodi4dh/

[31] Sampo-portaalit ja videot: https://seco.cs.aalto.fi/applications/sampo/; Hyvönen 2020b.

Käytettävissä on myös valmiita käyttöympäristöjä, esimerkiksi SPARQL-kieltä tukeva YASGUI-järjestelmä[32] ja Google Colab- ja Jupyter-järjestelmät[33] Python skriptien ja visualisointien käyttämiseen. Näin toteutettavien analyysien rajana on vain tutkijan mielikuvitus, ohjelmointitaito ja tietysti käytettävissä olevan datan ominaisuudet, kattavuus, ja laatu. Loppukäyttäjille suunnattujen Sampo-portaalien käyttö ei edellytä ohjelmointitaitoa.

Sampo-portaalien käyttöliittymien toteutuksessa keskeinen idea on ollut fasettihaun[34] yhdistäminen data-analyyttisiin työkaluihin. Ajatuksena on eräänlainen käyttölogiikan standardointi Sampo-UI-kehykseksi ja työkaluksi[35]. Ideana on, että hakukohteista suodatetaan ensin esiin kiinnostuksen kohteena oleva joukko kohteita tekemällä valintoja hierarkkisista faseteista. Tämän jälkeen tulosjoukkoa voidaan tutkia tarkemmin kohdistamme siihen mm. tilastollisia analyysejä, visualisointeja ja verkostoanalyysejä. Tämä ajatus sai innoitusta prosopografian tutkimuksesta[36].

Avointa yhteistä infrastruktuuria ja työkaluja yhä uudelleen hyödyntämällä ja asteittain kehittämällä uusien sampojen kehittäminen on saatu kustannustehokkaaksi, kun pyörää ei tarvitse keksiä joka kerta uudestaan ja ontologioita ja data-aineistoja voidaan hyödyntää uudelleen uusissa sovelluksissa. Lisäksi W3C:n standardeja käyttämällä voidaan rakentaa tiedon valtaväylää impivaarasta kansainvälisen semanttisen webin intrastruktuureihin. Suomeen vähitellen rakennettu infrastruktuuri ja linkitetyn datan hyödyntämismalli ovat kansainvälisesti poikkeuksellisia esimerkkejä kehitystyön systemaattisuuden ja pitkäjänteisyyden näkökulmista.

Kohti tekoälyperustaista kulttuuriaineistojen julkaisemisesta

Sampo-portaalit ja niiden pohjaksi kehitetty Sampo-malli ovat esimerkki paradigman muutoksesta, jossa kulttuurialalla on siirrytty ensin painettujen tekstien julkaisemista verkossa oleviin tietokantoihin hakukoneineen. Seuraavana askeleena ovat sampojen kaltaiset järjestelmät, joissa verkkojulkaisuun on integroitu saumattomasti data-analyyttisiä työkaluja digitaalisten ihmistieteiden tutkijoille.

Nyt ollaan ottamassa uutta askelta kohti tekoälyperustaisia järjestelmiä (knowledge discovery, computational creativity), joissa tietokone ei ole vain passiivinen työkalu, vaan voi osallistua aktiivisesti tutkimusongelmien etsimiseen, ratkaisemiseen ja jopa ratkaisujen selittämiseen.[37] Douglas Adamsin klassikkoromaanissa *Linnunradan käsikirja liftareille* (Hitchhikers Guide to the Galaxy) tietokoneelta haluttiin vastaus kysymykseen elämästä, maailmankaikkeudesta ja kaikesta muusta sellaisesta. Koneen antama vastaus ”42” voi olla oikea, mutta jäi epäselväksi, mikä oikeastaan oli kysymys, ja tutkija kuulisi mielellään myös perustelun vastaukselle.

[32] Rietveld & Hoekstra 2017.

[33] Google Golab -verkkotyökalu Python-ohjelmointiin: https://colab.research.google.com/notebooks/intro.ipynb

[34] Tunkelang 2009.

[35] Ikkala et al. 2021.

[36] Verboven et al. 2007.

[37] Hyvönen 2021.

Sampoja verkossa

Tarkastelen seuraavassa Sampo-mallin ajallista kehittymistä esittelemällä SeCo-tutkimusryhmän toimesta kehitettyjä ja verkossa julkaistuja Sampo-portaaleja:

Museosuomi – Suomen museot semanttisessa webissä (2004)

Museosuomi[38] oli ensimmäinen Sampo-mallia käyttävä järjestelmä. Sen ideana oli kerätä yhteen kokoelmatietoa eri museoiden tietokannoista, rikastaa aineistoja ja julkaista ne globaalina kansallisena verkkopalveluna. Tämä kansainvälisen Semantic Web Challenge palkinnon saanut verkkopalvelun prototyyppi julkistettiin vuonna 2004 Kansallismuseossa.

Kulttuurisampo – Suomalainen kulttuuri semanttisessa webissä (2008)

Kulttuurisampo.fi[39] yleisti MuseoSuomi-konseptin kaikenlaisten kulttuurialan sisältöjen julkaisemiseen linkitettynä datana, mukaan lukien kokoelmien ohella myös ei-aineellinen kulttuuri, kuten perinteiset taidot, musiikki ja kansanrunous. Järjestelmän keskiössä esimerkiksi oli suomalaisen kulttuurin ytimenä ”Semanttinen Kalavala”, Kalevalan ensimmäinen ”käännös” tietokoneiden ymmärtämään muotoon semanttisen webin RDF-kielellä[40], mistä ”Sampo”-nimi otettiinkin yleisempään käyttöön.

Tervesuomi (2008)

Tervesuomi.fi[41] sovelsi Sampo-mallia terveyden edistämisen verkkoaineistoihin, joita tuottaa Suomessa yli sata alan järjestöä ja toimijaa. Hanke oli Terveyden ja hyvinvoinnin laitoksen (THL) koordinoima, ja SeCo-ryhmässä kehitetty julkaisukonsepti ja prototyyppi tuotteistettiin yhdeksi THL:n virallisista tietoportaaleista. Tervesuomi on saanut Museosuomen tapaan kansainvälisen semanttisen webin tiedeyhteisön Semantic Web Challenge -sovelluspalkinnon.

Kirjasampo (2011)

Kirjasampo.fi[42] oli alun perin osa Kulttuurisampoa, mutta elää nykyään omaa elämäänsä yleisten kirjastojen ylläpitämänä erillisenä palveluna, jolla oli vuonna 2020 noin kaksi miljoonaan käyttäjää. Järjestelmän ytimessä on laaja tietämysgraafi, joka sisältää semanttisesti rikastettua dataa kaikesta suomalaisesta kertomakirjallisuudesta. Palvelussa on nykyisin myös tietokirjallisuutta. Tämän sammon käyttöliittymä on yleisten kirjastojen (Kirjastot.fi) toteuttama.

[38] Museosuomi-portaali vuodelta 2004: http://museosuomi.fi/; Hyvönen et al. 2005.

[39] Hyvönen et al. 2009; Mäkelä et al. 2012.

[40] Palonen et al. 2009.

[41] Hyvönen et al. 2007; Suominen et al. 2009.

[42] Mäkelä et al. 2011a.

Matkailusampo (2011)

Matkailusampo.fi[43] oli prototyyppi, jossa Sampo-konseptia tutkittiin kulttuurialan matkailukohteisiin liittyvän tiedon julkaisemisessa mobiilisti. Portaali joutui hakkereiden hyökkäyksen kohteeksi ja saastuttamaksi eikä se enää ole käytettävissä verkossa.

Sotasampo (2015–2019)

Tunnetuimpia sampoja on talvi- ja jatkosodan aineistoja julkaiseva Sotasampo.fi[44]. Sen semanttiseen verkkoon kuuluu laaja joukko käsitteitä ja 14,3 miljoonaa niiden välistä yhteyttä, ja datajoukko on otettu osaksi kansainvälistä Linked Open Data -pilveä. Data on saatavilla avoimesti Linked Data Finland -palvelusta, jonka SPARQL-rajapintaan Sotasammon yhdeksän sovellusnäkymää suoraan perustuvat. Järjestelmään on yhdistetty Kansallisarkiston tuottamat tiedot kaikista viime sodissamme menehtyneestä noin 95 000 sotilaasta ja tuotu eri lähteistä tietoja tuhansista muista sodasta selvinneistä tunnetuista sotilaista, kuten Mannerheimin ristin ritareista. Puolustusvoimien SA-Kuva-arkistosta on käytössä noin 160 000 autenttisen sota-ajan valokuvan kokoelma. Sotasammossa on tietoa tuhansista sodanajan tapahtumista ja kymmenistä tuhansista luovutetun alueen paikoista historiallisilla kartoilla. Aineistoja on linkitetty automaattisesti toisiinsa ja ulkoisiin aineistoihin, kuten Kansallisarkiston sotapäiväkirjoihin, Suomen Sotahistoriallisen Seuran verkossa julkaisemien *Kansa Taisteli* -lehtien (1957−1986) tuhansiin muisteluartikkeleihin ja Wikipediaan.

Vuonna 2015 valmistuneeseen Sotasammon 1. versioon on kehitetty myöhemmin uudet sovellusnäkymät puolustusvoimien valokuvista, sankarihautausmaista ja Neuvostoliittoon joutuneista suomalaisista sotavangeista yhteistyössä Kansallisarkiston, Kaatuneiden muistosäätiön ja Suomen Kameraseurojen liiton kanssa. Sotasampo oli yksi Suomen itsenäisyyden 100. juhlavuoden Suomi 100 -hankkeita. Sotasammosta on muodostunut suosittu palvelu, jota on käyttänyt yli 740 000 käyttäjää. Sovellus sai vuonna 2017 LODLAM Open Data Prize -palkinnon Venetsiassa.

Vanhat Norssit semanttisessa webissä (2017)

Vanhat Norssit semanttisessa webissä[45] on Helsingin Normaalilyseon 150-vuotisjuhlan kunniaksi tehty, historiallisiin oppilasmatrikkeleihin 1867−1992 perustuva verkkopalvelu noin 10 000 koulun oppilaasta.

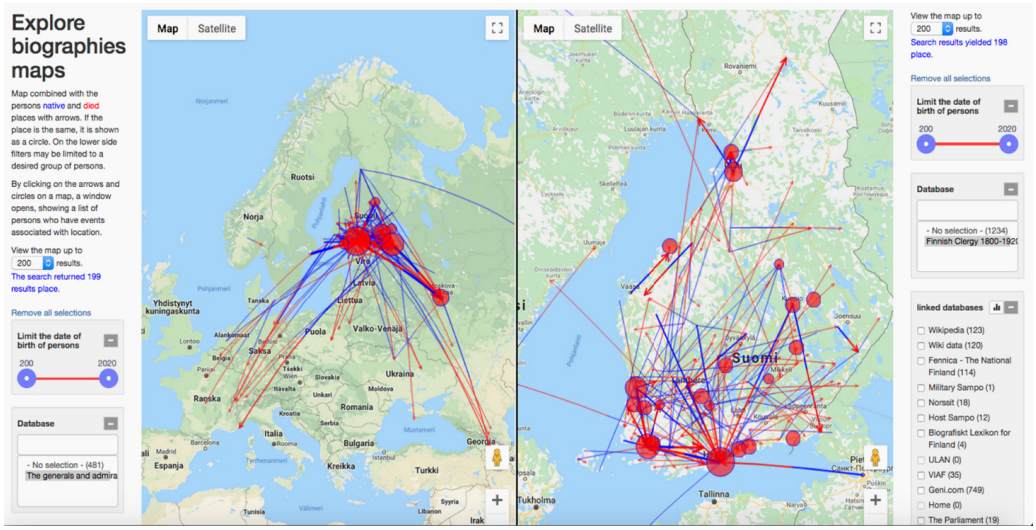
[43] Mäkelä et al. 2011b.

[44] Ks. hankkeen kotisivu https://seco.cs.aalto.fi/projects/sotasampo/ ; Hyvönen et al. 2016; Koho et al., 2021.

[45] Hyvönen et al. 2017.

U.S. Congress Prosopographer (2018)

U.S. Congress Prosopographer[46] on USA:n kongressin edustaja-aineistoihin perustuva portaali, joka kehitettiin yhteistyössä tokiolaisen Keio-yliopiston kanssa. Järjestelmän lähtökohtana oli Vanhat Norssit -sovellus, jota kehitettiin uudella aineistolla ja järjestelmään lisättiin uusia data-analyyttisiä visualisointeja ja toimintoja. Sovelluksen tietämysgraafi sisältää kattavasti tietoa n. 12 000 USA:n kongressin lainlaatijasta vuosilta 1789–2018 ja työkaluja aineistojen prosopografiseen tutkimiseen.



Kuva 1. Suomen suuriruhtinaskunnassa Venäjän sotavoimissa palvelleiden kenraalien ja amiraalin (vasemmalla) ja papiston (oikealla) elämänkaarien prosopografinen vertailu Biografiasammossa.

Biografiasampo (2018)

Biografiasampo[47] on Suomalaisen Kirjallisuuden Seuran (SKS) toimittamiin kansallisiin biografioihin perustuva datapalvelu ja portaali. Biografiasammon ydinaineistona ovat Kansallisbiografia ja muut SKS:n toimittamat ja julkaisemat pienoiselämäkerrat, yhteensä 13 100 elämäntarinaa. Niitä on kirjoittanut 980 suomalaista tutkijaa maamme suurimmaksi sanotussa historiantutkimuksen hankkeessa. Biografiasammossa elämäkerroista louhittua dataa on rikastettu linkittämällä sitä kuuteentoista muuhun tietolähteeseen ja automaattisen loogisen päättelyn avulla. Tietämysverkko on julkaistu linkitetyn avoimen datan palvelussa Linked Data Finland. Järjestelmän innovaationa on luoda kieliteknologian, tekoälyn ja semanttisen webin teknologioiden avulla elämäkertojen teksteistä ja niihin eri lähteissä liittyvistä tiedoista semanttinen verkko, linkitetyn avoimen datan palvelu ja siihen perustuvia sovelluksia historiasta kiinnostuneille tutkijoille ja kansalaisille.

[46] Miyakita et al. 2018.

[47] Ks. hankkeen kotisivu https://seco.cs.aalto.fi/projects/biografiasampo/; Hyvönen et al. 2019.

Biografiasammon yhteyshaku-sovelluksissa on otettu ensimmäisiä askeleita kohti selittävää tekoälyä. Siinä käyttäjä voi muotoilla hakufasettien avulla esimerkiksi hakukysymyksen ”Miten suomalaiset taidemaalarit liittyvät Italiaan”. Vastauksena on joukko semanttisen verkon kautta muodostettuja yhteyksiä selityksillä varustettuna, kuten että Elin Danielsson-Gambogi vastaanotti Firenzen kaupungin palkinnon vuonna 1899 tai ”Robert Wilhelm Ekman on luonut vuonna 1844 taideteoksen ’Maisema Subiacosta’, joka kuvaa paikkaa Italia”.

Datapalvelun avulla on toteutettu seitsemästä sovellusnäkymästä koostuva älykäs, avoin ja maksuton verkkopalvelu Biografiasampo.fi, jolla on ollut kymmeniä tuhansia käyttäjiä. Esimerkiksi kuvassa 1 käyttäjä vertaa toisiinsa kahden ihmisryhmän elämänlankoja, Venäjän sotavoimissa 1809−1917 palvelleita suomalaisia amiraaleja ja kenraaleja (vasemmalla) ja papistoa vuosina 1800−1920 (oikealla). Ryhmät on muodostettu kahdella rinnakkaisella fasettihaulla Biografiasammon elämäkarttojen vertailunäkymässä, jossa elämä kuvataan sinipunaisena nuolena syntymäpaikasta (sininen pää) kuolinpaikkaan (punainen pää). Yhdellä vilkaisulla selviää, että sotilaat liikkuivat pappeja kansainvälisemmin ja elämänsä loppuvaiheessa kohti etelää kuten eläkeläiset nykyään. Yhtä kaarta kartalla klikkaamalla pääsee käsiksi kaareen liittyviin elämäkertoihin tarkempaa tutkimusta varten. Esimerkiksi vasemmalla näkyvä poikkeava kaari Oulusta Länsi-Siperiaan osoittautuu Siperiaan maanmittaustöiden johtajaksi nimitetyn kenraali Gustav Adolf Silverhjelmin aiheuttamaksi.

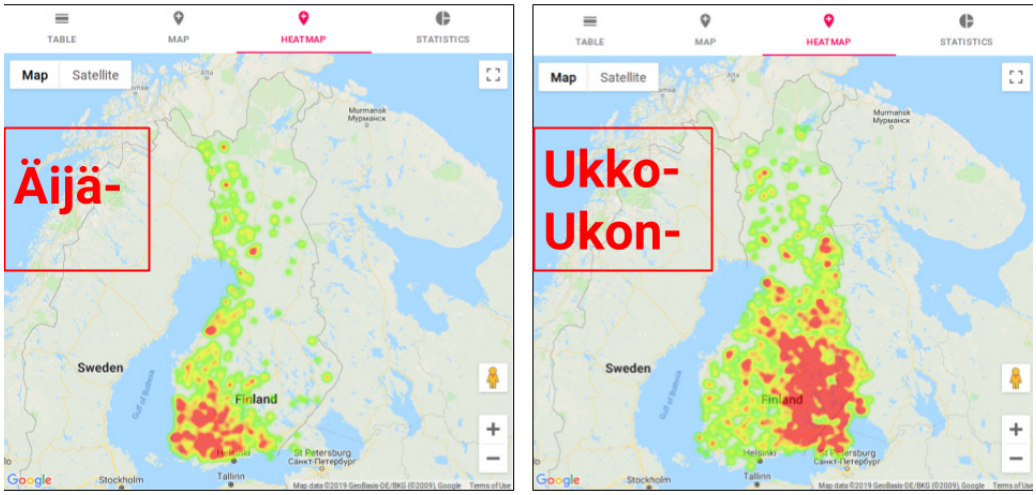
Nimisampo (2018)

Nimisampo.fi[48] julkaisee tietoa Suomen paikannimistä, aineistoina Kotimaisten kielten keskuksen Nimiarkiston kahden miljoonan nimikortin tietokanta, Maanmittauslaitoksen 800 000 paikan Paikannimirekisteri, Sotasammon luovutetun Karjalan paikat ja yhdysvaltaisen Getty-säätiön laaja historiallisten paikkojen TGN-rekisteri. Nimisampo tarjoaa käyttäjälleen älykkään käyttöliittymän, jonka avulla voidaan hakea ja tutkia eri lähteistä paikannimiä ja visualisoida niitä nykyisillä ja historiallisilla kartoilla. Kuvassa 2 Nimisammon käyttäjä esimerkiksi vertailee Äijä- ja Ukko-/Ukon-alkuisten paikannimien esiintymistä Suomessa lämpökarttojen avulla ja huomaa merkittävän eron niiden jakaantumisessa.

Järjestelmän datapalvelua kyselemällä voi myös selvittää esimerkiksi sen, mikä on Suomen yleisin paikannimi. Mitalistit ovat: Riihipelto (3699 kpl) kultaa, Mäkelä (3629 kpl) hopeaa ja Rantala (2872 kpl) pronssia. Tulos vastaa aiemmin eräässä väitöskirjassa saatua tulosta, jonka aikaansaamiseksi tutkijaparka joutui käymään läpi käsityönä pari miljoonaa paikannimikorttia. Nyt vastaava tulos löytyi hetkessä Nimisammon avulla! Järjestelmällä on selvitetty myös missä kunnissa on eniten kirosanoja sisältäviä paikannimiä; tässä kisassa pärjäävät jostain syytä Pohjois-Suomen kunnat. Nimisammolla on ollut kymmeniä tuhansia käyttäjiä. Nimisammosta on luotu vastaavanlainen palvelu Norske Stadnamn Norjaan[49].

[48] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/nimisampo/; Ikkala et al. 2018.

[49] Norjalainen versio Nimisammosta: https://toponymi.spraksamlingane.no/nb/app



Kuva 2. Nimisampo visualisoi lämpökartoilla Äijä- ja Ukko-/Ukon-alkuisten paikannimien jakautumia Suomessa.

Sotasurmasampo 1914–1922 (2019)

Sotasurmasampo 1914–1922[50] sisältää sisällissodan, heimosotien ja 1. maailmansodan 41 500 suomalaisen uhrin tiedot Kansallisarkiston tietokannoista, tietoa 1200 sisällissodan taistelusta sekä data-analyyttisiä työkaluja ja visualisointeja. Tälläkin Sampo-portaalilla on ollut kymmeniä tuhansia käyttäjiä ja sen data on julkaistu avoimesti LDF.fi-alustalla.

MMM (2020)

Mapping Manuscript Migrations (MMM) [51] on datapalvelu ja semanttinen sampo-portaali, joka sisältää metatietoa yli 220 000 keskiaikaisesta käsikirjoituksesta Oxfordin Bodleian kirjastosta, USA:n Schoenberg-instituutista ja Ranskan IRHT-tutkimuslaitoksesta. Järjestelmä on tarkoitettu työkaluksi esimodernin ajan käsikirjoitusten tutkijoille.

Akatemiasampo (2021)

Akatemiasampo[52] perustuu Turun akatemian ja Helsingin yliopiston Ylioppilasmatrikkeleihin, joista on louhittu ja semanttisesti rikastettu avoin datapalvelu ja portaali, hieman vastaavanlainen kuin Biografiasampo. Akatemiasammon aineistot sisältävät yksityiskohtais-

[50] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/; Rantala et al. 2020.

[51] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/mmm/; Burrows at al. 2020; Koho et al. 2021.

[52] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/yo-matrikkelit/; Leskinen ja Hyvönen, 2020; Hyvönen et al., 2021.

ta tietoa kaikista tiedossa olevista akateemisen koulutuksen Suomessa saaneista henkilöistä 1640–1899. Akatemiasammon tietämysgraafissa on noin 6,5 miljoonaa tietojen välistä yhteyttä eli kolmikkoa. Esimerkiksi graafista löytyy tieto siitä, että tutkimusmatkailija James Cookin miehistöön kuuluneen Turun akatemian ylioppilaan Herman Spöring nuoremman (1733–1771) kuolinpaikka on Intian valtamerellä.

Uusissa tutkimushankkeissa on syntymässä lisää sampoja alati kasvavan linkitetyn datan tietoinfrastruktuurin varaan. Näitä ovat arkeologiaan, kansalaistieteeseen ja Museoviraston aineistoihin perustuva Löytösampo[53], missä työssä tehdään myös kansainvälistä yhteistyötä yleiseurooppalaisen ARIADNEPlus-projektin ja British Museumin kanssa, oikeusministeriön kanssa taottava, Suomen lainsäädäntöä ja oikeustapauksia julkaiseva Lakisampo[54], eduskunnan avoimeen dataan perustuva Parlamenttisampo[55] ja valistuksen ajan kirjeaineistoihin perustuva Lettersampo[56]. Lettersammon perustana on yhteistyö Oxfordin yliopiston kanssa ja yleiseurooppalainen EU COST -hanke Reassembling the Republic of Letters 1500–1800, johon osallistui kolmisenkymmentä eri maata.

Datalukutaitoa tarvitaan

Sampojen kaltaisia, automaattisesti datasta muodostettuja järjestelmiä käytettäessä tarvitaan aiempaa enemmän lähdekritiikkiä ja ymmärrystä taustalla olevan datan luonteesta, kattavuudesta ja laadusta. Esimerkiksi Biografiasammon aineistot perustuvat asiantuntijoiden laatimiin kirjoituksiin, mutta tietojen rakenteistamisen, koostamisen, yhdistelyn, rikastamisen ja uuden tiedon muodostamisen on tehnyt paljolti tietokone. Koska aineistot ovat laajoja, ei kokonaisuuden virheettömyyttä voida tarkistaa käsin kuin testimielessä sieltä täältä. Linkityksistä löytyy siksi enemmän virheitä, kuin jos se olisi tehty käsityönä. Tilastollisten ja verkostoanalyysien johtopäätösten kanssa on syytä olla tarkkana ja pitää mielessä, mihin dataan ja laskentamenetelmään ne perustuvat. Tämä on tyypillistä digitaalisissa ihmistieteissä, jossa käsitellään usein niin laajoja aineistoja, ettei systemaattinen ihmistyö ole mahdollista aineistoja muodostettaessa. Laskennallisten tekniikoiden lisäarvo on kuitenkin erityisen suuri tällaista suurdataa (big data) käsiteltäessä ja puutteellinenkin tulos monasti parempi kuin ei mitään tulosta.

Dataa tutkittaessa on aina muistettava, että se heijastelee vain epäsuorasti reaalimaailman ilmiöitä. Esimerkiksi Biografiasammossa elämäkertojen toimituskunnan valinnat siitä, kenestä elämäkertoja kirjoitetaan, vaikuttaa ratkaisevasti dataan. Datan analyysi on siksi luonteeltaan historiografista avaten näkymiä elämäkertoihin ja elämäkertakokoelmien luomisprosessiin. Elämäkertojen biografinen ja prosopografinen data-analyysi nostaa kuitenkin esiin myös taustalla olevaan historiaan liittyviä kiinnostavia henkilöhistoriallisia ilmiöitä, joiden todellisuutta voidaan ryhtyä tutkimaan tarkemmin perinteisin historiantutkimuksen menetelmin. Vastaavanlaista historiografista tutkimusta on tehty myös Iso-Britannian ja Irlannin kansallisbiografioista[57].

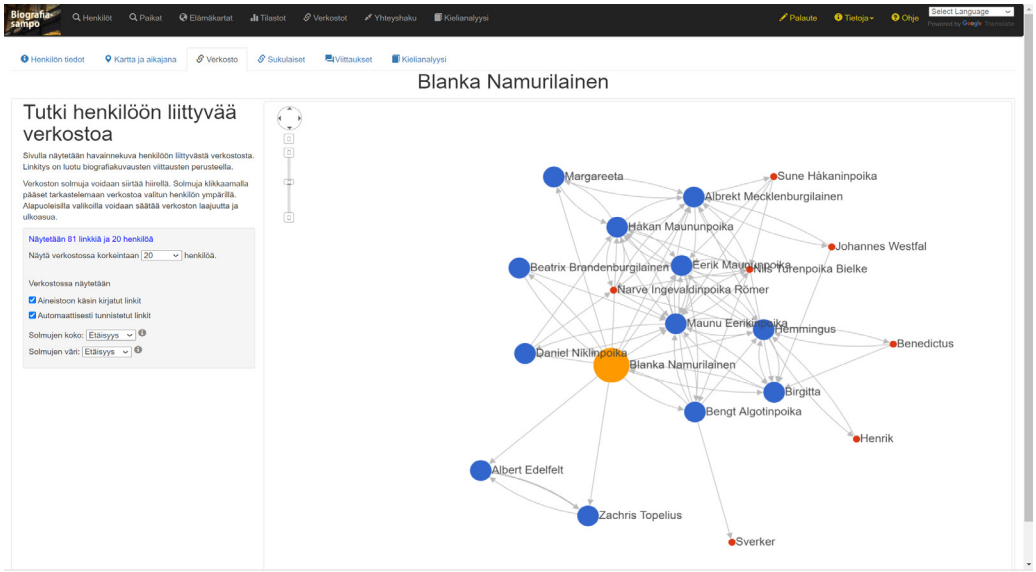
[53] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/sualt/; Hyvönen et al. 2021.

[54] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/lawlod/; Hyvönen et al. 2020.

[55] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/semparl/; Hyvönen et al. 2021.

[56] Ks. hankkeen kotisivu: https://seco.cs.aalto.fi/projects/rrl/; Tuominen et al. 2018.

[57] Warren 2018; Bhreathnach et al. 2019.



Kuva 3. Kuningatar Blanka Namurilaisen (1318–1363) egosentrinen verkosto Biografiasammossa.

Huomattava on myös esimerkiksi se, että Biografiasammon verkostoanalyysien perustana ovat Biografiakeskuksen elämäkertojen väliset linkit, jotka ovat osin aineistojen kirjoittajien ja toimittajien luomia aineistoja kirjoitettaessa, osin tekstianalyysiin perustuvia. Tällöin linkki ei tarkoita sitä, että yhdistetyt henkilöt esimerkiksi olisivat välttämättä edes tienneet toisistaan. Esimerkiksi Ruotsin kuningatar Blanka Namurilaisen (1318−1363), Ruotsin ja Norjan kuninkaan Maunu Eerikinpojan (1316−1374) ranskalaisen puolison egosentrinen verkosto kuvassa 3 esittää paitsi hänen lähipiiriään myös hänen jälkimainettaan. Oman aikansa ruotsalaisen ylimystön lisäksi kuningatar Blankan verkoston keskeisiä linkkejä ovat viisisataa vuotta myöhemmin eläneet Zachris Topelius (1818−1898) ja Albert Edelfelt (1854−1905). Topeliuksen Lukemisia lapsille ja Edelfeltin maalaus tekivät Blankasta 1800-luvun Suomessa yhden tunnetuimmista ruotsalaisista kuningattarista.

Tällaisten satunnaisten yhteyksien esille nostaminen on kuitenkin yksi osa dataanalyysin viehätysvoimaa. Esimerkiksi yllättävä yhteys keihäänheittäjä Tapio Rautavaaran (1915−1979) ja runoilija Aale Tynnin (1913−1997) välille syntyy siksi, että molemmat voittivat olympiakultaa Lontoossa, Rautavaara keihäänheitossa ja Tynni lyriikassa, joka oli vielä tuolloin olympialaji. Yhteyksien luonteen tarkempaa selvittämistä varten voidaan käyttää hyväksi Biografiasammon valmiiksi esiin louhimia lauseita, joissa linkit esiintyvät ja selittyvät.

Järjestelmien käyttäjältä edellytetään erityisen tarkkaa lähdekriittistä ymmärrystä siitä, millaista dataa tietoaineisto ja sovellus oikeastaan sisältää, onko tieto missä määrin epätäsmällistä tai puutteellista ja millaisia oletuksia järjestelmässä käytettävät ontologiat ja menetelmät kenties tekevät. Esimerkiksi Biografiasammon faseteissa käytetyssä historiallisten paikkojen ontologiassa luovutetun Karjalan paikat eivät löydy Suomen alta, vaikka niiden avulla kuvatut tapahtumat yleensä liittyvä aikaan, jolloin alue vielä oli osa Suomea. Kaikille asioille ei ole olemassa suoraviivaisia ratkaisuja.

Sampo-järjestelmät eivät korvaa perinteistä primaarilähteiden lähilukua ja tutkimusta, mutta tarjoavat tutkijan työkalupakkiin uudenlaisia välineitä, jotka helpottavat laajojen aineistojen hakua, selailua ja analyysiä ja auttavat löytämään niistä kiinnostavia ilmiöitä tarkempaa tutkimista varten.

Kiitokset

Kirjoittaja haluaa kiittää kaikkia Semanttisen laskennan tutkimusryhmän jäseniä ja alumneja[58], sekä yhteistyökumppaneita ja rahoittajia, joiden aineistoihin ja yhteistyöhön tässä artikkelissa kuvatut tutkimushankkeet perustuvat.

[58] SeCo-ryhmän nykyiset jäsenet ja alumnit: https://seco.cs.aalto.fi/people/

Lisätietoa

Eero Hyvönen: Semanttinen web. Linkitetyn avoimen datan käsikirja. Gaudeamus, 2018.

Lisätietoa Sampo-portaaleista: http://seco.cs.aalto.fi/applications/sampo/

Videoita sammoista ja suomalaisesta linkitetyn datan infrastruktuurista:

Semantic Web and AI for Digital Humanities: https://vimeo.com/470313703

WarSampo: Finnish Second World War on the Semantic Web: https://vimeo.com/212249404

LetterSampo – Historical Letters of the Semantic Web: https://vimeo.com/461293952

AcademySampo – Finnish Academic People 1640–1899: https://vimeo.com/462993654

AcademySampo – Akateemiset henkilöt Suomessa 1640–1899. Visio ja sen toteutus. https://vimeo.com/508756030

BiographySampo - AI Reading Biographies for the Semantic Web: https://vimeo.com/328419960

Building a National Level Linked Open Data Infrastructure for Digital Humanities in Finland: https://vimeo.com/460086143

Kirjallisuutta

Tim Berners-Lee, Mark Fischetti: *Weaving the Web. The original design and ultimate destiny of the World Wide Web, by its inventor*. Barnes & Noble, 1999.

Tim Berners-Lee, James Hendler and Ora Lassila: The Semantic Web. *Scientific American*, May 1, 2001.

Úna Bhreathnach, Cathal Burke, Jeig Mag Fhinn, Gearoid O. Cleircín, and Brian O Raghallaigh: A quantitative analysis of biographicaldata from Ainm, the Irish-language biographical database. In: *Proceedings of the Third Conference on Bio-graphical Data in a Digital World (BD 2019)*, 2019.

Toby Burrows, Douglas Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Andrew Morrison, Kevin Page, Lynn Ransom, Emma Cawlfield Thomson, Jouni Tuominen, Athanasios Ve-

lios, and Hanno Wijsman: Mapping Manuscript Migrations Knowledge Graph: Data for Tracing the History and Provenance of Medieval and Renaissance Manuscripts. *Journal of Open Humanities Data*, vol. 6, pp. 3, June, 2020. https://doi.org/10.5334/johd.14

Tom Heath, Christian Bizer: *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan-Claypool, 2011. https://doi.org/10.2200/s00334ed-1v01y201102wbe001

Pascal Hitzler: A review of the semantic web field. *Commun. ACM*, Vol. 64, Nr. 1, 2021.

https://doi.org/10.1145/3397512

Eero Hyvönen (ed.): *Semantic Web Kick-Off in Finland - Vision, Technologies, Research*, and Applications. HIIT Publications 2002-01, Helsinki, 2002.
https://seco.cs.aalto.fi/publications/2002/hyvonen-semantic-web-kick-off-2002.pdf

Eero Hyvönen: *Semanttinen web. Linkitetyn avoimen datan käsikirja*. Gaudeamus, 2018, 271 ss.
https://kauppa.gaudeamus.fi/sivu/tuotehaku/?action=search&search=semanttinen+web

Eero Hyvönen: Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web – Interoperability, Usability, Applicability* 11(1), 2020a, pp. 187–193. https://doi.org/10.3233/sw-190386

Eero Hyvönen: Sampo Model and Semantic Portals for Digital Humanities on the Semantic Web. DHN 2020 Digital Humanities in the Nordic Countries. *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pp. 373–378, CEUR Workshop Proceedings, vol. 2612, Riga, Latvia, October, 2020b.
http://ceur-ws.org/Vol-2612/poster1.pdf

Eero Hyvönen, Samppa Saarela and Kim Viljanen: Application of Ontology Techniques to View-Based Semantic Search and Browsing. The Semantic Web: Research and Applications. *Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, 2004.
https://doi.org/10.1007/978-3-540-25956-5_7

Eero Hyvönen, Kim Viljanen, Jouni Tuominen and Katri Seppälä: Building a National Semantic Web Ontology and Ontology Service Infrastructure--The FinnONTO Approach. *Proceedings of the European Semantic Web Conference ESWC 2008*, Springer, Tenerife, Spain, June 1–5, 2008
https://doi.org/10.1007/978-3-540-68234-9_10

Eero Hyvönen, Kim Viljanen and Osma Suominen: HealthFinland - Finnish Health Information on the Semantic Web. *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Busan , Korea, Springer-Verlag, Nov, 2007.
https://doi.org/10.1007/978-3-540-76298-0_56

Eero Hyvönen, Jouni Tuominen, Miika Alonen and Eetu Mäkelä: Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. *The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014* (Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I. and Tordai, A. (eds.)), pp. 226–230, Springer-Verlag, May, 2014. https://doi.org/10.1007/978-3-319-11955-7_24

Eero Hyvönen, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen and Eetu Mäkelä: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. *The Semantic Web – Latest Advances and New Domains (ESWC 2016)* (Harald Sack, Eva Blomqvist, Mathieu d Aquin, Chiara Ghidini, Simone Paolo Ponzetto and Christoph Lange (eds.)), pp. 758–773, Springer-Verlag, May, 2016. https://link.springer.com/chapter/10.1007%2F978-3-319-34129-3_46

Eero Hyvönen, Petri Leskinen, Erkki Heino, Jouni Tuominen and Laura Sirola: Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. *Proceedings, Language, Technology and Knowledge (LDK 2017)*, pp. 113-119, Springer-Verlag, Galway, Ireland, June, 2017. https://link.springer.com/chapter/10.1007/978-3-319-59888-8_9

Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen and Kirsi Keravuori: BiographySampo - Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. *The Semantic Web. ESWC 2019* (Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J.G. Gray, Vanessa Lopez, Armin Haller and Karl Hammar (eds.)), pp. 574–589, Springer-Verlag, June, 2019. https://link.springer.com/chapter/10.1007%2F978-3-030-21348-0_37

Eero Hyvönen, Minna Tamper, Esko Ikkala, Sami Sarsa, Arttu Oksanen, Jouni Tuominen and Aki Hietanen: Publishing and Using Legislation and Case Law as Linked Open Data on the Semantic Web. *The Semantic Web: ESWC 2020 Satellite Events* (Harth, Andreas, Presutti, Valentina, Troncy, Raphaël, Acosta, Maribel, Polleres, Axel, Fernández, Javier D., Xavier Parreira, Josiane, Hartig, Olaf, Hose, Katja and Cochez, Michael (eds.)), Lecture Notes in Computer Science, vol. 12124, pp. 110–114, Springer-Verlag, 2020.
https://doi.org/10.1007/978-3-030-62327-2_19

Eero Hyvönen, Petri Leskinen, Heikki Rantala, Esko Ikkala and Jouni Tuominen: Akatemiasammon käyttö henkilöiden ja henkilöryhmien historiallisessa tutkimuksessa. Informaatiotutkimus, 2021. *Informaatiotutkimus*, ilmestyvä.

Eero Hyvönen, Heikki Rantala, Esko Ikkala, Mikko Koho, Jouni Tuominen, Babatunde Anafi, Suzie Thomas, Anna Wessman, Eljas Oksanen, Ville Rohiola, Jutta Kuitunen and Minna Ryyppö: Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework. *Antiquity, A Review of World Archaeology*, 2021. Accepted. https://seco.cs.aalto.fi/publications/2020/hyvonen-et-al-findsampo-2020.pdf

Esko Ikkala, Eero Hyvönen, Heikki Rantala and Mikko Koho: Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web – Interoperability, Usability, Applicability*, 2021. Accepted. http://www.semantic-web-journal.net/content/sampo-ui-full-stack-javascript-framework-developing-semantic-portal-user-interfaces-0

Esko Ikkala, Jouni Tuominen, Jaakko Raunamaa, Tiina Aalto, Terhi Ainiala, Helinä Uusitalo and Eero Hyvönen: NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research. *GeoHumanities'18: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities*, pp. 1–9, ACM, Seattle, WA, USA, November, 2018. https://dl.acm.org/doi/10.1145/3282933.3282936

Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor: Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62(8), 2019, pp. 36–43. https://doi.org/10.1145/3331166

Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen and Eero Hyvönen: WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability*, Vol. 12, no. 2, pp. 265–278, 2021. http://www.semantic-web-journal.net/content/warsampo-knowledge-graph-finland-second-world-war-linked-open-data-0

Mikko Koho, Toby Burrows, Eero Hyvönen, Esko Ikkala, Kevin Page, Lynn Ransom, Jouni Tuominen, Doug Emery, Mitch Fraas, Benjamin Heller, David Lewis, Andrew Morrison, Guillaume Porte, Emma Thomson, Athanasios Velios and Hanno Wijsman: Harmonizing and Publishing Heterogeneous Pre-Modern Manuscript Metadata as Linked Open Data. 2021. Submitted for review. https://seco.cs.aalto.fi/publications/2021/koho-et-al-mmm.pdf

Petri Leskinen and Eero Hyvönen: Linked Open Data Service about Historical Finnish Academic People in 1640–1899. DHN 2020 Digital Humanities in the Nordic Countries. *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pp. 284–292, CEUR Workshop Proceedings, Vol. 2612, Riga, Latvia, October, 2020. http://ceur-ws.org/Vol-2612/short14.pdf

Goki Miyakita, Petri Leskinen and Eero Hyvönen: Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators. *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018*, Nicosia, Cyprus, Springer-Verlag, November, 2018. https://seco.cs.aalto.fi/publications/2018/miyakita-et-al-prosopographer-demo-2018.pdf

Eetu Mäkelä, Eero Hyvönen and Tuukka Ruotsalo: How to deal with massively heterogeneous cultural heritage data – lessons learned in CultureSampo. *Semantic Web – Interoperability, Usability, Applicability* 3(1) 2012. https://seco.cs.aalto.fi/publications/2012/makela-hyvonen-culturesampo-2011.pdf

Eetu Mäkelä, Kaisa Hypén and Eero Hyvönen: BookSampo--Lessons Learned in Creating a Semantic Portal for Fiction Literature. *Proceedings of ISWC-2011*, Bonn, Germany, Springer-Verlag, 2011a. https://seco.cs.aalto.fi/publications/2011/makela-hypen-hyvonen-booksampo.pdf

Eetu Mäkelä, Aleksi Lindblad, Jari Väätäinen, Rami Alatalo, Osma Suominen and Eero Hyvönen: Discovering Places of Interest through Direct and Indirect Associations in Heterogeneous Sources – The TravelSampo System. *Terra Cognita 2011: Foundations, Technologies and Applications of the Geospatial Web*, CEUR Workshop Proceedings, Vol-798, 2011b. https://seco.cs.aalto.fi/publications/2011/makela-et-al-subi.pdf

Tuomas Palonen, Jouni Hyvönen, Joeli Takala ja Eero Hyvönen: Semanttinen Kalevala - Kulttuurisammon taontaa. *eLore* 16 (2), 2009. https://seco.cs.aalto.fi/publications/2009/palonen-et-al-semanttinen-kalevala-2009.pdf

Heikki Rantala, Esko Ikkala, Ilkka Jokipii, Mikko Koho, Jouni Tuominen and Eero Hyvönen: WarVictimSampo 1914–1922: A Semantic Portal and Linked Data Service for Digital Humanities Research on War History. *The Semantic Web: ESWC 2020 Satellite Events* (Harth, Andreas, Presutti, Valentina, Troncy, Raphaël, Acosta, Maribel, Polleres, Axel, Fernández, Javier D., Xavier Parreira, Josiane, Hartig, Olaf, Hose, Katja and Cochez, Michael (eds.)), Lecture Notes in Computer Science, vol. 12124, pp. 191–196, Springer-Verlag, 2020. https://link.springer.com/chapter/10.1007%2F978-3-030-62327-2_33

L. Rietveld and R. Hoekstra: The YASGUI family of SPARQL clients. *Semantic Web – Interoperability, Usability, Applicability* 8(3), pp. 373–383, 2017. https://doi.org/10.3233/SW-150197

Katri Seppälä ja Eero Hyvönen: Asiasanaston muuttaminen ontologiaksi. Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista. National Library, Plans, Reports, Guides, March, 2014. https://www.doria.fi/handle/10024/96825

John F. Sowa: *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

Steffen Staab, Rudi Studer: *Handbook on Ontologies (2. edition)*. Springer, 2010.

Osma Suominen, Sini Pessala, Jouni Tuominen, Mikko Lappalainen, Susanna Nykyri, Henri Ylikotila, Matias Frosterus and Eero Hyvönen: Deploying National Ontology Services: From ONKI to Finto. Proceedings of the Industry Track at the International Semantic Web Conference 2014, CEUR Workshop Proceedings, Vol 1383. http://www.ceur-ws.org/Vol-1383/paper6.pdf

Osma Suominen, Eero Hyvönen, Kim Viljanen and Eija Hukka: HealthFinland – a National Semantic Publishing Network and Portal for Health Information. *Journal of Web Semantics* 7(4), pp. 287–297, 2009. https://doi.org/10.1016/j.websem.2009.09.003

Daniel Tunkelang: *Faceted search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan-Claypool, 2009. https://doi.org/10.2200/S00190ED1V01Y200904ICR005

Jouni Tuominen, Eetu Mäkelä, Eero Hyvönen, Arno Bosse, Miranda Lewis and Howard Hotson: Reassembling the Republic of Letters - A Linked Data Approach. Proceedings of the *Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, pp. 76–88, CEUR Workshop Proceedings, vol. 2084, Helsinki, Finland, March, 2018. http://www.ceur-ws.org/Vol-2084/paper6.pdf

K. Verboven, M., Carlier, J. and Dumolyn: A short manual to the art of prosopography. In: *Prosopography approaches and applications. A handbook*, pp. 35–70. Unit for Prosopographical Research, Linacre College, 2007.

Christopher Warren: Historiography's two voices: Data infrastructure and history at scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics*, 2018. https://hcommons.org/deposits/item/hc:21833

Kim Viljanen, Jouni Tuominen and Eero Hyvönen: Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI. Proceedings of the 6th European Semantic Web Conference (ESWC 2009), pp. 781–795, Springer-Verlag, 2009. https://doi.org/10.1007/978-3-642-02121-3_57