KATSAUS

Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen

Eero Hyvönen
eero.hyvonen@aalto.fi
https://orcid.org/0000-0003-1695-5840

Petri Leskinen
petri.leskinen@aalto.fi
https://orcid.org/0000-0003-2327-6942

Heikki Rantala
heikki.rantala@aalto.fi
https://orcid.org/0000-0002-4716-6564

Esko Ikkala
esko.ikkala@aalto.fi
https://orcid.org/0000-0002-9571-7260

Jouni Tuominen
jouni.tuominen@aalto.fi
https://orcid.org/0000-0003-4789-5676

Semanttisen laskennan tutkimusryhmä (SeCo) ***http://seco.cs.aalto.fi***
Aalto-yliopisto, tietotekniikan laitos ja
Helsingin yliopisto, Digitaalisten ihmistieteiden keskus (HELDIG)

AcademySampo – Academic people in Finland 1640–1899 is a portal and a Linked Open Data service on the Semantic Web. AcademySampo contains richly interlinked open data about all people that have got academic education in Finland in 1640–1899. The system is targeted to researchers and the general public for biographical and prosopographical research. This review gives an overview on how AcademySampo can be utilized in practise with its novel digital humanities tools included in the portal and by using the data service via APIs.

Asiasanat: semanttinen web, linkitetty data, portaalit (tietotekniikka), datapalvelu, elämäkerrat, prosopografia

Artikkeli on lisensoitu Creative Commons Nimeä-EiKaupallinen-JaaSamoin 4.0 Kansainvälinen -lisenssillä

Pysyvä osoite: https://doi.org/10.23978/inf.102656

Johdanto

Akatemiasampo[1] (Leskinen ja Hyvönen, 2020) koostuu kahdesta osasta,

1. Akatemiasampo-portaalista[2] ja
2. linkitetyn avoimen datan palvelusta[3], joka on julkaistu Linked Data Finland -alustalla.

Akatemiasampo-portaali tarjoaa käyttäjälleen älykkäät haku- ja selailutoiminnot, joihin on saumattomasti integroitu joukko data-analyyttisiä työkaluja ja visualisointeja henkilöiden ja henkilöryhmien prosopografista (Verboven et al., 2007) tutkimista ja analysointia varten tilastoina, verkostoina, erilaisina graafeina ja kartoilla (Leskinen et al., 2018). Portaalin käyttö ei edellytä erityistä tietoteknistä osaamista. Akatemiasammon datapalvelun avoimet rajapinnat ja SPARQL-palvelupiste puolestaan tarjoavat helppokäyttöisen mahdollisuuden uusien data-analyysien toteuttamiseen digitaalisten ihmistieteiden tutkijoille, joilla on jonkin verran kokemusta semanttisen webin SPARQL-kyselykielestä ja ohjelmoinnista. Tässä voidaan käyttää esimerkiksi YASGUI-käyttöliittymää (Rietveld and Hoekstra, 2017) tai Jupyter- ja Google Colab -dokumentteja ja Python-skriptejä. Akatemiasampo perustuu Sampo-malliin (Hyvönen, 2020b) ja sen portaaliosa on toteutettu Sampo-UI-ohjelmointikehyksen avulla (Ikkala et al., 2021) esimerkkinä datapalvelun hyödyntämismahdollisuuksista sovellusten kehittämisessä.

Akatemiasammon data muodostaa laajan semanttisen webin tietämysgraafin (knowledge graph), joka on tuotettu algoritmisesti Turun akatemian ja Helsingin yliopiston digitoiduista Ylioppilasmatrikkeleista 1640–1852 ja 1853–1899[4] louhimalla tietoa teksteistä ja tietokannan rakenteista. Dataa on lisäksi rikastettu linkittämällä sitä sisäisesti ja ulkoisesti muihin aineistoihin sekä tekoälyperustaisella päättelyllä. Järjestelmän ydinaineistona olevat ylioppilasmatrikkelit sisältävät tietoa kaikista tiedossa olevista Suomessa akateemisen koulutuksen saaneista henkilöistä 1640–1899, sillä tuohon aikaan ei Suomessa ollut muita yliopistoja.

1 Akatemiasampo-hankkeeen kotisivu: https://seco.cs.aalto.fi/projects/yo-matrikkelit/

2 Portaali avattiin Runebergin päivänä 5.2.2021 osoitteessa https://akatemiasampo.fi

3 Akatemiasampo-datapalvelu: https://www.ldf.fi/dataset/yoma

4 Helsingin yliopiston Ylioppilasmatrikkelit: https://www.helsinki.fi/fi/yliopisto/ylioppilasmatrikkelit-1640-1907

Kuvassa 1.1 on esitetty esimerkkinä ylioppilas Johan Ludvig Runebergin (1804–1877) henkilötiedot Helsingin yliopiston vuonna 2005 julkaisemassa Ylioppilasmatrikkelit-palvelussa[5]. Ylioppilaiden matrikkelikuvaukset sisältävät runsaasti myöhemmin lisättyä tietoa heidän urastaan, sukulaisistaan ja elämästään opintojen jälkeen sekä viitetietoja kirjallisuuteen. Turun akatemian alkuperäinen matrikkeli tuhoutui Turun palossa 1827, mutta se rekonstruoitiin 1800-luvun lopulla Vilhelm Laguksen toimesta. Tätä työtä on täydennetty 1900-luvulla eri lähteistä, ja lopulta tiedot toimitettiin noin 10 henkilötyövuoden urakalla nykyisiksi, verkossa oleviksi Ylioppilasmatrikkeleiksi Yrjö Kotivuoren ja Veli-Matti Aution toimesta[6].

5 Ylioppilasmatrikkelien verkkopalvelu: https://www.helsinki.fi/fi/yliopisto/yliopiston-matrikkelit

6 Ylioppilasmatrikkelien seloste: https://ylioppilasmatrikkeli.helsinki.fi/esipuhe/aluksi.php

Ylioppilasmatrikkeli 1640–1852

Etusivu > Henkilötiedot

Sisältö:

- Matrikkelin etusivu
- Esipuhe
- Ylioppilaat aikajärjestyksessä
- Ohjattu haku
- Hakulomake
- Muistiin tallennetut tietueet
- Hakuohje
- Viittausohje
- Viimeisimmät päivitykset

Muut matrikkelit:

- Ylioppilasmatrikkeli 1853–1899

Henkilötiedot

2.10.1822 **Johan Ludvig Runeberg** 13687. * Pietarsaaressa 5.2.1804. Vht: pietarsaarelainen merikapteeni *Lorentz Ulrik Runeberg* 10660 (yo 1791, † 1828) ja *Anna Maria Malm*. Oulun triviaalikoulun oppilas 6.3.1813 (cl. I) – 1814 (avg.). Vaasan triviaalikoulun oppilas 30.1.1815 – 24.7.1822. Pääsykuulustelu 30.9.1822. Ylioppilas Turussa 2.10.1822. Pohjalaisen osakunnan jäsen 5.10.1822 *[1822] Johannes Ludovicus Runeberg. die 5 Octobris. Natus die V Februarii anno MDCCCIV. | Promotus Philos. Doctor anno 1827. Eloquentiae Docens 1829. Lector Gymnasii Borgoënsis 1837. | De stella polari eques 1844. Professor honorarius nominatus 1845. Utgaf "Dikter" 1829. "Elgskyttarne" s.å. Dikter 2:dra h. 1832. "Hanna" 1836. "Nadeschda" 1841. "Dikter" 3:dje h. 1843. "Kung Fjalar" 1844. "Fänrik Ståls Sägner" I 1848.* Respondentti 17.6.1826 pro exercitio, pr. *Anders Johan Lagus* 10906. Stipendiaattiteesi 6.10.1826, pr. *Anders Johan Lagus* 10906. FK 13.6.1827. Respondentti 23.6.1827 pro gradu, pr. *Carl Reinhold Sahlberg* 10919. FM 10.7.1827. Preeses 16.6.1830 pro venia docendi. Preeses 16.6.1830 pro venia docendi. Preeses 30.11.1833 pro munere. Vihilty papiksi Porvoon hiippakunnassa 19.12.1838. TT h.c. 28.5.1857 (keis. maj.:n nimittämänä 14.5.1857). — Aleksanterin yliopiston konsistorin amanuenssi 1830–34, virkavapaa 1831 ja 1832–33, kaunopuheisuuden dosentti 1830, ero 1836. Samalla Helsingin yksityisyliseon opettaja 1831–36. Sanomalehtimies vuodesta 1832. Porvoon lukion Rooman kirjallisuuden lehtori 1837, Kreikan kirjallisuuden lehtori 1842, ero 1857. Samalla Porvoon yksityisen esilukion opettaja 1837–42. Professorin arvonimi 1844. Halvaantui 1863. Saavuttanut Suomen kansallisrunoilijan aseman. † Porvoossa 6.5.1877.

Pso: 1831 *Fredrika Charlotta Tengström* († 1879).

Poika: Kuopion lukion lehtori, FMM *Ludvig Michael Runeberg* 17281 (yo 1854, † 1902).

Poika: piirilääkäri Helsingissä, LKT *Lorenzo Runeberg* 17282 (yo 1854, † 1919).

Poika: kuvanveistäjä, FT *Walter Magnus Runeberg* 17759 (yo 1858, † 1920).

Poika: sisätautiopin professori, valtioneuvos, LKT *Johan Wilhelm Runeberg* 17983 (yo 1860, † 1918).

Poika: LL *Fredrik Karl Runeberg* 18818 (yo 1868, † 1884).

Lanko: kokkolalainen asianajaja, varatuomari, FM *Karl Tengström* 12646 (yo 1813, † 1853).

Yksityistod. saaja 16.6.1828: *Roland Konstantin Tillman* 14531; *Anders Oskar Roos* 14588; *Karl Gustaf Favorin* 14671; *Karl August Adrian Hagelstam* 14673; *Karl Adolf Antell* 14715; *Karl Otto Ramstedt* 14717; *Josef Vilhelm Fontell* 14808; *Karl Emil af Enehjelm* 14810; *Georg Edvard af Enehjelm* 14811; *Hans Johan Berg* 14812; *Karl Fredrik Duckman* 14853; *Oskar Vilhelm Forsman* 14856; *Rudolf Israel Holsti* 14937; *Karl August Londicer* 14938; *Johan August von Essen* 14941; *Henrik Schwartzberg* 14947; *Bror Niklas Sandman* 14948; *Karl Johan Krabbe* 14958; *Magnus Arvid Krabbe* 14959; *Mauritz Hasselblatt* 14985; *Ernst Odert Reuter* 15046; *Johan Justus Staudinger* 15052; *Zachris Topelius* 15088; *Johan Edvard Thodén* 15095.

Viittauksia: HYK ms., Pohj. osak. matr. #2001; HYKA, Album 1817–65 s. 78; HYKA TAA Bb, Tutkintoja koskevat luettelot 1796–1828; HYKA TAA Ef, Ylioppilaiden kirjoittautumisasiakirjat v. 1822; KA Ansioluettelokokoelma; KA Colliianderin kokoelma, Suomen kirkon paimenmuisto #2920. — V. Lagus, Studentmatrikel II (1892–95) s. 644 (CLXXIV); A. Dahl, Kort historik över Uleåborgs pedagogi och dess efterföljare trivialskolan... (1926) #1228, s. 36; O. Wanne, Liber scholae Wasensis 1722–1830. SSJ 18 (1947) #2153; E. Kojonen, Pohjalaisen osakunnan nimikirja V 1809–1827 (1957) #2001. — M. Akiander, Skolverket. BNF 9 (1866) s. 202, 207; A. Bergholm, Sukukirja II (1901) s. 1125 (Runeberg Taulu 12); G. Heinricius, Skildringar från Åbo akademi 1808–1828. SSLS 101 (1911) s. 44; H. Söderstéen, Bidrag till Johan Ludvig Runebergs ättartal. SSV 2 (1918) s. 1–9; T. Carpelan ja L. Tudeer, Helsingin yliopisto. Opettajat ja virkamiehet vuodesta 1828. II (1925) s. 817; J. Vallinkoski, Turun akatemian väitöskirjat 1642–1828. HYKJ 30 (1962–66) #2133R, 2134R, 3422R; Sursillin suku (täyd. ja toim. E. Kojonen, 1971) #2052; Suomen kirjailijat 1809–1916. SKST 570 (toim. M. Hirvonen, 1993) s. 640; Suomen kansallisbiografia 8 (2006) s. 407.

Päivitetty 15.7.2013.

Lähdeviite

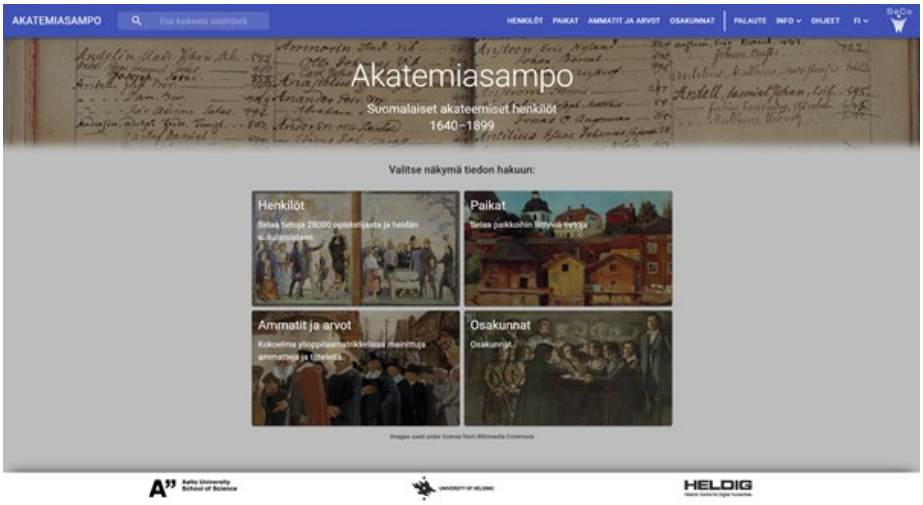
Yrjö Kotivuori, *Ylioppilasmatrikkeli 1640–1852: Johan Ludvig Runeberg*. Verkkojulkaisu 2005 <https://ylioppilasmatrikkeli.helsinki.fi/henkilo.php?id=13687>. Luettu 21.1.2021.

Ks. myös viittausohje.

Näytä / piilota lisätiedot

Kuva 1.1. Johan Ludvig Runebergin (1804–1877) henkilötiedot Helsingin yliopiston alkuperäisessä Ylioppilasmatrikkelit 1640–1852 -verkkopalvelussa.

Akatemiasammon data on luotu rakenteistamalla Ylioppilasmatrikkelin 1640–1852 n. 9 500 henkilön ja matrikkelin 1853–1899 n. 18 450 henkilön tekstikuvaukset linkitetyksi dataksi (Linked Data)[7] (Heath and Bizer, 2011). Tämä on tehty tunnistamalla säännöllisten lausekkeiden avulla matrikkelitekstistä henkilöiden biografiset perustiedot, matrikkelin ulkopuoliset sukulaiset (n. 47 000 kpl), henkilöiden välisiä sukulaissuhteita (n. 130 000 kpl), historiallisia paikkoja (n. 3 000 paikkaa), ammatteja ja arvoja (n. 10 000 kpl) ja akateemisia oppilas-opettaja-suhteita (n. 4 000 kpl). Kokonaisuuden ”semanttisena liimana” ovat teksteistä tunnistetut, henkilöiden ammatilliseen ja perhe-elämään liittyvät tapahtumat (n. 175 000 kpl), jotka linkittävät yhteen eri rooleissa osallistuvia henkilöitä ja organisaatioita, paikkoja ja aikoja kulttuuriperintöalan CIDOC CRM -ontologian[8] ja -ISO-standardin mukaisesti. Dataa on lisäksi rikastettu linkityksillä ulkoisiin tietoaineistoihin, kuten Biografiasammon/Kansallisbiografian elämäkertoihin (Hyvönen et al., 2019) ja Wikidataan[9], sekä päättelemällä henkilöiden välisiä sukulaissuhteita. Data on julkaistu ja se on käytettävissä linkitetyn avoimen datan palveluna Linked Data Finland -alustalla[10] (Hyvönen et al., 2014), josta löytyy datan ohella siihen liittyvää dokumentaatiota ja työkaluja datan hyödyntämistä varten uusissa tutkimuksissa ja sovelluksissa.



Kuva 1.2: Akatemiasampo.fi-portaalin neljä sovellusnäkymää ovat valittavissa pääsivulla.

7 W3C Linked Data: `https://www.w3.org/standards/semanticweb/data`
8 CIDOC CRM: `http://cidoc-crm.org`
9 Wikidata: `https://wikidata.org`
10 Linked Data Finland -alusta: `http://ldf.fi`

Akatemiasampo.fi-portaali tarjoaa Sampo-UI-ohjelmointikehyksen periaatteiden mukaisesti erilaisia sovellusnäkymiä Ylioppilasmatrikkeleiden aineistoihin. Pääsivulla esitellään kuvan 1.2. mukaisesti neljä sovellusnäkymää: Henkilöt, Paikat, Ammatit ja arvot sekä Osakunnat. Vastaavaa kuvaketta klikkaamalla avautuvat näkymät henkilöiden, paikkojen, ammattien ja arvojen sekä osakuntien hakemista ja tutkimista varten joko yksittäin tai ryhminä digitaalisten ihmistieteiden keinoin.

Akatemiasampo kuuluu Aalto-yliopiston ja Helsingin yliopiston (HELDIG) Semanttisen laskennan tutkimusryhmässä (SeCo) kehitettyyn Sampo-sarjaan[11] ja digitaalisten ihmistieteiden linkitetyn avoimen datan infrastruktuuriin Suomessa (LODI4DH)[12], joka on osa Suomen Akatemian tiekartalla olevaa digitaalisten ihmistieteiden tietoinfrastruktuuria[13]. Sampo-järjestelmät perustuvat semanttisen webin teknologioihin ja linkitettyyn avoimeen dataan (Hyvönen, 2018).

Alla kuvataan ensin lyhyesti Akatemiasampo.fi-portaalin sovellusnäkymien tarjoamia haku- ja selailutoimintoja sekä näihin saumattomasti integroituja data-analyysin työkaluja ja visualisointeja. Kaikki portaalin toiminnallisuus on kehitetty Akatemiasammon avoimen datapalvelun varaan *vain* SPARQL-rajapintaa[14] käyttäen. Tämä tarkoittaa sitä, että kenellä tahansa on sama mahdollisuus kehittää vastaavia uusia sovelluksia tai tehdä digitaalisten ihmistieteiden tutkimusta kun Akatemiasampo.fi-demonstraattorin takojilla.

Sovellusnäkymät tutkimuskäytössä

Tässä luvussa esitellään lyhyesti Akatemiasampo.fi-portaalin eri sovellusnäkymien käyttömahdollisuuksia.

Henkilöt-näkymä

Portaalin tärkein sovellusnäkymä on Henkilöt, jossa voi hakea ontologiaperustaisella fasettihaulla tiettyä yksittäistä henkilöä tai henkilöryhmää prosopografista analyysiä varten.

Henkilöt-kuvaketta pääsivulla klikkaamalla avautuu kuvan 2.1 (fasetti) hakunäkymä Akatemiasammon matrikkelihenkilöihin. Hakufasetit näkyvät

11 Tietoa Sampo-portaaleista: https://seco.cs.aalto.fi/applications/sampo/

12 Linked Open Data Infrastructure for Digital Humanities: https://seco.cs.aalto.fi/projects/lodi4dh/

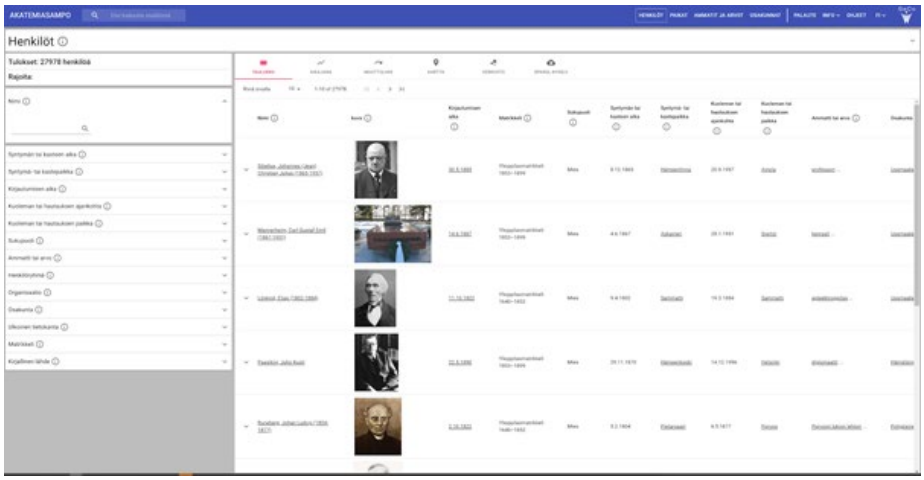
13 FIN-CLARIAH: https://www.aalto.fi/fi/uutiset/digitaalisten-ihmistieteiden-tietoinfrastruktuurit-mukaan-suomen-akatemian-uudelle

14 SPARQL-kyselykieli: https://www.w3.org/TR/sparql11-query/

vasemmalla ja haun tulos, alussa kaikki henkilöt, oikealla taulukkomuodossa. Tulosjoukon henkilöt listataan oletusarvoisesti heidän verkostoitumisasteensa perusteella, jolloin kärkipaikkoja pitävät kuvan 2.1 mukaisesti Sibeliuksen, Mannerheimin ja Lönnrotin kaltaiset ylioppilaat. Tulosjoukon rivit vastaavat henkilöitä ja sarakkeet fasettien arvoja kunkin henkilön osalta. Tulokset voidaan aina järjestää uudelleen ylärivissä olevien, fasetteja vastaavien sarakkeiden painikkeiden avulla esimerkiksi syntymäajan tai kuolinpaikan mukaan.

Fasettihaussa tulosjoukkoa rajataan tekemällä valintoja hierarkkisista faseteista. Kuvassa 2.2 on esimerkkinä avattu kaksi fasettia, joista ylempi luettelee henkilöihin liittyviä organisaatioita, joista on valittu ”Hornin rykmentti”, ja alempi osakuntia, joista Hornin rykmenttiin liittyviä henkilöitä ylipäätään löytyy. Näistä on valittu ”Pohjalainen osakunta”, jolloin kahdella fasettivalinnalla on löydetty kaikki neljä Hornin rykmenttiin ja Pohjalaisen osakuntaan liittyvät henkilöt. Huomattava on, että todellisuudessa ylioppilaissa voi toki olla muitakin tällaisia henkilöitä. Heitä haku ei löydä, jos tästä Akatemiasammon aineistoissa ei jostain syystä ole mainintaa (historiallinen tietohan on usein puutteellista) tai jos Akatemiasammon algoritmit eivät ole kaikkea tietoa jostain syystä onnistuneet tekstistä erottamaan. Akatemiasammon haut ja kaikki data-analyysit rajoittuvat luonnollisesti vain siihen dataan, joka sillä on käytettävissä.

Myös perinteinen tekstihaku on mahdollista erillisen tekstifasetin kautta ja se kohdistuu henkilöiden kaikkiin tiedossa oleviin nimiin. Myös tekstihaku kaikkiin datapalvelussa oleviin kohteisiin samalla kertaa (henkilöt, paikat, osakunnat jne.) onnistuu portaalin yläpalkissa olevan erillisen hakukentän avulla. Tekstihaussa on mahdollista käyttää säännöllisiä lausekkeita ja korvausmerkkejä ? (mikä tahansa merkki), + (ainakin yksi merkki) ja * (mikä tahansa määrä merkkejä).



Kuva 2.1: Henkilöiden ja henkilöryhmien sovellusnäkymä: hakufasetit vasemmalla ja tulos oikealla.

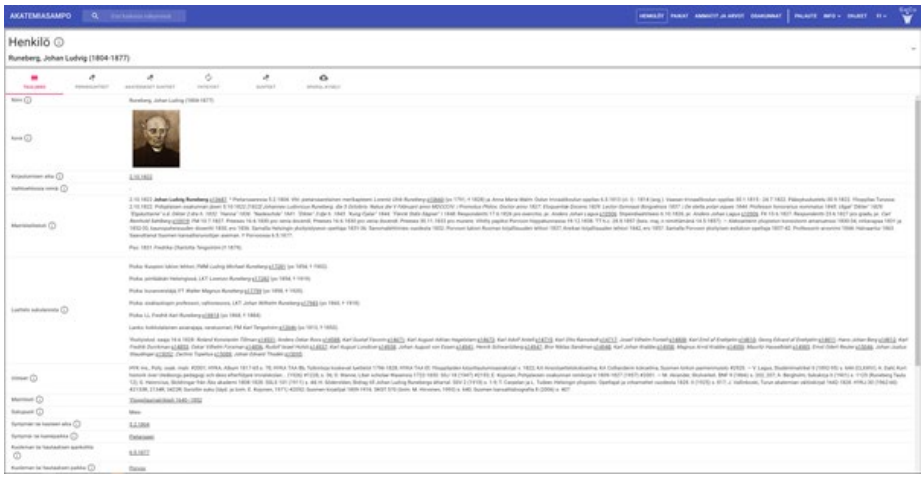
Kuva 2.2: Henkilöiden haku fasettivalintojen avulla; tässä etsitään Hornin rykmenttiin ja Pohjalaiseen osakuntaan liittyviä henkilöitä.

Akatemiasammon jokaiselle henkilölle on muodostettu oma ”kotisivu”, johon on kerätty paitsi hänen biografiset tietonsa perinteiseen tapaan, myös linkitetty yhteen henkilöön liittyvää tietoa ja analyysejä koko aineiston perusteella sekä matrikkelien ulkopuolista täydentävää dataa. Kotisivuille pääsee klikkaamalla hakutuloksessa näkyvää henkilöä. Esimerkiksi kuvassa 2.3 on Johan Ludvig Runebergin kotisivu. Kotisivut esitetään taulukkona, jossa vasen sarake luettelee henkilöön liittyvät ominaisuudet, esimerkiksi yliopistoon kirjautumisaika, ammatti/arvo ja osakunta. Oikealla ovat ominaisuuksien arvot, esimerkiksi yliopistoon kirjautumisaika 10.2.1822 ja

ammatit/arvot ”Porvoon lukion lehtori”, ”dosentti” ja ”kirjailija”, sekä linkkejä lisätietoon.

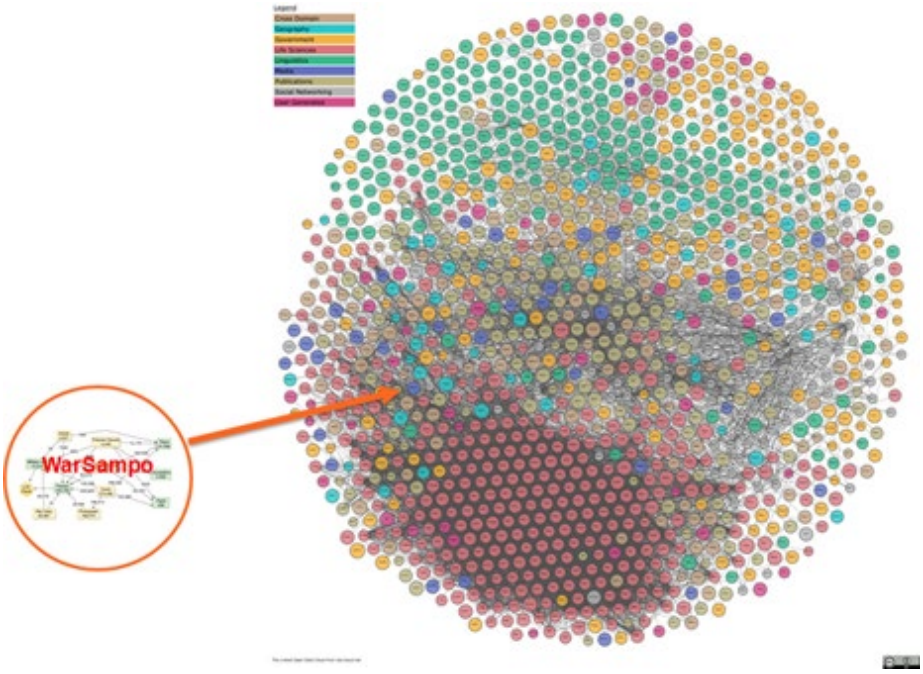
Akatemiasammon henkilöt on linkitetty Wikidataan ja Biografiasampoon. Wikidatan-linkkien perusteella on edelleen haettu henkilöiden Wikipedia-osoitteet ja Wikimedia Commons -aineistosta mahdolliset valokuvat henkilöistä. Vastaavasti linkit Kansallisbiografiaan on haettu Biografiasammon tietokannasta. Linkit Turun Akatemian väitöskirjat -materiaaliin Kansallisarkiston Doria-julkaisuarkistossa olivat valmiina ylioppilasmatrikkelien lähdeaineistossa.

Kuvan 2.4. sivulla on esimerkiksi linkki laajempaan elämäkertaan ja lisätietoihin Runebergistä Biografiasammossa. Se taas perustuu Suomalaisen Kirjallisuuden Seuran Kansallisbiografiaan ja muihin elämäkertoihin ja näihin edelleen linkitettyihin muihin aineistoihin suomalaisen linkitetyn datan infrastruktuurissa (LODI4DH), esimerkiksi Sotasampoon. Näin eri sammoista ja niiden perustana olevista ontologioista ja datasta on vähitellen muodostumassa yhä laajempi kansallinen linkitetyn avoimen datan pilvi (Linked Open Data Cloud), eräänlainen ”Samposampo”, joka yhdistyy myös kansainväliseen Linked Open Data (LOD) -pilveen[15]. Kuva 2.4 esittää kansainvälisen LOD-pilven aineistoja ja siinä olevia, keskenään linkitettyjä datajoukkoja. Esimerkiksi Sotasammon[16] tietämysverkko, n. 14 miljoona tietojen välistä yhteyttä, on yksi pallukka tässä datajoukkojen pallomeressä, jonka ytimessä on Wikipedioiden datasta louhittu linkitetty data.

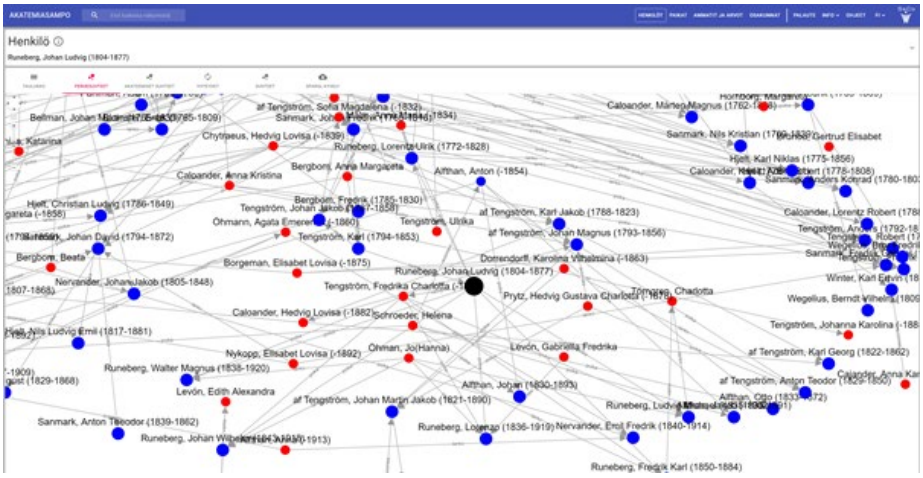


Kuva 2.3: Johan Ludvig Runebergin (1804–1877) kotisivu Akatemiasammossa.

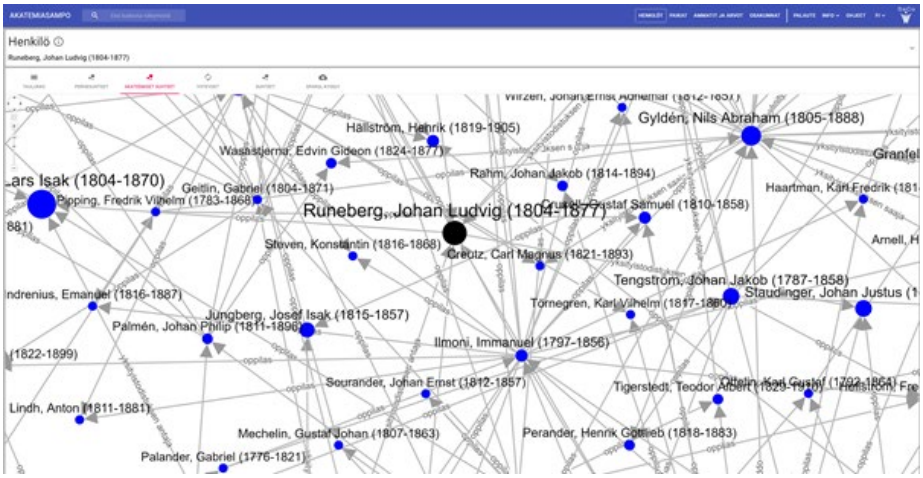
15 Linked Open Data Cloud: `https://www.lod-cloud.net`
16 Sotasampo-hankkeen kotisivu: `https://seco.cs.aalto.fi/projects/sotasampo/`



Kuva 2.4: Kansainvälinen Linked Open Data -pilvi, johon sisältyy myös mm. Sotasammon oma tietämysverkko. Akatemiasampo linkittyy mm. Wiki-dataan ja sitä kautta moniin muihinkin datajoukkoihin kuten Sotasampoon.



Kuva 2.5: Johan Ludvig Runebergin sukulaisverkostoa Akatemiasammon visualisoimana.



Kuva 2.6: Osa Johan Ludvig Runebergin akateemista verkostoa Akatemiasammossa.

Yksittäisen henkilön data-analyysi ja visualisoinnit

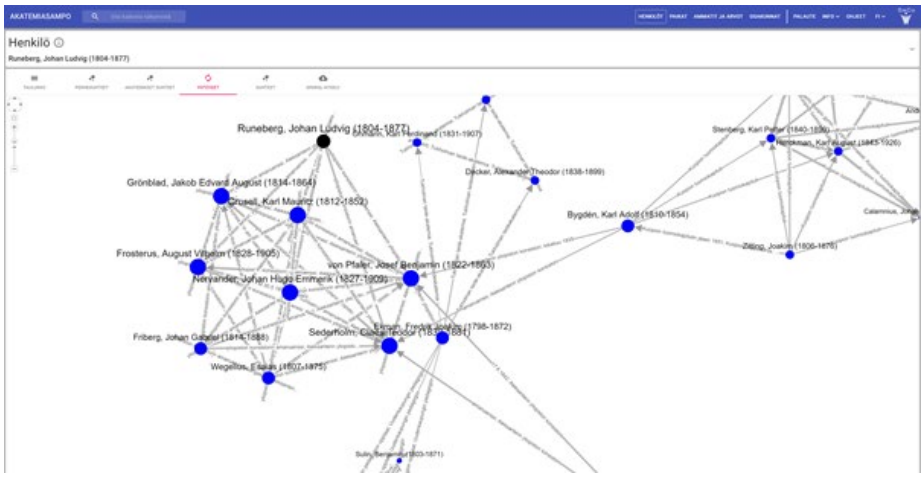
Sampo-portaalien yhtenä innovaationa on tarjota käyttäjälle älykkäiden haku- ja selailutoimintojen ohella data-analyyttisiä työkaluja ja visualisointeja sisältöjen tarkempaa tutkimista ja tietämyksen muodostamista (knowledge discovery) varten (Hyvönen, 2020a). Työkalut voidaan valita hakunäkymien yläreunassa olevilta välilehdiltä, jolloin välinettä sovelletaan faseteilla rajattuun joukkoon kulloisenkin näkymän hakuobjekteja. Myös hakuobjektien kotisivuilta löytyy joukko data-analyyttisiä työkaluja välilehdillä kyseisen yksittäisen objektin tutkimista varten. Esimerkiksi henkilöiden kotisivuilla valittavana on paitsi oletusarvoinen taulukkonäkymä (kuva 2.3) (välilehti TAULUKKO) myös mahdollisuus tutkia henkilön sukulaissuhteita, akateemisia yhteyksiä toisiin ylioppilaisiin, suhteita toisiin henkilöihin yhteisten työorganisaatioiden kautta tai henkilöön liittyvää tapahtumien verkostoa.

Kuvassa 2.5 käyttäjä on valinnut Johan Ludvig Runebergin kotisivulla PERHESUHTEET-välilehden, joka visualisoi hänen sukulaistensa verkostoa. Sen data on louhittu Akatemiasammossa eri yhteyksissä mainituista henkilöistä, ja sukulaissuhteita on lisäksi rikastettu päättelemällä. Verkostosta selviää esimerkiksi se, että Runebergin pojan, Lorenzo Runebergin (1836–1919) vaimo oli Gabriella Fredrika Levón.

Kuvassa 2.6 taas näkyy Johan Ludvig Runebergin akateemista verkostoa Akatemiasammon AKATEEMISET SUHTEET -välilehden visuali-

sointina, jossa suunnatut kaaret kuvaavat opettaja–oppilas-suhteita ja yksityistodistusten antajia ja saajia.

Henkilöiden välisiä verkostoja voidaan muodostaa heitä eri tavoin yhdistävien tekijöiden kautta. Akatemiasammon henkilöiden kotisivulta löytyy esimerkiksi myös välilehti, joka visualisoi henkilön yhteyksiä toisiin henkilöihin, sillä perusteella, ovatko he olleet saman organisaation jäseniä. Kuvassa 2.7 on esimerkkinä tämän kriteerin avulla muodostuvia klustereita Johan Ludvig Runebergin kotisivulla.



Kuva 2.7: Runebergin verkostoon liittyviä henkilöiden klustereita YHTEYDET-välilehdellä.

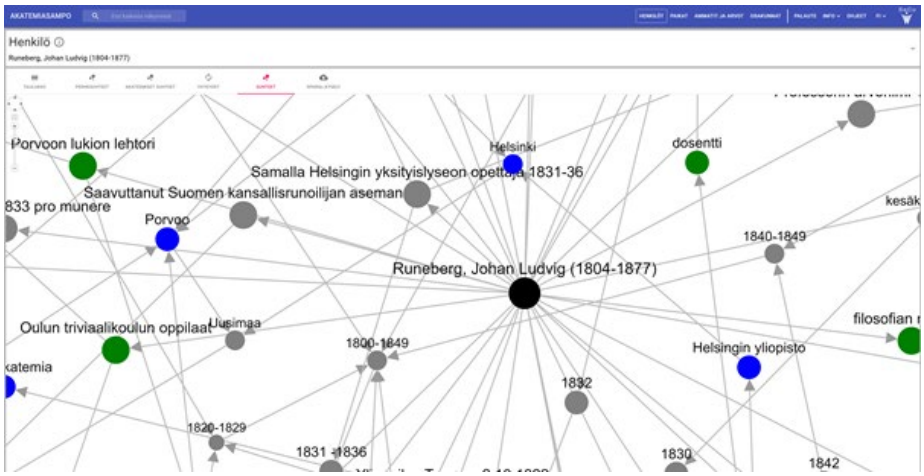
Myös matrikkelidatasta louhittujen tapahtumien kautta on muodostettu oma verkostonsa erilliselle SUHTEET-välilehdelle. Siitä selviää esimerkiksi Runebergin tapauksessa se, että hän oli Oulun triviaalikoulun oppilas 1813–1814 ja valmistui maisteriksi promootiossa 10.7.1827 (kuva 2.8).

Henkilöryhmien prosopografinen tutkiminen

Valintoja faseteista tekemällä voi myös rajata henkilöryhmiä ja tutkia niitä. Muutamalla klikkauksella voidaan esimerkiksi hakea yliopistoon 1700-luvulla kirjautuneet ylioppilaat (kirjautumisajan fasetti), joista tuli professoreita (ammatit ja arvot -fasetti), ja jotka ovat olleet Pohjalaisen osakunnan jäseniä (osakunnat-fasetti). Tällaisia henkilöitä, kuten Suomen historian isäksi mainittu Henrik Gabriel Porthan (1739–1804) ja Carl von Linnén oppilas, professori Pehr Kalm (1716–1779), löytyy 45 kpl. Hakutulos perustuu Akatemiasammon tiedossa olevaan dataan, joka voi olla epätäydellistä, ja käyttäjän on myös

aina otettava huomioon virheiden ja puutteiden mahdollisuus erityisesti vanhempaan Turun akatemian ajan aineistoon liittyen. Nämä voivat johtua alkuperäisestä datasta tai Akatemiasammon automaattisen käsitteiden tunnistamisen ja linkityksen yhteydessä tapahtuneista algoritmien tekemistä virheistä.

Fasettivalinnoilla löydettyjen henkilöryhmien (TAULUKKO-välilehti) prosopografiseen analyysiin on tarjolla omia työkaluja Henkilöt-näkymän välilehdillä AIKAJANA, MUUTTOLIIKE, KARTTA ja VERKOSTO.



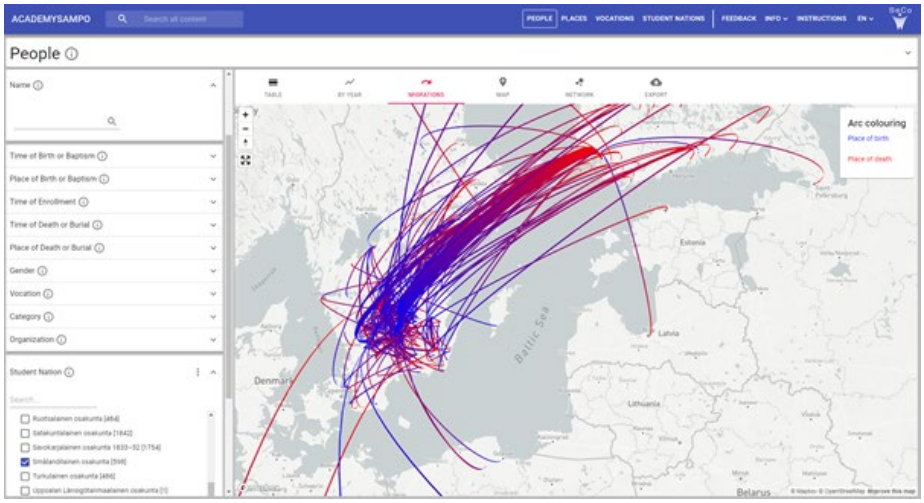
Kuva 2.8: Osa J. L. Runebergin elämään liittyvien tapahtumien verkostosta Akatemiasammossa.

Kuva 2.9 esittää Akatemiasammon naisylioppilaiden vuosittaiset syntymät, kirjautumiset yliopistoon ja kuolemat AIKAJANA-välilehdellä.

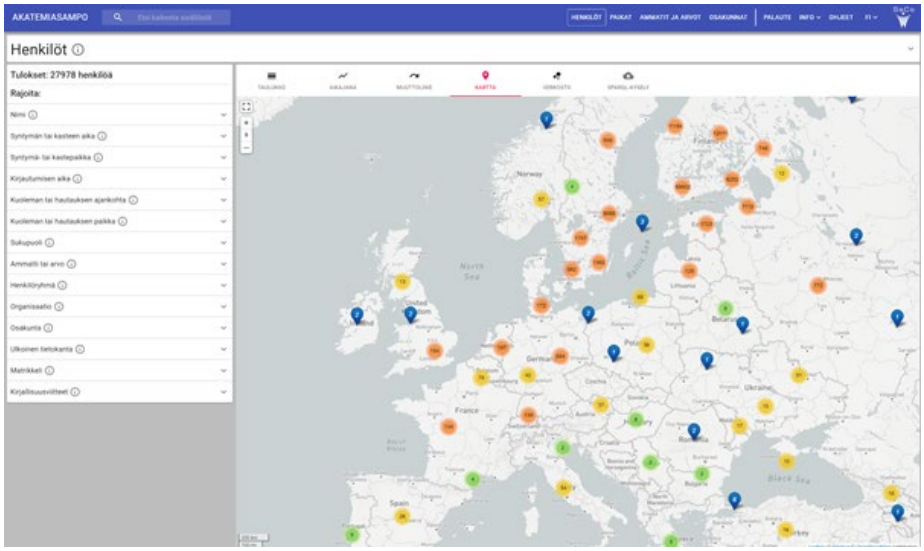
Kuvassa 2.10 on visualisoitu Smålandilaisen osakunnan ylioppilaiden (597 kpl) liikkuvuutta ja maahanmuuttoa elämänkaaria kuvaavilla kaarilla MUUTTOLIIKE-välilehdellä. Kaaren sininen pää osoittaa syntymäpaikkaa ja punainen kuolinpaikkaa, joka useimmiten on nykyisen Suomen alueella, ja kaaren paksuus kuvastaa kaareen liittyvien henkilöiden lukumäärää. Jos henkilö on syntynyt ja kuollut samassa paikassa, kaarta ei näytetä. Kaarta klikkaamalla löytyvät siihen liittyvät linkit henkilöiden kotisivuille.

Kuvan 2.11 visualisointi KARTTA-välilehdellä näyttää Akatemiasammon paikat (yli 3000 paikkaa), joihin ylioppilaat tällä kertaa ilman mitään fasettirajausta liittyvät n. 175 000 tapahtuman kautta. Esimerkiksi Irlannissa olevaa markkeria klikkaamalla löytyy kaksi Irlantiin liittyvää henkilöä. Näistä toinen on sinne matkan tehnyt kuuluisa turkulainen ylioppilas, kemisti Johan Gadolin (1760–1855), joka löysi sittemmin uuden alkuaineen, yttriumin.

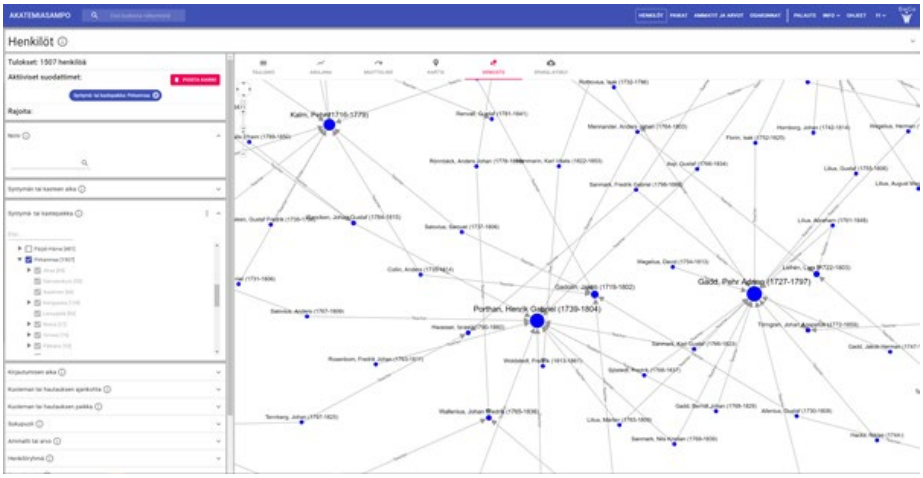
Välilehden VERKOSTO kautta avautuu näkymä faseteilla rajatun henkilöjoukon sisäisen akateemisen verkoston tutkimiseen. Kuvassa 2.12 käyttäjä on selvittämässä Pirkanmaalla syntyneiden ylioppilaiden välisiä opettaja–oppilasverkostoja.



Kuva 2.10: Akatemiasammon käyttöä prosopografisessa tutkimuksessa. Smålandin osakunnan jäsenten elämänkaaret syntymäpaikoista (kaaren sininen pää) kuolinpaikkoihin (punainen pää).



Kuva 2.11: KARTTA-välilehden visualisointi paikoista, joihin Akatemiasammon ylioppilaat liittyvät n. 175 000 tapahtuman kautta.

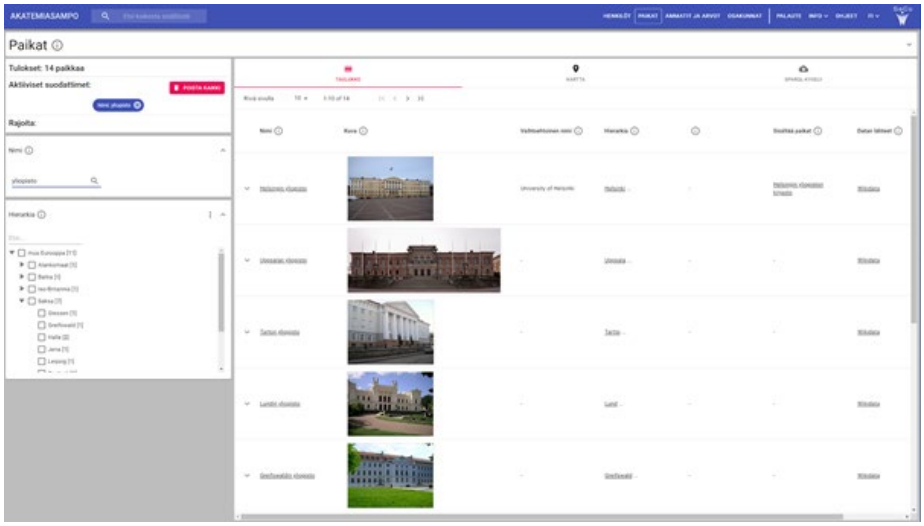


Kuva 2.12: VERKOSTO-välilehden visualisointi Pirkanmaalla syntyneiden ylioppilaiden välisistä akateemisista opettaja-oppilas-verkostoista.

Paikat-näkymä

Akatemiasammon Paikat-näkymän kautta tarjoutuu käytettäväksi vastaavanlainen fasettihaku kuin henkilöt-näkymässä, mutta hakukohteena ovat nyt historialliset paikat. Paikkoja voidaan hakea hierarkkisen fasetin avulla tai tekstihaulla ja nähdä hakutulos taulukkona. Kuvassa 2.13 käyttäjä on hakenut Akatemiasammon yli 3000 historiallisen paikan joukosta ne, joiden nimessä esiintyy sana ”yliopisto”. Tuloslistalle oikealla on tarttunut 14 yliopistoa Suomesta, Ruotsista ja Virosta ja muualta Euroopasta Wikidatasta löytyvine valokuvineen, esimerkiksi seitsemän matrikkeleissa mainittua saksalaista yliopistoa. Paikka-aineisto on suomalaisten paikannimien osalta haettu Maanmittauslaitoksen PNR-tietokannan (Paikannimirekisteri) linkitetyn datan versiosta[17] sekä Kansalliskirjaston Finto.fi-ontologiapalvelun YSO-paikat[18] ontologiasta. Akatemiasammon varhaisemmassa aineistossa on runsaasti mainintoja ruotsalaisista paikannimistä, joiden geokoodaukseen käytettiin kansainvälisen GeoNames-tietokannan[19] aineistoa. Ulkomaalaisten paikkojen osalta tärkein lähde on kuitenkin ollut Wikidata, josta on myös haettu myös paikkasivujen mahdolliset valokuvat ja vaakunat.

17 https://www.ldf.fi/dataset/pnr/index.html
18 YSO-paikat: https://finto.fi/yso-paikat/fi/
19 GeoNames: https://www.geonames.org/



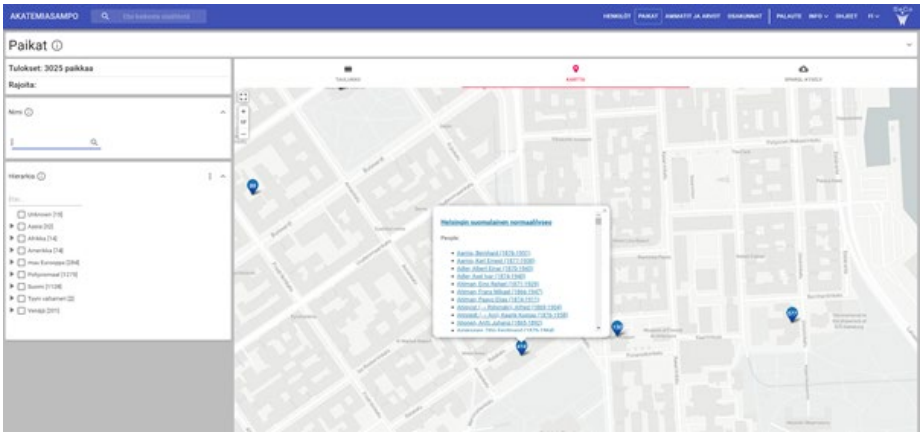
Kuva 2.13: Paikat-näkymän avulla voidaan hakea paikkoja ja nähdä ne taulukkona tai kartalla omalla välilehdellä KARTTA.

Hakutulos voidaan myös visualisoida kartalla erillisellä KARTTA-välilehdellä. Paikkaa kartalla klikkaamalla, avautuu ponnahdusikkuna, joka luettelee linkkeinä ne henkilöt, joihin liittyviä tapahtumia, esimerkiksi kuolema tai uraan liittyvä tapahtuma, tiedetään tapahtuneen valitussa paikassa. Ponnahdusikkunassa oleva henkilölinkki johtaa henkilön kotisivulle tapahtuman tarkempaa tutkimista varten alkuperäisestä matrikkeliaineistosta.

Kuvassa 2.14 käyttäjä on avannut Paikat-näkymän KARTTA-välilehden ja zoomannut Helsinkiin, josta löytyy paikkana mm. Ratakadulla oleva Helsingin normaalilyseo ”Norssi” markkerilla merkittynä. Norssin sijoittaminen osoitteeseen Ratakatu 6 kartalle on esimerkki linkitetyn datan mahdollisuuksista: matrikkeleista löytyy vain mainintoja normaalilyseosta, mutta Wikidatasta[20] dataa rikastamalla selviää myös mm. koulun sijainti kartalla ja saadaan Akatemiasammon käyttöön myös koulun valokuva avoimella lisenssillä. Markkeria klikkaamalla avautuu ponnahdusikkuna, joka luettelee kaikki Akatemiasammosta löytyneet 415 Norssin oppilasta ja muuta kouluun liittyvää henkilöä, kuten Suomen urheilun isänä tunnettu koulun voimistelunopettaja ja professorin arvon saanut Ivar Edvard Wilskman (1854–1932), ja linkit heidän kotisivuilleen. Norssin oppilaista löytyy lisää tietoa koulun historiallisesta matrikkelista aiemmin tehdystä verkkopalvelusta ”Vanhat

20 Helsingin normaalilyseo Wikidatassa: `https://www.wikidata.org/wiki/Q3269135`

Norssit semanttisessa webissä" (Hyvönen et al., 2017), joka on yksi Akatemiasampoa edeltäneistä henkilöhistoriallisista sammoista.



Kuva 2.14: Paikat-näkymän KARTTA-välilehdellä näkyvät paikat ja niihin tapahtumien kautta liittyvät henkilöt. Kuvassa löydetään Ratakadulta Helsingin normaalilyseoon tapahtumien kautta liittyvät 415 oppilasta ja muuta henkilöä, kuten Suomen urheilun isänä tunnettu voimistelun opettaja Ivar Edvard Wilskman (1854–1932).

Ammatit ja arvot -näkymä

Ammatit ja arvot -näkymä tarjoaa mahdollisuuden henkilöiden ja henkilöryhmien hakemiseen ammattien ja arvojen sekä henkilöön liittyvien paikkojen avulla. Käytetty ammattien ja arvojen luokitus perustuu SeCo-ryhmässä kehitettyyn historiallisten ammattien ja arvojen AMMO-ontologiaan (Koho et al., 2019), joka on linkitetty mm. kansainväliseen HISCO-luokitukseen[21]. AMMO-ontologia tarjoaa mahdollisuuksia tutkia esimerkiksi ylioppilaiden sosiaalista asemaa tai ammattien periytyvyyttä sukupolvien yli.

Osakunnat-näkymä

Osakunnat ovat muodostaneet tärkeän osan yliopistojen elämää keräten yhteen samalta alueelta kotoisin olevat ylioppilaat ja luomalla yhteyksiä opiskelijoista yliopistojen hallintoon. Osakunta-instituutio perustettiin Turun akatemiaan vuonna 1643. Helsingin yliopiston nykyiset osakunnat ovat alkuperäisten

21 HISCO classification: `https://iisg.amsterdam/en/data/data-websites/history-of-work`

osakuntien perillisiä, mutta monet osakunnat ovat aikojen kuluessa jakaantuneet tai yhdistyneet uusiksi osakunniksi. Kokonaan uusia osakuntia on myös perustettu ja vanhoja lakkautettu.[22] Akatemiasammon Osakunnat-näkymässä voidaan hakea osakuntia, ja niille on luotu omat kotisivut samaan tapaan kuin henkilöille ja paikoille. Näille on kerätty esimerkiksi osakunnan jäsenet eri aikoina, kuraattorit, inspehtorit ja kunniajäsenet linkkeinä heidän kotisivuilleen, sikäli kun heistä on mainintoja matrikkelin artikkelien teksteissä. Datassa on suomalaisten osakuntien ohella viittauksia myös ulkomaisiin osakuntiin esimerkiksi Uppsalan ja Lundin yliopistoissa.

Akatemiasammon datapalvelun rajapinnat

Akatemiasampo.fi-portaali tarjoaa edellä kuvattuja valmiiksi toteutettuja haku-, selailu- ja data-analyyttisiä työkaluja datan tutkimista varten. Palvelun taustalla oleva data on myös vapaasti käytettävissä muunlaisten analyysien tekemiseen ja sovellusten kehittämiseen. Akatemiasammon data on muiden sampojen tapaan julkaistu verkossa Linked Data Finland -palvelussa W3C:n standardien ja linkitetyn datan julkaisukäytäntöjen mukaisella tavalla. Data on avattu käytettäväksi avoimella CC BY-4.0 -lisenssillä, ja portaali sekä portaalin käyttöliittymän kehittämiseen käytetty Sampo-UI-ohjelmointikehys ovat avointa koodia vastaavalla lisenssillä. Tärkein rajapinta dataan on Akatemiasammon SPARQL-palvelupiste[23], josta dataa voi kysellä ja analysoida joustavasti semanttisen webin SPARQL-kyselykielen avulla.

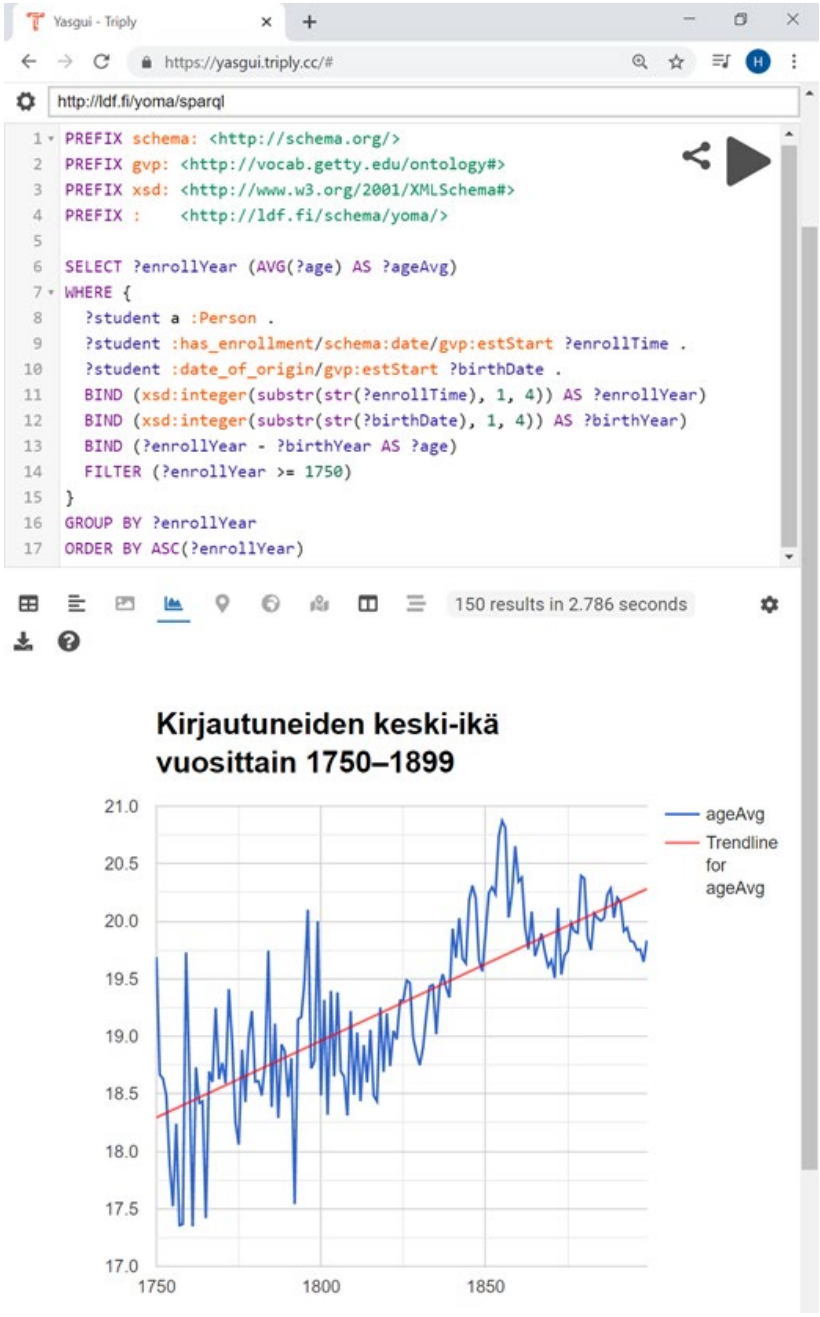
Tämän artikkelin tavoitteena on kannustaa digitaalisten ihmistieteiden tutkijoita hyödyntämään linkitetyn avoimen datan tarjoamia uusia mahdollisuuksia. Esittelemme siksi seuraavassa lyhyesti esimerkkeinä näistä mahdollisuuksista YASGUI-työkalun ja Google Colab / Jupyter-dokumenttien käyttöä. Nämä työkalut tarjoavat kevyen ja yksinkertaisen tavan luoda analyysejä datasta ja jakaa niitä toiminnallisina dokumentteina muidenkin arvioitavaksi ja käytettäväksi avoimen tieteen periaatteiden mukaisesti. SPARQL-rajapinnan kautta on mahdollista luoda datasta halutun muotoinen taulukkoesitys, jota voi analysoida millä tahansa kullekin tutkijalle sopivimmilla työkaluilla. Akatemiasammon haku- ja kotisivuilta löytyy erillinen SPARQL-KYSELY-välilehti, jolta löytyy linkki YASGUI-työkaluun, jossa on valmiiksi ohjelmoituna sivuun liittyvä kysely, sekä linkki linkitetyn datan selaimeen, jonka avulla voi tutustua tarkemmin Akatemiasammon datan rakenteisiin.

22 Suomalaisten osakuntien kehittymistä vuodesta 1643 on kuvattu kattavasti verkkosivustolla `https://osakunta.fi`.

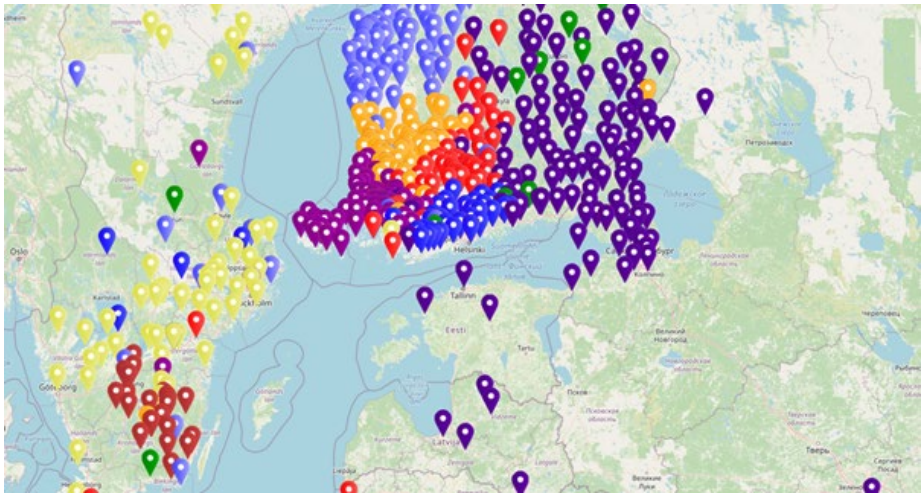
23 Akatemiasammon SPARQL-palvelupiste: `https://ldf.fi/yoma/sparql`

YASGUI tarjoaa helppokäyttöisen selaimessa toimivan verkkotyökalun SPARQL-kyselyiden kirjoittamista varten. Kyselyiden vastaukset ovat tutkittavissa mm. taulukkomuodossa ja niitä voidaan myös helposti visualisoida valmiiksi ohjelmoiduilla tavoilla, esimerkiksi näyttää paikkatietoa sisältävää dataa kartalla, tai muodostaa erilaisia kaavioita. Esimerkiksi kuvassa 3.1 näkyy kysely, jossa on laskettu opiskelijoiden keski-ikä vuosittain karkeasti syntymä- ja kirjautumisvuoden perusteella vuodesta 1750 eteenpäin. Kirjautumisiän keskiarvon voi helposti nähdä kasvaneen tänä aikana. Erityisen merkittävä kasvu vaikuttaisi tapahtuneen noin vuosina 1825–1850. Itse kysely vaatii tässä vain hieman yli kymmenen riviä SPARQL-kyselykieltä. Kyselyn tuloksia on visualisoitu YASGUI-sovelluksen valmiilla työkalulla. Visualisoinnin asettaminen on vaatinut vain muutaman painalluksen: on valittu valikosta visualisoinnin tyypiksi viivakaavio, sekä lisätty selitysteksti ja asetettu käyttöön trendilinja.

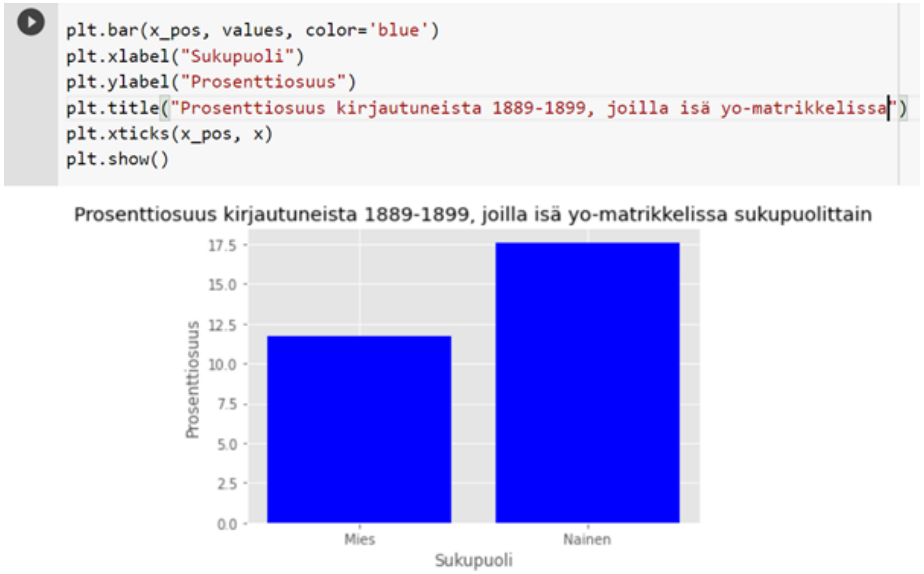
Kuvassa 3.2 on vastaavasti visualisoitu paikkatietoa YASGUI-editorilla. Kuvassa on merkitty paikkoja symboleilla, jotka on väritetty sen mukaan, mihin osakuntaan siellä syntyneistä enemmistö on kuulunut. Kuvasta voi nähdä osakuntien kanta-alueiden muodostuvan melko selvärajaisesti. Voi esimerkiksi nähdä, että Baltia ja Venäjä ovat selkeästi olleet Viipurilaisen osakunnan kanta-aluetta. Tämän visualisoinnin toteuttaminen on vaatinut vain hieman monimutkaisemman kyselyn, jonka avulla data on noudettu sopivassa muodossa. Tämän jälkeen riittää valita ”Geo”-välilehti editorin valikosta, ja visualisointi syntyy automaattisesti.



Kuva 3.1: YASGUI-editorilla voi hakea dataa joustavasti Akatemiasammosta semanttisen webin SPARQL-kyselykielen avulla ja myös visualisoida tuloksia graafisesti.



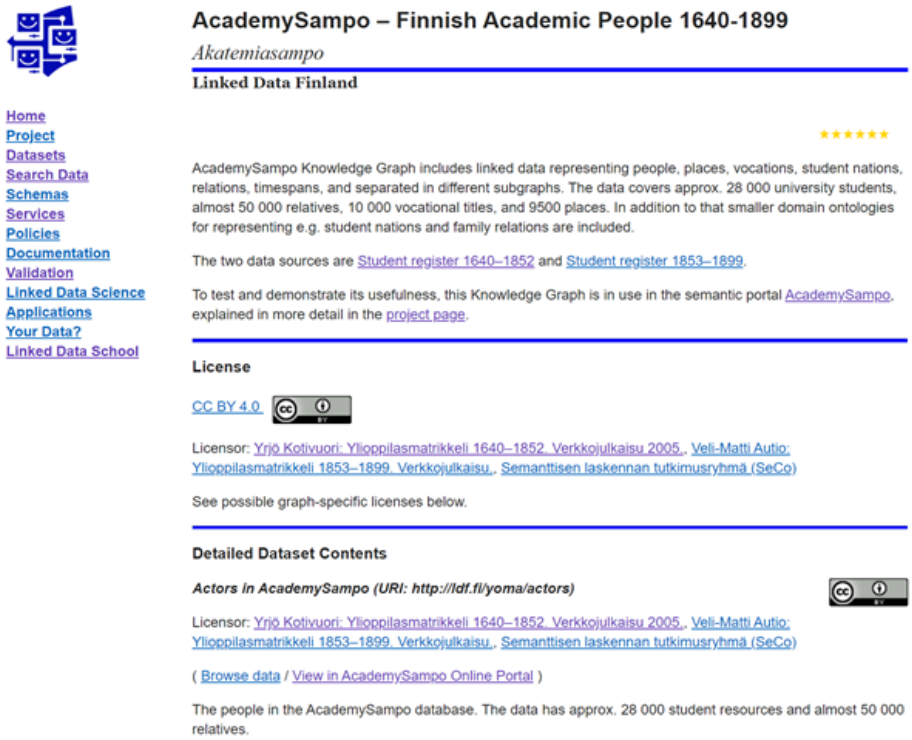
Kuva 3.2: YASGUI-editorilla toteutettu paikkatietoa käyttävä karttavisualisointi, jossa ylioppilaiden syntymäpaikat on väritetty sen mukaan, mistä osakunnasta on ollut eniten siellä syntyneitä. Kysely ja visualisointi on testattavissa osoitteessa `https://api.triplydb.com/s/xcJe_Hj0n`.



Kuva 3.3: Data-analyysiä Akatemiasammosta Google Colab -dokumentin avulla.

Monimutkaisempaa analyysiä tai laajempia räätälöintimahdollisuuksia varten voi SPARQL-kyselyn tuloksia analysoida erilaisten ohjelmointikielten kirjastojen avulla. Google Colab -palvelu tarjoaa vaivattoman mahdollisuuden kirjoittaa ja ajaa Python-ohjelmia verkossa Jupyter-dokumentteina pelkän verkkoselaimen avulla, editoida niitä yhteistyössä muiden kanssa ja jakaa tuloksia helposti ja visuaalisesti. Dokumentti voi koostua selittävistä tekstiosuuksista, tulkittavista koodiosuuksista ja Python-ohjelmien ja kirjastojen tekemistä visualisoinneista datalle, joka haetaan käyttöön kyselyillä SPARQL-palvelupisteeseen.

Kuvassa 3.3 on laskettu esimerkkinä sukupuolittain matrikkelin vuosina 1889–1899 kirjautuneille henkilöille prosenttiosuudet siitä, kuinka suurella osalla ylioppilaista myös isä on ylioppilasmatrikkelissa. Kuvasta voi nähdä, että naisilla tämä osuus on suurempi. Kuvassa näkyy myös yläpuolella osa ohjelmakoodia. Visualisointi on luotu Python-ohjelmointikielen Matplotlib-kirjaston avulla. Koodiin voi helposti tehdä muutoksia. Vasemmassa yläkulmassa näkyvää ”play”-symbolia painamalla tämä koodiosuus suoritettaisiin uudestaan ja mahdolliset muutokset näkyisivät kaaviossa.



Kuva 3.4 Osa Akatemiasampo-datapalvelun kotisivusta (`http://www.ldf.fi/dataset/yoma`) Linked Data Finland -alustalla (`http://ldf.fi`), jolta löytyvät mm. muidenkin sampojen datapalvelut. LDF.fi käyttää hyväkseen CSC – Tieteen tietotekniikan keskus Oy:n tarjoamaa kansallista palvelininfrastruktuuria.

Lisätietoa ja dokumentaatiota Akatemiasammon linkitetyn avoimen datan julkaisusta ja SPARQL-palvelupisteestä löytyy sille luodulta kotisivulta Linked Data Finland -palvelussa (kuva 3.4). Julkaisu on Tim Berners-Leen viiden tähden mallin[24] mukainen, mutta LDF-alustassa ehdotettu seitsemän tähden malli antaa vielä kuudennen tähden, koska datajulkaisussa on mukana myös tietomallin kuvaus helpottamassa datan uudelleenkäyttöä. Seitsemäs tähti edellyttäisi datan validointia, mitä tässä vaiheessa ei ole systemaattisesti tehty.

24 Linkitetyn datan viiden tähden julkaisumalli: `https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/`

Datalukutaitoa tarvitaan

Akatemiasampo-projektin käyttöönsä saaman alkuperäinen tietokannan matrikkelidata koostui CSV-muotoisista taulukoista. Matrikkelit 1640–1852 ja 1853–1899 ovat eri henkilöiden digitoimia ja toimittamia ja niiden taulukkomuotoinen data poikkeaa jossain määrin toisistaan.

Matrikkelitaulukosta 1640–1852 löytyviä tietoja ovat eräiden tietokannan teknisten tietojen ohella 1) henkilön matrikkelinumero, 2) HTML-muotoinen teksti, josta käy ilmi henkilön nimi, syntymäpaikka ja aika, vanhemmat, uraan liittyviä tapahtumia, kuolinpaikka ja vuosi, sukulaisia, oppilaita, viitetietoja kirjallisuuteen ja 3) tietueen luomisen päiväys. Jos matrikkelissa 1640–1852 mainittu henkilö löytyy jommastakummasta matrikkelista, on hänen mainintansa yhteydessä manuaalisesti luotu HTML-linkki ko. henkilön sivulle matrikkelinumeron avulla. Matrikkelissa 1853–1899 tällaisia linkkejä ei kuitenkaan ole, vaan viittaukset on jouduttu tulkitsemaan koneellisesti. Lisäksi henkilöön on voitu kirjata tekstimuotoista lisätietoa muista matrikkelista. Esimerkiksi edellä esitellyn Johan Ludvig Runebergin kohdalla on lisätietoja Laguksen ja Carpelanin matrikkeleista (ruotsin kielellä).

Akatemiasampo-hankkeen käyttämä primaaridata oli siis lähinnä tekstiä HTML-muodossa ilman rakenteista metadataa, kuten syntymäpaikka/aika, ammatti jne. Linkitetyn datan muodostamisen ensimmäisenä teknisenä haasteena olikin tunnistaa yksikäsitteisesti tekstissä mainitut nimetyt entiteetit, kuten henkilöt, paikat ja organisaatiot, ajan ilmaukset, erilaiset tapahtumat, kuten avioliitot, palkitsemiset ja promootiot sekä matrikkelitiedon kannalta keskeiset käsitteet kuten ammatit ja arvot. Omat haasteensa tiedon irrottamisessa muodostivat samannimisten henkilöiden erottaminen toisistaan, sukupuolen päättely nimen perusteella ja erilaisten sukulaissuhteiden, kuten pikkuserkku ja lanko, päätteleminen toisten sukulaissuhteiden avulla.

Matrikkeliteksteistä tunnistetut entiteetit, käsitteet ja näiden väliset suhteet muodostavat perustan Akatemiasammon linkityksille, hakutoiminnoille, data-analyyseille ja visualisoinneille. Akatemiasammon kaltaisen järjestelmän rakenteinen metadata data on tuotettu suureksi osaksi automaattisesti ja käyttäjältä edellytetään sen käyttämisessä uudenlaista datalukutaitoa (Koltay, 2015). Järjestelmän esiin nostamat rakenteet ja linkitykset perustuvat alkuperäisiin teksteihin, jotka voivat olla osin puutteellisia ja virheellisiäkin. Tämän lisäksi käytetyt algoritmit eivät välttämättä ole onnistuneet tunnistamaan tekstistä kaikkia haluttuja ilmauksia ja aineistojen entiteettien tunnistuksessa ja niiden merkitysten yksilöinnissä voi tapahtua virheitä. Myös tunnistettavat käsitteet voivat olla keskenään epäyhteentoimivia (esimerkiksi eri aikakausien ammattinimikkeet) ja niiden merkitys voi

muuttua ajan kuluessa (esimerkiksi historialliset paikat ja alueet). Linkitettyyn ontologiseen dataan perustuvat järjestelmät nostavat datan epäjohdonmukaisuudet, virheet ja puutteet voimakkaasti esille käyttöliittymässä.

Esimerkiksi Akatemiasammon osakunnista löytyy sekä lakkautettuja että edelleen toimivia osakuntia ja paikkaontologiassa on mukana Neuvostoliitolle menetettyjä alueita, jotka kuitenkin olivat aiemmin osa Suomea ja Ruotsia. Tällaiset tiedon esittämisen haasteet eivät niinkään johdu linkitetyn datan menetelmistä kuin kuvattavan reaalimaailman ontologisesta monimutkaisuudesta ja historialliseen matrikkelitietoon liittyvistä puutteista ja epätarkkuuksista, mutta ontologioiden määritteleminen ja käyttö haussa, selailussa ja data-analyyseissä paljastaa datan rakenteet. Perinteisissä hakujärjestelmissä ongelmat tavallaan jäävät maton alle piiloon tekstimuotoiseen dataan ja ihmisen tulkittavaksi aineistojen lukemisen yhteydessä.

Akatemiasammon tavoitteena on helpottaa tutkijan työtä matrikkeliaineiston läpikäymisessä ja tutkimisessa louhimalla automaattisesti kokoon kiinnostavia viittauksia, linkkejä ja visualisointeja silloin kuin se on teknisesti mahdollista. Tämä on esimerkki digitaalisten ihmistieteiden ”kaukoluvusta” (distant reading) (Moretti, 2013). Tällaista semanttisesti linkitettyä, rikasta dataa käytettäessä syntyy helposti sellainen harhakuva, että data ja linkitykset olisivat täydellisiä ja tietojen puutteet tuntuvat virheiltä. Muistettava on, Akatemiasammon kaltaisen järjestelmän esiin louhima tietämys perustuu luonnollisesti vain käytettävissä oleviin aineistoihin. Esimerkiksi henkilöistä, joista ei ole omaa artikkelia matrikkelissa, kuten useimmista ylioppilaiden vaimoista ja sukulaisista tai James Cookin kaltaisista hahmoista, ei käytettävissä ole muuta tietoa kuin maininnat matrikkelihenkilöiden kuvausten yhteydessä.

Biografisten aineistojen data-analyyttiseen tutkimiseen on maailmalla kehitetty ja käytetty erilaisia järjestelmiä (Larson, 2010; ter Braake et al., 2015; Warren et al., 2016; Fokkens et al., 2017; Warren, 2018; Bhreathnach et al., 2019; Jatowt et al., 2019). Akatemiasammon innovatiivisuus perustuu linkitetyn datan käyttöön aineistojen yhdistämisessä ja rikastamisessa, tapahtumaperustaiseen (event-based) CIDOC CRM -standardia laajentavaan tiedon esittämistapaan ja Sampo-mallin käyttöön. Työ on jatkoa SeCo-tutkimusryhmän aiemmille prosopografisille datapalveluilla ja portaaleille Vanhat Norssit semanttisessa webissä (Hyvönen et al., 2017), U. S. Congress Prosopographer (Miyakita et al., 2018) ja Biografiasampo (Hyvönen et al., 2019). Akatemiasammon ja Sampo-sarjan kaltaista järjestelmää, kokonaisuutta ja semanttisen webin kansallista tietoinfrastruktuuria ei parhaan tietomme mukaan muualla maailmassa ole olemassa, mutta semanttisen webin

teknologioita on alettu käyttää yhä enemmän kulttuurialan järjestelmissä ja digitaalisten ihmistieteiden tutkimuksessa (Bikakis et al., 2021).

Haasteista huolimatta semanttisen webin ontologisten käsiterakenteiden käyttö laajojen aineistokokonaisuuksien luomisessa, kyselyiden muodostamisessa sekä hakutulosten jäsentämisessä, visualisoinnissa ja tutkimisessa on hyödyllistä, kuten esimerkiksi A. Fokkensin ja kumppaneiden (2017), C. Warrenin (2018) ja Ú. Bhreathnachin ja kumppaneiden analyysit Alankomaiden, Iso-Britannian ja Irlannin kansallisbiografioista sekä oma työmme prosopografisten Sampo-portaalien parissa osoittavat. Tässä katsauksessa on pyritty havainnollistamaan näitä uusia mahdollisuuksia esimerkeillä ja kuvakaappauksilla Akatemiasammosta. Akatemiasampo tarjoaa uudenlaisia tapoja etsiä laajasta matrikkelidatasta (big data) tehokkaasti kiinnostavia ilmiöitä ja osajoukkoja ja analysoida niitä (distant reading), mutta tulosten tulkintaan tarvitaan edelleen aineistojen lähilukua ja datan lukutaitoa oikeiden johtopäätösten tekemistä ja tieteellistä perustelemista varten.

Lisätietoa Akatemiasammosta

WWW-osoitteet

- Akatemiasampo-portaali: `https://akatemiasampo.fi`
- Akatemiasampo-datapalvelu: `https://ldf.fi/dataset/yoma/`
- Akatemiasampo-hankkeen kotisivu: `https://seco.cs.aalto.fi/projects/yo-matrikkelit/`

Videoita Akatemiasampo-hankkeesta ja järjestelmästä

- Akatemiasampo – Akateemiset henkilöt Suomessa 1660–1899: visio ja sen toteutus: `https://vimeo.com/508756030`
- AcademySampo – Finnish Academic People 1640–1899: `https://vimeo.com/462993654`

Lähteet

Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B., & Mosca A. (2021). Editorial: Special Issue on Semantic Web for Cultural Heritage. *Semantic Web 12*(2). https://doi.org/10.3233/sw-210425

Bhreathnach, Ú., Burke, C., Fhinn, J. M., Cleircín, G. Ó., & Raghallaigh, B. Ó. (2019). A Quantitative Analysis of Biographical Data from Ainm, the Irish-language Biographical Database. *Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019)*.

Fokkens, A., ter Braake, S., Ockeloen, N., Vossen, P., Legêne, Schreiber, G., & de Boer, V. (2017). BiographyNet: Extracting Relations Between People and Events. Teoksessa *Europa baut auf Biographien* (pp. 193–224). New Academic Press.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hyvönen, E., Tuominen, J., Alonen, M., & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. Teoksessa *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers* (pp. 226–230). Springer. https://doi.org/10.1007/978-3-319-11955-7_24

Hyvönen, E. (2018). *Semanttinen web. Linkitetyn avoimen datan käsikirja*. Gaudeamus.

Hyvönen, E. (2020a). Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web, 11*(1), 187–193. https://doi.org/10.3233/sw-190386

Hyvönen, E. (2020b). "Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web. Teoksessa *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. http://ceur-ws.org/Vol-2612/poster1.pdf

Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., & Keravuori, K. (2019). BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. Teoksessa *The Semantic Web: ESWC 2019* (pp. 574-589). Springer. https://doi.org/10.1007/978-3-030-21348-0_37

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2021). Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web*, accepted. http://www.semantic-web-journal.net/

Jatowt, A., Kawai, D., & Tanaka, K. (2019). Time-focused Analysis of Connectivity and Popularity of Historical Persons in Wikipedia. *International Journal on Digital Libraries, 20*(4), 287–305. https://doi.org/10.1007/s00799-018-0231-4

Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., & Hyvönen, E. (2019). AMMO Ontology of Finnish Historical Occupations. Teoksessa *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 19)* (pp. 91–96). CEUR Workshop Proceedings, Vol. 2375. http://ceur-ws.org/Vol-2375/

Koltay, T. (2015). Data Literacy for Researchers and Data Librarians. *Journal of Librarianship and Information Science, 49*(1), 3–14. https://doi.org/10.1177/0961000615616450

Larson, R. (2010). *Bringing Lives to Light: Biography in Context. Final Project Report*. University of Berkeley. http://metadata.berkeley.edu/Biography_Final_Report.pdf

Leskinen, P., Hyvönen, E. & Tuominen, J. (2018). Analyzing and Visualizing Prosopographical Linked Data Based on Biographies. Teoksessa *BD2017 Proceedings of the Second Conference on Biographical Data in a Digital World 2017* (pp. 39–44). CEUR Workshop Proceedings, Vol. 2119. http://ceur-ws.org/Vol-2119/

Leskinen, P., & Hyvönen, E. (2020). Linked Open Data Service about Historical Finnish Academic People in 1640–1899. Teoksessa *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. http://ceur-ws.org/Vol-2612/

Miyakita G., Leskinen P., & Hyvönen E. (2018). Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators. Teoksessa M. Ioannides et al. (eds), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection; EuroMed 2018* (pp. 150–162). Lecture Notes in Computer Science, vol 11197. Springer. https://doi.org/10.1007/978-3-030-01765-1_18

Moretti, F. (2013). *Distant Reading*. Verso Books.

Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M, . . . Nevalainen, T. (2020). Wrangling with Non-standard Data. Teoksessa *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference* (pp. 81–96). CEUR Workshop Proceedings, Vol. 2612. http://ceur-ws.org/Vol-2612/

Rietveld. L., & Hoekstra, R. (2017). The YASGUI Family of SPARQL Clients. *Semantic Web 8*(3), 373–383. https://doi.org/10.3233/SW-150197

ter Braake, S., Fokkens, A., Sluijter, R., Declerck, T., & Wandl-Vogt, E. (eds) (2015). *BD2015 Biographical Data in a Digital World 2015*. CEUR Workshop Proceedings, Vol. 1399. http://ceur-ws.org/Vol-1399/

Verboven, K., Carlier, M., & Dumolyn, J. (2007). A Short Manual to the Art of Prosopography. Teoksessa *Prosopography Approaches and Applications. A Handbook* (pp. 35–70). Unit for Prosopographical Research (Linacre College).

Warren, C. N. (2018). Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics, 1*(2). https://doi.org/10.22148/16.028

Warren, C., Shore, D., Otis, J., Wang, L, Finegold, M., & Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *Digital Humanities Quarterly, 10*(3). http://www.digitalhumanities.org/dhq/vol/10/3/000244/000244.html