Yhteyshaku semanttisessa webissä

Heikki Rantala

Helsinki 16.1.2019

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen osasto

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Avdelning — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen osasto

Tekijä — Författare — Author
Heikki Rantala

Työn nimi — Arbetets titel — Title

Yhteyshaku semanttisessa webissä

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	16.1.2019	60 sivua

Tiivistelmä — Referat — Abstract

Tavallisessa haussa etsitään yksilöitä, kuten henkilöitä tai paikkoja. Joissain tilanteissa esimerkiksi historian tutkija voi olla kiinnostunut myös etsimään yhteyksiä henkilöiden ja paikkojen välillä. Tässä työssä esitetään metodi tällaisen yhteyshaun toteuttamiseksi käyttäen semanttisen webin sisältämää avointa dataa. Työssä muodostettiin graafi, joka sisältää kuvauksia Suomen kulttuurihistorian henkilöiden ja paikkojen välisistä kiinnostaviksi arvioiduista yhteyksistä. Graafi luotiin SPARQL CONSTRUCT -kyselyillä. Yhteyksien hakemista varten luotiin web-sovellus, joka hyödyntää fasettihakua.

Tarvittavien SPARQL CONSTRUCT -kyselyjen luominen ei osoittautunut erityisen hankalaksi, mutta niiden soveltaminen yleisemmin eri aineistoihin vaatii jonkin verran työtä. Yhteyksien fasettihaku osoittautui mielenkiintoiseksi. Fasettihaku mahdollistaa haun tarkentamisen askel kerrallaan. Lisäksi yhteyksien suhteellisia määriä on mahdollista vertailla erilaisten rajausten mukaan. Tämä tarjoaa aineistoon uusia näkökulmia.

ACM Computing Classification System (CCS):
Information systems → World Wide Web → Web data description languages → Semantic web description languages → Resource Description Framework (RDF)
Information systems → Information retrieval

Avainsanat — Nyckelord — Keywords
linkitetty data, SPARQL, RDF, yhteyshaku

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Internet sisältää valtavan määrän avointa dataa erilaisista aiheista kuten esimerkiksi kulttuurihistoriasta. Tätä suurta tietomäärä voi ihminen selata webin avulla, mutta ihmisen kyky suurten tietomäärien käsittelyyn on rajallinen. Tiedon automaattisen käsittelyn helpottamiseksi tietoa voidaan julkaista *linkitettynä datana* (linked data) [5], jossa datan suhteet muuhun dataan ilmaistaan koneluettevassa muodossa. Tällaisen linkitetyn datan muodostamaa koneymmärrettävää webiä kutsutaan *semanttiseksi webiksi* (Semantic Web) [4]. Tässä työssä esitetään menetelmä asioiden välisten yhteyksien löytämiseen semanttisen webin avulla ja luodaan tätä menetelmää soveltaen sovellus Suomen historian henkilöiden ja paikkojen välisten kiinnostavien yhteyksien etsimiseen, erityisesti keskittyen kulttuurihistoriaan.

Historiallisten henkilöiden ja paikkojen välisistä yhteyksistä voidaan olla kiinnostuneita monesta syystä. Kulttuurihistorian tutkija saattaisi esimerkiksi tietää, että eräs taiteilija on kuvannut tiettyä paikkaa teoksissaan, ja hän saattaisi olla kiinnostunut tietämään, onko tämän henkilön ja paikan välillä jotain yllättäviä yhteyksiä, jotka saattaisivat selittää taiteilijan kiinnostuksen kyseiseen paikkaan. Tutkija saattaisi olla kiinnostunut myös laajemmista kokonaisuuksista kuten siitä, ketkä suomalaiset taiteilijat ovat kuvanneet Italiaa teoksissaan. Myös esimerkiksi Hämeen maakunnan historiasta kiinnostunut henkilö voisi olla kiinnostunut siitä, millaisia yhteyksiä Suomen historian merkittävillä henkilöillä on Hämeeseen. Tavallisilla hakumetodeilla voidaan hakea esimerkiksi kiinnostavaa henkilöä tai paikkaa, mutta tällaisessa tilanteessa olisi hyödyllistä kyetä suoraan hakemaan yhteyksiä asioiden välillä ja vertailemaan niitä. Oleellista on myös kyetä hakemaan yhteyksiä laajempiin kokonaisuuksiin. Esimerkiksi Hämeestä kokonaisuutena kiinnostuneen tutkijan olisi vaivalloista etsiä yhteydet jokaiselle Hämeessä sijaitsevalle paikkakunnalle erikseen. Koneymmärrettävä data voi auttaa tietokonetta ymmärtämään Hämeenlinnan kuuluvan Hämeeseen ja siten mahdollistaa tällaisen laajemman haun.

Tavallisilla hakumetodeilla voidaan hakea esimerkiksi kiinnostavaa henkilöä tai paikkaa, mutta tällaisessa tilanteessa olisi hyödyllistä kyetä suoraan hakemaan yhteyksiä asioiden välillä ja vertailemaan niitä. Tällaiselle *yhteyshaulle* (relational search) on luotu erilaisia menetelmiä, joilla on omia heikkouksiaan ja vahvuuksiaan [39, 8]. Tässä työssä esitetään menetelmä tietämykseen perustuvalle kulttuurihistorian yhteyshaulle. Tämä tarkoittaa sitä, että tarkoitus ei ole kehittää kaikille sovellusaloille soveltuvaa yleispätevää ja täysin automaattista järjestelmää, vaan yhteyksien kiinnostavuuden arviointi perustuu ihmisen ymmärrykseen.

Työhypoteesina on muuntaa datasta SPARQL CONSTRUCT -kyselyjen avulla uusi graafi, joka koostuu Suomen kulttuurihistorian henkilöiden ja paikkojen välisiä yhteyksiä kuvaavista, etukäteen lasketuista yhteys-luokan instansseista. Näiden kriittisinä ominaisuuksina ovat semanttisen yhteyden päätepisteet, paikka tai henkilö, sekä näiden välisen yhteyden selitys luonnollisella kielellä. Hypoteesina on, että kulttuurihistorian kannalta mielenkiintoiset yhteydet voidaan määritellä yleisten SPARQL-hahmojen avulla, joita voi soveltaa laajasti eri aineistoihin. Lisäksi hypoteesin mukaan näiden hahmojen avulla muodostettuja yhteyksiä voi kulttuurihistoriasta kiinnostunut käyttäjä selata tavalla, joka tarjoaa mahdollisuuden löytää uutta ja yllättävää tietoa.

Yhteyksien hakemista varten toteutettiin web-sovellus. Sovelluksen avulla yhteyksiä voi hakea fasettihaun avulla. Yhteyksistä näytetään käyttäjälle luonnollisen kielen selitys, sekä linkkejä yhteyteen liittyviin web-sivuihin. Sovellus toteutettiin osana elämäkertojen tutkimiseen tarkoitettua Biografiasampo-portaalia[1] [26, 42, 43, 27]. Biografiasampo, mukaan lukien tässä tässä työssä toteutettu yhteyshaku-näkymä, julkaistiin avoimeen käyttöön 27.9.2018.

Biografiasampo perustuu suurimmaksi osaksi Suomalaisen Kirjallisuuden Seuran julkaisemiin elämäkertoihin, joista on luohittu tietoa ja muutettu se koneymmärrettävään muotoon. Tässä työssä toteutettu yhteyshaku-sovellus käyttää lähteenään suurimmaksi osaksi Biografiasammon dataa, mutta dataa on rikastettu myös muista lähteistä, jotka liittyvät erityisesti kulttuurihistoriaan ja luovan työn tuloksiin. Näitä lähteitä ovat Kirjasampo[2] [22, 48, 49, 51, 50], Suomen kansallisbibliografia Fennica[3], HISTO-ontologia[4] [25, 28] sekä Kansallisgallerian tietokanta[5]. Osana työtä suoritettiin myös J. V. Snellmanin koottujen teosten[6] datan konversio RDF-graafiksi, ja käytettiin sitä yhtenä aineistona yhteyshaussa.

Tämä työ edustaa metodologialtaan *suunnittelutiedettä* [57] (design science). Työssä pyritään luomaan toimiva järjestelmä, arvioimaan sitä ja oppimaan saaduista kokemuksista.

Tämän työn tutkimuskysymykset liittyvät siihen, kuinka helppoa ja tehokasta on luoda SPARQL CONSTRUCT -kyselyillä yhteyksiä sisältävä graafi, sekä siihen,

[1] http://biografiasampo.fi/
[2] https://seco.cs.aalto.fi/applications/kirjasampo/
[3] https://www.kansalliskirjasto.fi/fi/palvelut/metadatan-muunto-ja-valityspalvelut/avoin-data
[4] https://seco.cs.aalto.fi/ontologies/histo/
[5] https://www.kansallisgalleria.fi/avoin-data/
[6] http://snellman.kootutteokset.fi/

miten käyttäjä voi hakea yhteyksiä tällaisesta graafista. Tämän työn tutkimuskysymykset voidaan muotoilla seuraavasti:

- Miten helppoa on muodostaa linkitetystä datasta henkilöiden ja paikkojen yhteyksiä kuvaava graafi SPARQL CONSTRUCT -kyselyillä, ja kuinka helppoa on yleistää näitä kyselyitä?
- Mitä etuja saadaan fasettihaun soveltamisesta yhteyshakuun?

Tämän työn toisessa luvussa käydään lyhyesti läpi perusteet semanttisesta webistä, tiedonhausta ja kiinnostavuudesta. Lisäksi esitetään lyhyt katsaus aikaisemmin toteutetuista yhteyshaku-sovelluksista. Kolmannessa luvussa käsitellään tässä työssä käytetyt datan lähteet ja niiden tämän työn kannalta oleellisimmat ominaisuudet. Neljännessä luvussa esitetään erikseen tarkemmin J. V. Snellmanin koottujen teosten aineiston muunnos linkitetyn datan muotoon. Viidennessä luvussa esitetään yhteyksiä kuvaavan RDF-graafin tietomalli ja datan louhimisen vaiheet. Kuudennessa luvussa käsitellään yhteyshaun demonstroimiseksi luotua web-sovellusta ja sen käyttöliittymää. Seitsemännessä luvussa arvioidaan saatuja tuloksia ja mahdollisuuksia jatkokehitykselle sekä esitetään yhteenveto työstä.

2 Yhteydet semanttisessa webissä

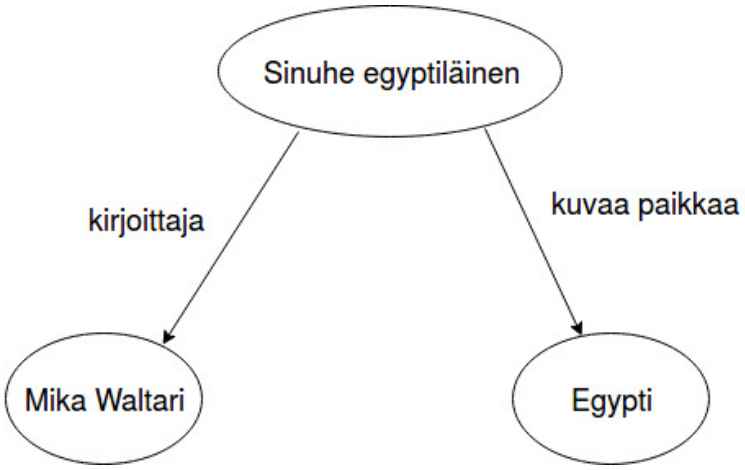
2.1 Semanttinen web

Semanttinen web [4] on World Wide Webin kehittäjänä tunnetun Tim Berners-Leen ja kumppaneiden ideoima ajatus webistä, jossa tiedon metadata on kuvattu rakenteisella tavalla linkitettynä datana, joka on helposti käsiteltävissä tietoteknisillä menetelmillä. RDF [40, 56] eli *Resource Description Framework* on yleiskäyttöinen tietomalli ja kieli informaation kuvaamiseksi semanttisessa webissä. Kaikkia RDF:llä kuvattavia asioita kutsutaan resursseiksi (*resource*), ja niillä on yksilölliset URI tunnisteet, jotka mahdollistavat niihin viittaamisen. *Ominaisuudet* (property) ovat suhteita tai ominaisuuksia, joilla resursseja kuvataan. RDF kuvaa tietoa *kolmikkojen* (triple) joukkona. Jokainen kolmikko on ilmaisu, joka koostuu lauseen subjektia, predikaattia ja objektia vastaavista osista. Ilmaisussa subjekti on kuvattava resurssi, predikaatti on ominaisuus, joka ilmaisee millainen suhde subjekti-resurssin ja objekti-resurssin välillä on. Objekti on resurssi tai literaali arvo, johon ominaisuus

viittaa. Yksinkertaistettu esimerkki RDF-kolmikosta ilman URI-tunnisteita voisi olla muotoa:

```
<Akseli Gallen-Kallela> <on maalannut teoksen> <Symposion>.
```

RDF-kolmikkojen joukkoa voi visualisoida verkkona jossa resurssit liittyvät toisiinsa eri tavoin. Esimerkiksi kuvassa 1 on esitetty yksinkertainen verkko, joka kuvaa romaania ”Sinuhe egyptiläinen”.



Kuva 1: Graafinen kuvaus verkosta, joka kuvaa romaanin ”Sinuhe egyptiläinen” aihetta ja kirjoittajaa.

Resurssien voidaan määritellä kuuluvan tiettyyn *luokkaan* (class). Esimerkiksi jos tietty resurssi kuvaa henkilöä, sen voi määritellä kuuluvan henkilöitä kuvaavaan luokkaan. Tällaisia luokkien *yksilöitä* voidaan kutsua myös *instansseiksi* (instance).

Tässä työssä pääasiassa käytetty syntaksi RDF-muotoisen tiedon kuvaamiseen on Turtle [6]. Turtle on helposti luettava syntaksi, jossa resurssien URI-tunniste kirjoitetaan lyhennetyssä muodossa. URI-tunnisteet voidaan jakaa kahteen osaan. Tunnisteen alkuosaa kutsutaan *nimiavaruudeksi* (namespace) ja loppuosaa *paikallisnimeksi* (local name). Nimiavaruutta voidaan kuvata vapaasti määriteltävän *etuliitteen* (prefix) avulla. Etuliitteet määritellään Turtle-notaatiolla ilmauksella @prefix näin:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
```

Ilmausta @prefix seuraa siis haluttu lyhenne, ja viimeisenä on kuvattava nimiavaruus. Esimerkiksi resurssi, jonka URI on

```
<http://xmlns.com/foaf/0.1/Person>,
```

voidaan kuvata Turtle-notaatiolla lyhyemmin muodossa

```
foaf:Person,
```

kun on määritelty aluksi etuliite yllä esitetyllä tavalla. Taulukko 1 sisältää keskeisimmät tässä työssä käytetyt nimiavaruudet ja niiden etuliitteet.

Etuliite	**Nimiavaruus**
rdf:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs:	<http://www.w3.org/2000/01/rdf-schema#>
schema:	<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
foaf:	<http://xmlns.com/foaf/0.1/>
skos:	<http://www.w3.org/2004/02/skos/core#>
dc:	<http://purl.org/dc/elements/1.1/>
crm:	<http://www.w3.org/2004/02/skos/core#>
snellman:	<http://ldf.fi/snellman/>
relse:	<http://ldf.fi/relse/>

Taulukko 1: Tässä työssä käytettävien nimiavaruuksien etuliitteitä

2.2 SPARQL-kyselykieli

SPARQL [60][24, s. 89–109] on kyselykieli tiedolle, joka on ilmaistu RDF-muodossa. SPARQL muistuttaa syntaksiltaan relaatiotietokantojen SQL-kieltä. SPARQL tarjoaa välineitä RDF-muotoisen datan kyselemiseen sekä graafien luomiseen, poistamiseen ja muokkaamiseen. SPARQL-kielen tämän työn kannalta keskeisin kyselytyyppi on CONSTRUCT, mutta myös SELECT-kysely esitellään sen yleisen tärkeyden vuoksi. SELECT-kysely palauttaa kyselyssä esiintyville muuttujille arvosijoituksia. CONSTRUCT-kysely taas muodostaa uutta RDF-muotoista dataa. SPARQL-kielessä haluttu tieto ilmaistaan graafihahmona (*graph pattern*). Hahmo on Turtle-notaatiolla kirjoitettu kolmikkojoukko, jossa osa resursseista on korvattu muuttujilla. Muuttujien nimen edessä on merkki ”?” tai ”$”.

Henkilöiden tietojen kuvaamiseen voidaan käyttä esimerkiksi FOAF-ontologiaa[7]. Esimerkiksi seuraavalla SELECT-kyselyllä voi hakea tietokannasta kaikki henkilöt, joiden sukunimi on ”Korhonen”:

```
SELECT ?person WHERE {
    ?person a foaf:Person .
    ?person foaf:familyName "Korhonen" .
}
```

Yllä olevassa kyselyssä ilmauksen ”SELECT” jälkeen tulevat muuttujat ovat ne, joiden tiedot halutaan kyselyn vastaukseksi. Tässä on merkitty ainoastaan muuttuja ?person, joten kyselyn vastaukseksi tulee listaus muuttujan ?person arvoja. Ilmausta ”WHERE” seuraa aaltosulkeissa oleva lohko, jonka sisältö määrää graafihahmon, joka rajoittaa kyselyn tulosjoukon. Yllä olevassa graafihahmossa määritellään ensin, että muuttuja ?person on tyyppiä

```
foaf:Person.
```

Ilmaisu ”a” on lyhenne ominaisuudelle

```
rdf:type,
```

joka ilmaisee resurssin olevan tietyn luokan yksilö. Seuraavalla rivillä rajataan tulosjoukkoa lisää siten, että kaikilla muuttujan ?person arvoksi tulevilla entiteeteillä tulee olla myös ominaisuus

```
foaf:familyName,
```

ja sen arvona literaali ”Korhonen”. Siten tulosjoukkoon rajautuvat vain ne henkilöä kuvaavan luokan yksilöt joiden sukunimenä on lisäksi ”Korhonen”.

CONSTRUCT-kysely palauttaa vastauksena RDF-muotoista dataa, josta voi muodostaa uuden graafin. Tässä työssä luotu yhteyksiä kuvaava RDF-graafi on muodostettu suurimmaksi osaksi tällaisilla kyselyillä. Esimerkiksi seuraava kysely valitsee tietokannasta kaikki resurssit, jotka ovat tietyn henkilöä kuvaavan luokan yksilöitä, ja luo uuden graafin jossa nämä resurssit on määritelty toisen henkilöä kuvaavan luokan yksilöiksi:

[7] http://xmlns.com/foaf/spec/

```
CONSTRUCT {
    ?person a schema:Person .
}
WHERE {
    ?person a foaf:Person .
}
```

Osa graafihahmosta on mahdollista merkitä valinnaiseksi ilmaisulla OPTIONAL. Tätä ominaisuutta voi käyttää, kun halutaan saada tietty tieto, mutta ei kuitenkaan haluta rajata tuloksia sen perusteella. Esimerkiksi seuraavassa graafihahmossa valitaan tietokannasta kaikki henkilöt, ja annetaan tuloksena kaikkien henkilöiden URI:t sekä sukunimi niille, joille sellainen on merkitty. Ilman OPTIONAL ilmaisua tulosjoukko rajautuisi pelkkiin henkilöihin, joilla on sukunimi.

```
SELECT ?person ?familyName WHERE {
    ?person a foaf:Person .
    OPTIONAL {
        ?person foaf:familyName ?familyName .
    }
}
```

SPARQL mahdollistaa myös vaihtoehtoisten ratkaisujen ilmaisemisen. Tässä työssä osaan CONSTRUCT-hahmoista on lisätty tällaisia vaihtoehtoisia ratkaisuja, jotta hahmo toimisi erilaisissa tietokannoissa. Esimerkiksi ilmaisulla UNION on mahdollista määrittää vaihtoehtoisia hahmoja. Lisäksi useamman vaihtoehtoisen ominaisuuden voi ilmaista erottamalla ne operaattorilla ”|”.

2.3 Julkaiseminen semanttisessa webissä

Tim Berners-Lee on esittänyt neljä periaatetta linkitetyn datan julkaisemiselle: asioiden nimeäminen URI-tunnisteilla, HTTP-URI-tunnisteiden käyttäminen, tunnisteisiin liittyvän tiedon tarjoaminen standardien avulla ja datan linkittämisen toisiin URI-tunnisteisiin [3]. Berners-Lee on esittänyt avoimen datan palveluiden arvioimisen viiden tähden mallin[8], joka kuvaa datan avoimuuden asteita [24]. Yhden tähden saasaseen riittää datan julkaiseminen webissä avoimella lisenssillä. Kaksi tähteä vaatii datan julkaisemisen rakenteisessa muodossa. Kolme tähteä ansaitsee käyttämällä

[8]https://5stardata.info/en/

avointa formaattia. Neljä tähteä saadakseen täytyy merkitä asioita URI:en avulla. Täydet viisi tähteä saa, kun edeltävien lisäksi linkittää datansa muuhun avoimeen dataan. Hyvönen ja kumppanit [32] ovat esittäneet tämän mallin laajentamista seitsemään tähteen. Kuusi tähteä saisi dokumentoimalla datan ja seitsemän tähteä vaatisi lisäksi datan oikeellisuuden validoimista.

Linkitetyn datan sovelluksien keskeinen piirre on niiden datalähtöisyys [24, s. 189-194]. RDF-muotoista linkitettyä dataa voi julkaista SPARQL-palvelupisteen avulla, josta sovellukset hakevat tietoa SPARQL-rajapinnan kautta. Hyvin toteutettu data mahdollistaa monipuolisten sovellusten luomisen helposti. Sovellukset voivat olla myös sellaisia, joita dataa luotaessa ei ole pystytty kuvittelemaan.

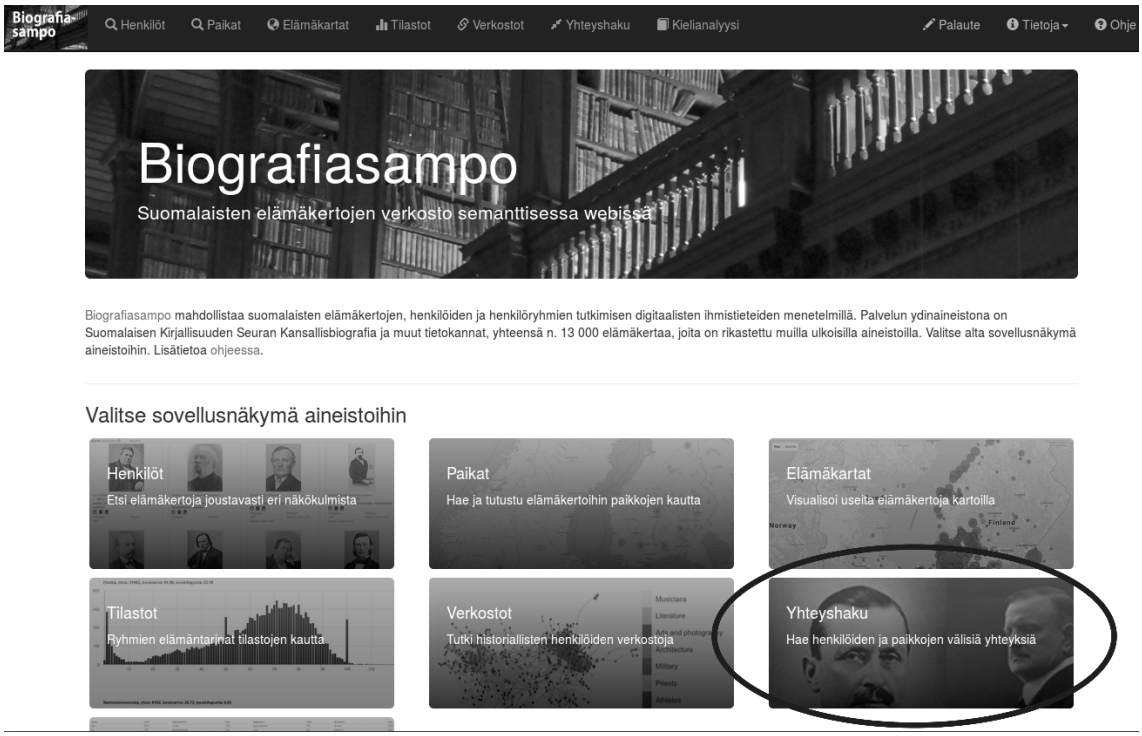
Kulttuurihistoria on erityisen hedelmällinen sovellusalue semanttisen webin sovelluksille, koska alaan liittyy paljon hyödyllisiä sanastoja ja yleisesti saatavilla olevia kokoelmatietoja [23, s. 121]. Tähän liittyen keskeinen sovellustyyppi ovat olleet semanttiset portaalit, jotka voivat koota yhteen dataa erilaisista lähteistä. Semanttinen web tarjoaa luontevan perustan tällaisten portaalien luomiseen ja ylläpitoon. Suurin haaste tällaisten portaalien luomisessa liittyy heteregoonisten aineistojen yhdistämiseen. Datan laadussa on myös usein puutteita.

Tässä työssä kehitetty demonstraattori julkaistiin osana Biografiasampo-sivustoa[9]. Biografiasampo on toteutettu Sampo-mallin [24, s. 199-202] mukaan. Kuvassa 2 näkyy Biografiasammon portaali, ja kuvaan on ympyröity Yhteyshaku-sovellukseen johtava linkki. Sampo-malli on sisällöntuotantoon tarkoitettu hajautettu malli semanttiselle portaalille, jossa eri sovellukset käyttävät yhteistä tietoinfrastruktuuria ja niille on luotu yhteinen portaali. Samojen tietojen perusteella voi luoda helposti useita erilaisia sovelluksia tai näkymiä. Eräs suosittu Sampo-mallilla toteutettu portaali on esimerkiksi Sotasampo[10], joka tarjoaa tietoa suomalaisesta sotahistoriasta toisessa maailmansodassa.

Biografiasampoon on esivaiheessa yhteyshaun lisäksi toteutettu näkymät esimerkiksi henkilöiden elämäkertojen visualisoimiseen kartalla sekä elämäkertojen kielianalyysiin. Elämäkertoja kartalla visualisoiva näkymä muistuttaa tietyssä mielessä tässä työssä toteutettavaa yhteyshaku-näkymää. Tässä elämäkartat-näkymässä voi fasettien avulla rajata henkilöitä ja nähdä kartalta, minkä tyyppisiä tapahtumia liittyy mihinkin paikkaan. Tämä näkymä eroaa yhteyshausta, ainakin ensimmäisessä vaiheessa toteutetussa muodossaan siinä, että näitä tapahtumia ei voi rajata pai-

[9] http://biografiasampo.fi/

[10] https://www.sotasampo.fi/fi/



Kuva 2: Biografiasammon Sampo-mallin mukainen portaali

kan mukaan. Yhteyksien sanallisten selitysten saaminen on myös rajatumpaa ja vaikeampaa. Lisäksi elämäkartat-näkymää ei ole rikastettu ulkopuolisista datajoukoista. Elämäkartat on visuaalisempi näkymä kuin yhteyshaku-näkymä. Molemmat tarjoavat erilaisen näkymän osin samaan tietoon.

2.4 Ontologiat

Tietotekniikassa ontologialla tarkoitetaan formaalia ja yhteisesti sovittua käsitteellistä kuvausta maailmasta [18][24, s. 133-134]. Ontologia määrittelee käsitteet, joita tarvitaan tietyn sovellusalan kuvaamiseen. Semanttiseen webiin liittyen ontologialla voidaan tarkoittaa sekä metadatamalleja että käsitteitä määritteleviä aiheontologioita. Kun tieto on kuvattu ontologiassa määriteltyjen käsitteiden avulla, on sitä mahdollista tulkita tietoteknisin menetelmin ja tehdä päätelmiä. Klassisen esimerkin mukaan, jos tiedetään ihmisten olevan eläimiä ja eläimien olevan kuolevaisia, voidaan päätellä ihmisten olevan kuolevaisia. Yhtenäisten ontologioiden käyttö mahdollistaa eri lähteiden datan helpon yhdistämisen. Ontologiat voidaan myös *sillata* (mapping)

toisiinsa, jolloin ilmaistaan tietyn käsitteen yhdessä ontologiassa vastaavan käsitettä toisessa ontologiassa.

Suomalaisia ontologioita on kehitetty FinnONTO-projektissa [33]. Ihannetilanteessa entiteettejä, tämän työn kannalta erityisesti henkilöitä ja paikkoja, on kuvattu datassa yhtenäisten ontologioiden avulla tai ne on sillattu julkisiin ontologioihin. Usein näin ei ole. Käytettyä henkilö-ontologiaa ei välttämättä ole sillattu mihinkään ulkopuolelle tai paikkoja ei edes ole kuvattu entiteetteinä vaan pelkkinä literaaliarvoina. Tällaisessa tapauksessa siltaukset on tarpeen tehdä erikseen kiinnostavilta osin.

Entiteettien disambiguointiin liittyy paljon ongelmia, esimerkiksi paikat muuttuvat ajan kuluessa [35]. Esimerkiksi Helsinki on tarkoittanut eri aikoina hiukan eri asioita. Tässä työssä paikkojen muuttumiseen ajan myötä liittyvät ongelmat suurimmaksi osaksi sivuutetaan, mutta koska käytetään historiallista materiaalia, on hyvä ymmärtää, että jotkut tulokset voivat olla omituisia tämän vuoksi.

Ontologioiden heterogeenisuus on ongelma tämän työn kannalta. Tieto on yleensä kuvattu eri tavalla eri paikoissa. Kulttuuriperinnön alan toimijoista saattaa osa jopa suhtautua negatiivisesti ajatukseen yhteisen kontrolloidun sanaston käytöstä [10]. Ongelma on sekä henkilöitä ja paikkoja kuvaavien aihe-ontologioiden että metadatamallien heterogeenisuus. Tämän työn kannalta asiaa helpottaa se, että Biografiasammon dataan on valmiiksi tehty siltaukset useiden Suomen kulttuurihistoriaan liittyvien aineistojen käyttämiin henkilö-ontologioihin. Haasteita aiheuttaa myös se, että muistiorganisaatioiden, kuten museoiden ja kirjastojen, käyttämät katalogisointijärjestelmät eivät usein tue ontologioihin perustuvaa annotointia [23, s. 121]. Siten asioita ei välttämättä ole lainkaan kuvattu ontologioiden avulla, mikä lisää työtä eri aineistojen yhteensovittamisessa. Kun tällainen yhdistäminen tehdään automaattisesti, lisää se myös virheiden mahdollisuutta.

Eräs tämän työn tavoite on arvioida, kuinka helppoa on muodostaa yleisiä muotoja, joilla löytää tietämykseen perustuvan arvion mukaan olennaista kulttuurihistoriaan liittyvää tietoa tietokannoista. Tuntuu ilmeiseltä, että samaa ontologiaa käyttävien tietokantojen välillä voi käyttää samaa muotoa tiedonhakuun. Tässä tapauksessa kysymykseksi jää vain se, kuinka helppo muotoja on muodostaa ja kuinka paljon erilaisia ontologioita on käytössä. Mielenkiintoisempi ja vaikeampi kysymys on, kuinka helppo samoja muotoja on käyttää eri tietokannoissa, jos ne käyttävät eri ontologioita metadatamalleina. Vaikka ontologiat poikkeaisivat toisistaan käyttämiensä ominaisuuksien ja luokkien osalta, ne saattavat noudattaa jonkinlaisia yhteneviä suunnitteluperiaatteita. Eli esimerkiksi ontologiassa voi olla jonkinlainen luokka ku-

vaamaan kirjaa, toinen luokka kuvaa henkilöä ja kirja-luokan instanssilla voi olla ominaisuus, joka tarkoittaa kirjan kirjoittajaa ja viittaa henkilö-luokan instanssiin. Toinen ontologia saattaisi käyttää eri nimisiä luokkia ja ominaisuuksia, mutta jos niiden merkitys ja suhde on samanlainen, voidaan yhdellä sopivasti muodostetulla SPARQL-muodolla hakea sama tieto molempia ontologioita käyttävistä tietokannoista. SPARQL mahdollistaa esimerkiksi vaihtoehtoisten ominaisuuksien merkitsemisen, mikä mahdollistaa tämän. Formaalisti määriteltynä tällaisia ontologioiden suunnitteluun liittyviä yleisiä malleja kutsutaan *käsitteellisiksi ontologian suunnittelumalleiksi* (Conceptual (or Content) Ontology Design Pattern) [15, 16]. Tämän työn kannalta ei ole kuitenkaan oleellista, onko ontologioiden muodon yhdenmukaisuus syntynyt formaalin suunnitteluprosessin tuloksena vai pelkästään sattumalta, kun tietynlainen käsitteellinen kuvaus ilmiöstä on niin luonteva, että sitä käytetään useissa ontologioissa toisistaan riippumatta.

2.5 Kiinnostavuus ja serenpiditeetti

Frawleyn ja kumppaneiden mukaan [14] tietämyksen muodostaminen on prosessi jossa datasta löydetään epätriviaalilla tavalla aikaisemmin tuntematonta ja potentiaalisesti hyödyllistä tietoa. *Kaavalla* (pattern) tarkoitetaan tässä yhteydessä tietokannasta louhittua väitettä asioiden välisistä suhteista. Riittävän varmaa ja käyttäjän näkökulmasta kiinnostavaa kaavaa voi kutsua tietämykseksi.

Eri tietokannoissa on saatavilla valtava määrä tietoa ja erilaisia yhteyksiä henkilöiden ja paikkojen välille. On kuitenkin selvää, että eräät yhteydet ovat mielenkiintoisempia kuin toiset. Esimerkiksi jos J. V. Snellmanin lapsenlapsenlapsi on kerran käynyt Tampereella, se tietyssä mielessä loisi yhteyden Snellmanin ja Tampereen välille, ja sellaisen yhteyden voisi teoriassa avoimesta datasta löytää. Tällainen yhteys tuskin olisi kuitenkaan erityisen mielenkiintoinen. Jos käyttäjälle näytetään paljon tällaisia turhia yhteyksiä, hukkuvat mielenkiintoiset yhteydet muiden joukkoon. Kiinnostavat yhteydet olisi tärkeää pystyä tunnistamaan. Siksi on tärkeä myös ymmärtää, mitä kiinnostavuudella tarkoitetaan ja miten sitä voi mitata.

Silberschatz ja Tuzhilin jakavat artikkelissaan [65] kiinnostavuuden mittaamiseen käytetyt metriikat objektiivista ja subjektiivista kiinnostavuutta mittaaviin metriikoihin. Objektiiviset metriikat mittaavat vain datan ominaisuuksia, kun taas subjektiiviset metriikat riippuvat tulkitsijasta. Kaksi ominaisuutta, jotka tekevät tiedon käyttäjän kannalta kiinnostavaksi, ovat *odottamattomuus* ja *hyödyllisyys* (actionability). Tieto on käyttäjälle kiinnostava, jos se on jotenkin odottamaton eli yllättävä.

Toisaalta tieto on käyttäjälle kiinnostava, jos hän pystyy tekemään sillä tiedolla jotain. Tieto sitä kiinnostavampaa mitä ”hyödyllisempää” tämä jokin on.

Geng ja Hamilton erottavat yhdeksän kiinnostavuuden kriteeriä: *suppeus* (conciseness), *kattavuus* (generality/coverage), *luotettavuus* (reliability), *erikoisuus* (peculiarity), *monipuolisuus* (diversity), *uutuus* (novelty), *yllättävyys* (surprisingness), *hyödyllisyys* (utility), *käyttökelpoisuus* (actionability/applicability) [17]. He jakavat nämä yhdeksän kriteeriä kolmeen luokkaan: objektiivisiin, subjektiivisiin ja semantiikkaan perustuviin kriteereihin. Semanttinen metriikka tarkastelee Gengin ja Hamiltonin mukaan kaavojen semantiikka ja selityksiä. He erottavat tällaiset metriikat subjektiivisista, mutta toisenlaisissa luokitteluissa tällaiset metriikat saatetaan laskea subjektiivisten metriikoiden alalajiksi.

Ken McGarry erottaa artikkelissaan [54] kuusi kiinnostavuuden metriikkaa, jotka hän jakaa objektiivisiin ja subjektiivisiin. Objektiivisia metriikoita ovat hänen mukaansa *kattavuus* (coverage), *tuki* (support) ja *tarkkuus* (accuracy). Subjektiivisia metriikoita ovat McGarryn mukaan *odottamattomuus* (unexpectedness), *käyttökelpoisuus* (actionability) ja *uutuus* (novelty).

Tässä työssä on yksinkertaisuuden vuoksi jaettu kiinnostavuuden metriikat objektiivisiin ja subjektiivisiin. Tämän työn kannalta oleellisia eivät ole objektiiviset metriikat, sillä tämän työn lähestymistavassa datan sisäisellä rakenteella ei ole kiinnostavuuden kannalta merkitystä. Tässä työssä oleellisia ovat kiinnostavuuden subjektiiviset ominaisuudet.

Subjektiivisista metriikoista yllättävyyden arvioimiseen on kehitetty joitain tapoja [44]. Tässä työssä ei kuitenkaan pyritty arvioimaan yllättävyyttä formaalisti. Hypoteesina on, että käyttäjä pystyy fasettihaun avulla rajaamaan tuloksia riittävästi itselleen yllättävään suuntaan. Saattaa silti olla asioita, jotka ovat niin odotettuja että ne on ovat useimmiten ilmeisen epäkiinnostavia. Esimerkki tästä voisi olla suomalaisten henkilöiden yhteydet paikkaan ”Suomi”. Tällaisia yhteyksiä on pyritty jättämään pois.

Kiinnostavuuden ominaisuuksista käyttökelpoisuus on asia, johon on tarpeen muodostaa tässä työssä jonkinlainen kanta. Haettavat yhteydet valittiin niiden oletetun käyttökelpoisuuden perusteella. Ei ole kuitenkaan ilmeistä, mitä ”käyttökelpoisuus” tarkoittaa kulttuurihistorian alalla. Gengin ja Hamiltonin [17] mukaan kaava on käyttökelpoinen, jos se mahdollistaa päätöksenteon tulevista toimista. Tällainen määritelmä on lähinnä tekninen eikä erityisen hyvin sovellu kulttuurihistorian alalle. Saksalainen filosofi Jürgen Habermas erottaa kolme erilaista *tiedonintressiä*: *tekni-*

sen-, *käytännöllisen-* ja *emansipatorisen* tiedonintressin [2]. Tekninen tiedonintressi liittyy Habermasin mukaan työhön. Se koskee tietoa, jota hankitaan jotta pystyttäisiin manipuloimaan luontoa. Tekninen tiedonintressi on luonnollinen esimerkiksi insinööritieteille. Käytännöllinen tiedonintressi taas koskee tarvetta ymmärtää esimerkiksi toisia ihmisiä ja perinnettä. Tällainen tiedonintressi on luonteva esimerkiksi humanistisille tieteille ja historialle. Emansipatorinen tiedonintressi liittyy Habermasin mukaan tarpeeseen vapautua perinteen kahleista. Se koskee erityisesti yhteiskuntatieteitä. Tämän Habermasin jaottelun hengessä tässä työssä ei pyritä löytämään yhteyksiä, joista saatua tietoa voi käyttää jotenkin hyödyllisesti, vaan tavoitteena on löytää yhteyksiä, jotka auttavat käyttäjää ymmärtämään Suomen kulttuurihistoriaa.

Mitään tarkkaa sääntöä sille, millainen yhteys tämän työn kannalta koetaan kiinnostavaksi, on vaikea määritellä. Yhteyksien tyypit valittiin sen mukaan, mitä on helposti saatavilla sekä subjektiivisen kiinnostavuusarvion mukaan. Ohjenuorana kiinnostavuuden arviointiin on tässä työssä käytetty ymmärryksen parantamista. Voidaan ehkä ajatella, että yhteydet jotka kertovat henkilön vaikutuksesta paikkaan tai paikan vaikutuksesta henkilöön, parantavat ymmärrystä henkilöstä tai paikasta. Mitä suurempaa vaikutusta yhteys vihjaa, sitä kiinnostavampana yhteyttä voi pitää. Esimerkiksi jos henkilö on vastaanottanut kirjeen tietystä paikasta, voi hänelle tulkita olevan jotain yhteyksiä kyseiseen paikkaan. Siten kyseinen paikka on ehkä vaikuttanut häneen ja hän kyseiseen paikkaan. Vaikutusta ei kuitenkaan voi pelkästään tällä perusteella olettaa kovin suureksi. Jos henkilö on itse lähettänyt kirjeen jostain paikasta, voi hänen tulkita vähintään käyneen kyseisessä paikassa. Tämä on ilmeisellä tavalla merkittävämpi yhteys kuin vain kirjeen vastaanottaminen jostain, mutta ei välttämättä tarkoita merkittävää vaikutusta henkilön ja paikan välille. Jos henkilö on syntynyt jossain paikassa, voi paikan vaikutusta häneen pitää jo paljon suurempana. Lisäksi paikoille on usein suuri merkitys, jos jokin suurmies on syntynyt siellä. Jos taas henkilö on kuvannut paikkaa taiteessaan voi paikan tulkita olleen hänelle merkityksellinen jollain tapaa, ja mahdollisesti tällä on ollut myös vaikutuksia kyseiseen paikkaan.

Seredendipisyys eli *seredipiteetti* (serendipity) liittyy läheisesti tiedon etsintään ja kiinnostavuuteen. Serenpisyydellä tarkoitetaan sellaisen hyödyllisen tiedon onnekasta löytämistä, jota ei oltu suoraan etsimässä [70]. Tässä työssä toteutetun sovelluksen on tarkoitus mahdollistaa serendipisyyttä tiedonhaussa. Serendipisyydelle ei ole tarkkaa yleisesti hyväksyttyä määritelmää. Sitä on myös vaikea mitata, koska se on määritelmän mukaisesti odottamatonta. Serendipisyyttä voi silti mahdollistaa.

Järjestelmä, joka mahdollistaa tiedon etsimisen helposti ilman liian tiukkoja ennalta määrättyjä rajoitteita, luultavasti edistää serenpidisyyttä. Lisäksi asiantunteva ja avoimen mielen omaava ihminen kykenee tekemään serendipisiä löytöjä helpommin.

2.6 Tiedonhaku

Ihmiskäyttäjän näkökulmasta *haku* ja *selailu* ovat webin pääasialliset käyttötavat tiedon etsimisessä [24, s. 32-47]. Haun laatua voidaan mitata *tarkkuudella* ja *saannilla*. Tarkkuus mittaa sitä kuinka suuri osa hakutuloksista on oikeita, kun taas saanti mittaa sitä, kuinka moni mahdollisista hakutuloksista löydettiin. Yhteydet ovat yksi asia jota haku voi koskea, ja tässä työssä toteutetussa yhteyshaussa onkin kyse hausta. Toisaalta tässä työssä toteutetun sovelluksen on tarkoitus mahdollistaa käyttäjälle yhteyksien selailu ja suodattaminen fasettien avulla. Lisäksi yhteyshaku toimii tässä osana laajempaa Biografiasampo-kokonaisuutta. Tarkoitus on tarjota käyttäjälle linkkejä henkilöiden, paikkojen ja esimerkiksi taideteosten tarkempia tietoja sisältäville sivuille. Siten kysymys on osin myös selailusta.

Tutkiva haku (exploratory search) tarkoittaa tiedon hakua jossa ei tarkkaan tiedetä mitä ollaan hakemassa [52, 1]. Tiettyyn asiaan liittyvän faktatiedon etsimistä voi kutsua *etsiväksi hauksi* (look-up search). Sen sijaan tutkivassa haussa pyritään ymmärtämään jotain asiaa. Haun tavoite on epäselvä ja avoin. Lisäksi ei ole selvää, mitä tarkalleen etsitään ja milloin se on tullut onnistuneesti löydetyksi. Tässä työssä toteutetussa järjestelmässä on kyse tutkivasta hausta, koska käyttäjän tavoitteena on hankkia epämääräisesti määriteltävissä oleva ymmärrys historiasta henkilöihin tai paikkoihin liittyen.

Tuntuu luontevalta yhdistää etsivä haku Habermasin tekniseen tiedonintressiin ja tutkiva haku käytännölliseen tiedonintressiin. Yhteys tuntuu niin luontevalta, että herää kysymys onko Habermasin kolmannelle tiedonintressille, eli emansipatoriselle tiedonintressille, vastaavaa hakutyyppiä. Tällainen voisi olla haku, jossa pyritään löytämään yhteiskunnassa vallitsevia sortavia lainalaisuuksia. Tietyssä mielessä Biografiasampoa on mahdollista käyttää tällaiseen hakuun. Esimerkiksi Biografiasammossa on kielianalyysi-näkymässä[11] mahdollisuus vertailla sitä, miten eri tavalla naisista ja miehistä on käytetty erilaista kieltä elämäkerroissa, ja löytää siten mahdollisia yhteiskunnan piilotettuja valtarakenteita.

Fasettihaku [19, 67, 21] tarkoittaa hakua, jossa hakutulosta rajataan askel kerrallaan

[11] http://biografiasampo.fi/kielianalyysi

niiden hierarkisesti järjestettyjen ominaisuuksien arvojen, eli *fasettien*, perusteella. Fasettihaku mahdollistaa monimutkaisten hakujen toteuttamisen tavalla, jota on yksinkertaista käyttää myös maallikolle. Hakua on mahdollista rajata yksi ominaisuus kerrallaan. Käyttäjä voi jokaisen valinnan jälkeen arvioida hakuaan tarkemmin. Rajaamalla hakua eri suodattimilla, voidaan luoda valtava määrä erilaisia kyselyitä. Tyypillinen sovellusala jossa tavallinen kansalainen kohtaa fasettihaun ovat verkkokaupat. Erityisen hyvin fasettihaku soveltuu tutkivan haun toteuttamiseen [53].

Jos käyttäjä tietää tarkkaan mitä hän on hakemassa, on tavallinen sanahaku usein parempi, mutta fasettihaku soveltuu hyvin tilanteisiin joissa käyttäjä ei pysty tarkkaan määrittämään tiedontarvettaan heti [11]. Jos tiedon tarve on yksinkertainen ja helposti määriteltävissä, voi monimutkainen käyttöliittymä fasetteineen vain haitata käyttäjän kokemusta turhaan. Fasettihaku saattaa myös vähentää tarvetta haun kohtein asettamiseen kiinnostavuusjärjestykseen automaattisesti. Fasettihaun avulla käyttäjä pystyy kohdistamaan hakua juuri itseään kiinnostaviin kohteisiin.

Fasettihaku sopii hyvin semanttisen webin ja linkitetyn datan sovelluksiin [31]. Linkitetty data käyttää hyväkseen ontologioita, joiden soveltaminen faseteiksi on usein luontevaa.

Kules ja kumppanit ovat käsitelleet artikkelissaan [38] sitä, miten faseteja käytetään tutkivassa haussa. Heidän mukaansa käyttäjän hakuun käyttämästä ajasta merkittävä osa kului fasettien tarkastelemiseen. Joissain haun vaiheissa fasetit ovat käyttäjälle jopa keskeisempi kuin itse tulokset. Käyttävät myös ohjaavat hakuaan fasettien sisällön perusteella.

2.7 Yhteyshaku

Semanttisten assosiaatioiden tunnistaminen on konsepti, jota Sheth ja kumppanit [64] ovat soveltaneet kansallisen turvallisuuden alalla. Kurki & Hyvönen [39] kutsuvat tätä *semanttiseksi yhteyshauksi* (relational semantic search) ja ovat soveltaneet sitä kulttuurihistorian alaan. Semanttiset assosiaatiot ovat entiteettien, tapahtumien ja käsitteiden välillä olevia monimutkaisia ja merkityksellisiä yhteyksiä. Semanttisella polulla tarkoitetaan tässä RDF-graafissa kahden entiteetin välillä olevaa katkeamatonta polkua, joka voi muodostua mielivaltaisesta määrästä entiteettejä ja ominaisuuksia. Semanttisessa yhteyshaussa käyttäjä haluaa saada vastaukseen kysymykseen, joka on tyyppiä ”Miten A liittyy B:hen”. Kun käyttäjä valitsee semanttisen yhteyden päätepisteet A ja B hänelle palautetaan esitys semanttisista poluista

A:n ja B:n välillä. Keskeinen ongelma tällaisessa yhteyshaussa on epäkiinnostavien yhteyksien eliminointi siten, että datan sisältämät arvaamattomat kiinnostavat yhteydet tulevat havaittaviksi [47].

Cheng ja kumppanit [7] ovat hiljattain toteuttaneet laajan katsauksen erilaisiin menetelmiin semanttisten assosiaatioiden asettamisesta kiinnostavuusjärjestykseen. He erottavat viisi datan ominaisuutta, joiden avulla yhteyden kiinnostavuutta on yleensä arvioitu: polun pituuden, frekvenssin, keskeisyyden, informatiivisuuden ja spesifisyyden. Lisäksi he ehdottavat metriikaksi homogeenisuuttä. Cheng ja kumppanit suorittivat tekniikoiden vertailun käyttämällä asiantuntija-arvioita. Tämän perusteella he pitävät polun pituutta ja homogeenisuutta parhaina tekniikoina. Keskeisyys, informatiivisuus ja spesifisyys eivät heidän mukaansa ole toimivia tekniikoita ainakaan siinä muodossa, missä niitä on yleensä käytetty. Sekä liian harvinaiset että liian yleiset asiat vaikuttavat epäkiinnostavilta. Parasta olisi löytää jonkinlainen keskitie.

2.8 Esimerkkejä yhteyshakusovelluksista

Tässä aliluvussa esitellään esimerkkejä linkitettyä dataa hyödyntävistä yhteyshakusovelluksista. Tarkoitus ei ole käydä kattavasti läpi kaikkia mahdollisia esimerkkejä, vaan antaa karkea kuva siitä, millaisia sovelluksia on aikaisemmin toteutettu.

Keskeinen esimerkki yhteyshausta on Shethin ja kumppaneiden artikkeli [64], jossa he soveltavat yhteyshakua kansallisen turvallisuuden alaan ja esittelevät monia yhteyshakuun liittyviä peruskäsitteitä. Shethin ja kumppaneiden kehittämä prototyyppi seuloo semanttisen yhteyshaun perusteella automaattisesti mahdollisesti vaarallisia henkilöitä lentokoneen matkustajista. Heidän mukaansa kansallinen turvallisuus on ala, jolla semanttisten assosiaatioiden löytämisestä on erityisen paljon hyötyä. Kyseinen artikkeli on julkaistu vuonna 2005, joten on syytä olettaa, että tällainen arvio liittyy osin vuoden 2001 jälkeiseen yleiseen kansallisen turvallisuuden korostamiseen Yhdysvalloissa.

Ehkä eniten tässä työssä toteutettavaa yhteyshakua muistuttava sovellus luonteeltaan ja sovellusalaltaan on Kulttuurisampo-portaalin[12] osana toteutettu yhteyshaku-palvelu, jossa voi hakea yhteyksiä taiteilijoiden välillä [30, 39, 29]. Kyseinen sovellus antaa ketjun ihmisten välisiä yhteyksiä jostain taiteilijasta toiseen taitelijaan. Yhteyksille annetaan sanalliset luonnollisen kielen selitykset. Lisäksi on mahdollista

[12]http://www.kulttuurisampo.fi/

tarkastella graafista esitystä yksittäisen taiteilijan yhteysverkosta. Sovellus perustuu ULAN-sanastoon[13], joka kuvaa taiteilijoiden nimiä ja muuta tietoa, kuten yhteyksiä toisiin taiteilijoihin. Tämä sovellus ei mahdollista eri tyyppisten yhteyksien hakemista. Käytännössä sovellus näyttää kahta henkilöä yhdistävän opettajalinjan.

RelFinder [20, 21, 46, 45] on esimerkki yhteyksien visualisointityökalusta. RelFinder on tarkoitettu yleiskäyttöiseksi lähestymistavaksi, mutta sen julkinen demo käyttää datan lähteenä DBpediaa[14]. RelFinder ottaa syötteekseen entiteettejä ja laskee näiden entiteettien välisiä yhteyksiä. Käyttäjälle näytetään visualisointi kahden entiteetin välisistä poluista. Käyttäjän on mahdollista rajata näkyviin vain tietynlaisia polkuja käyttämällä suodattimia. Tätä voi pitää fasettihaun soveltamisena, vaikkakin hyvin rajatulla tavalla. RelFinder järjestelmälle on myös ehdotettu parannuksia [63].

Explass on Chengin ja kumppaneiden toteuttama prototyyppi tutkivalle haulle assosiaatioiden eli yhteyksien löytämiseen [8]. Explass soveltaa osin fasettihakua, joten se on tämän työn kannalta mielenkiintoinen. Myös Explass on yleiseen tapaukseen tarkoitettu järjestelmä, jonka demo käyttää DBpediaa. Käyttäjä kirjoittaa kahden entiteetin nimet ja järjestelmä hakee niille tietyn määrän yhteyksiä. Vain kiinnostavimmiksi arvioidut yhteydet näytetään. Kiinnostavuuden kriteerinä käytetään tässä kaavojen frekvenssiä, informatiivisuutta ja vähäistä päällekkäisyyttä. Kaavan frekvenssi viittaa tässä sen relevanttiuteen suhteessa kyselyn kontekstiin. Informatiivisuus riippuu luokkien ja yhteyksien yleisyydestä, eli harvinaisempia luokkia ja yhteyksiä pidetään informatiivisempina. Vähäinen päällekkäisyys taas tarkoittaa sitä, että hyvin paljon toisiaan muistuttavista yhteyksistä pyritään näyttämään vain yksi. Käyttäjälle täytyy tehdä haku kahden entiteetin välille niiden nimien perusteella. Sen jälkeen Explass luo fasetit yhteyksien tyypeille ja niihin liittyville entiteeteille, joiden avulla voi tarkentaa hakua haluamaansa suuntaan. Samoilta tutkijoilta on myös toinen käytännössä hyvin samanlainen yhteyshaku-sovellus nimeltä RelClus [71].

RECAP [58, 59, 13] on sovellus, joka visualisoi yksinkertaisella tavalla paremmuusjärjestykseen sijoitettuja selityksiä syötteenä annetun kahden entiteetin välisille yhteyksille. Lisäksi RECAP tarjoaa käyttäjälle SPARQL kyselyn, jolla voi etsiä entiteettejä, jotka liittyvät syötteenä annettuihin entiteetteihin.

Hiljattain esitelty Tartarin ja Hoganin [66] WiSP on ratkaisu mielenkiintoisten se-

[13]http://www.getty.edu/research/tools/vocabularies/ulan/
[14]https://wiki.dbpedia.org/

manttisten polkujen löytämiseen kahden asian välillä. Tässä ratkaisussa mielenkiintoisimmat yhteydet valitaan painotetun lyhyimmän polun perusteella.

Voskarides ja kumppanit ovat käsitellet luonnollisen kielen selityksen antamista kahden entiteetin väliselle yhteydelle [69]. He pyrkivät lähestymistavassaan poimimaan tekstistä koneoppimisen menetelmin selityksiä annetulle yhteydelle ja valitsemaan niistä parhaan. Yhteys itsessään on siis annettu graafissa ja sen selitys poimitaan tekstistä.

Mainitsemisen arvoinen on lisäksi Lehmannin ja kumppaneiden [41] esittelemä DBpedian yhteyshaku, joka näyttää syötteenä annettavan kahden entiteetin välisen semanttisen yhteyden polun. Myös Rex [12] on järjestelmä entiteettien välisten yhteyksien selittämiselle ja sijoittamiselle paremmuusjärjestykseen. On hyvä myös todeta, että SPARQL ei suoraan tue entiteettien välisten yhteyspolkujen löytämistä yksinkertaisella tavalla, mutta siihen on esitetty laajennuksia, jotka mahdollisesti helpottavat asiaa [36].

Miao ja kumppanit [55] ovat jakaneet yhteyksien visualisoimiseen käytetyt tavat kahteen luokkaan: graafiin perustuviin ja listaan perustuviin tapoihin. RelFinder ja RECAP ovat esimerkkejä graafiin perustuvista visualisoinnin tavoista. Explass ja Rex taas edustavat listaan perustuvaa tapaa. Miaon ja kumppaneiden mukaan bisimulaatioon perustuva tapa olisi kolmas mahdollinen. Siinä graafiin perustuva visualisointi otetaan lähtökohdaksi, mutta graafia tiivistetään ja yksinkertaistetaan algoritmisesti, jotta oleelliset asiat on helpompi nähdä.

Tietämykseen perustuvaa kulttuurihistorian henkilöiden ja paikkojen välisten yhteyksien etsimistä ovat käyttäneet hyödyksi ainakin Kaminskas ja kumppanit [34]. He ovat hakeneet musiikin suosittelujärjestelmää varten yhteyksiä muusikoiden ja paikkojen välillä. Kyseessä on siis suosittelujärjestelmä, mutta sen yhtenä vaiheena muodostetaan graafi kuvaamaan henkilöiden ja paikkojen välisiä yhteyksiä. Muusikoiden ja paikkojen välisten yhteyksien perusteella järjestelmä suosittelee tiettyyn paikkaan sopivaa musiikkia.

Yhteenvetona voi todeta, että täysin tässä työssä toteutetun järjestelmän kaltaista sovellusta yhteyshakuun ei ilmeisesti ole toteutettu. Sen sijaan on useita sovelluksia, jotka ovat osin samankaltaisia. Järjestelmä muistuttaa visualisoinniltaan listaan perustuvia yhteyksien visualisointijärjestelmiä. Yhteyksien tarkempaan suodattamiseen on käytetty fasetteja aikaisemmissa sovelluksissa kuten Explass ja RelFinder, mutta ilmeisesti sovellusta, joka mahdollistaa yhteyksien hakemisen heti fasettien avulla ei ole aikaisemmin toteutettu. Vastaavaa laajasti eri lähteistä kulttuurihis-

toriaan liittyviä yhteyksiä hakevaa sovellusta ei ilmeisesti myöskään ole toteutettu aikaisemmin.

3 Datan lähteet

3.1 Biografiasampo

Biografiasampo[15] on hanke, jonka työnimenä on ollut myös Semanttinen kansallisbiografia [26, 42, 43, 27]. Hankkeen tarkoituksena on muuntaa Suomalaisten elämäkertoja linkitetyksi dataksi, rikastaa dataa eri lähteistä ja julkaista web-pohjaisia palveluita, jotka perustuvat biografioiden dataan. Biografiasammon datan lähteenä on käytetty Suomalaisen Kirjallisuuden Seuran Biografiakeskuksen luomia elämäkertoja. Taulukossa 2 on listattu Biografiasammossa käytetyt elämäkerrat tietokannoittain. Yhteensä elämäkertoja on noin 13000. Mukana on suomalaisten lisäksi myös eräitä Suomen historian kannalta merkittäviä ulkomaalaisia henkilöitä. Suurin osa biografioista kuvaa yksittäisiä henkilöitä, mutta osa kuvaa myös sukuja tai esimerkiksi pariskuntia. Mukana on myös joitain kansallisesti merkittäviä fiktiivisiä hahmoja kuten esimerkiksi Väinämöinen. Biografiasammon RDF-mallissa jokaiselle biografialle on oma henkilö-luokan instanssinsa, eli esimerkiksi suvut ovat tässä mielessä henkilöitä tietomallissa.

Kokoelma	**Koko**
Kansallisbiografia	6478
Talouselämän vaikuttajat	2235
Kenraalit ja amiraalit 1809–1917	481
Turun hiippakunnan paimenmuistio 1554–1721	2716
Suomen papisto 1800–1920	1234
Yhteensä elämäkertoja	13144

Taulukko 2: SKS:n Biografiakeskuksen pienoiselämäkertojen tietokannat

Henkilöiden elämäkertojen kuvaamiseen linkitettynä datana Biografiasampo käyttää tapahtumaperusteista metadatamallia nimeltä Bio CRM [68]. Bio CRM on elämäkertojen kuvaamiseen tarkoitettu laajennus CIDOC CRM [9] metadatamallille, joka on tapahtumapohjainen malli kulttuuriperintöä koskevan tiedon kuvaamiseen. Tähän työhön on Biografiasammon datasta louhittu yhteyksiä, joissa henkilön

[15]https://seco.cs.aalto.fi/projects/biografiasampo/

ura liittyy paikkaan, ja yhteyksiä, joissa henkilö on saanut johonkin paikkaan liittyvän kunnianosoituksen.

3.2 HISTO-ontologia

HISTO-ontologia[16] [25, 28] eli Historia-ontologia, on Suomen historian tapahtumia kuvaava ontologia, joka perustuu Agricola-verkoston[17] keräämiin tietoihin Suomen historiasta. Ontologia sisältää 1198 tapahtumaa, joille annetaan metadataa kuten tapahtumaan osallistuneet henkilöt ja siihen liittyvät paikat. Tässä työssä HISTO-ontologiasta louhittiin yhteyksiä joissa henkilö on osallistunut historialliseen tapahtumaan, joka liittyy paikkaan.

HISTO-ontologia käyttää tiedon kuvaamiseen CIDOC CRM -metadatamallia. CIDOC CRM nimiavaruuden etuliite määritellään tässä työssä seuraavasti:

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/>.
```

Historiallisia tapahtumia kuvataan luokalla

```
crm:E5_Event.
```

Tapahtuman ominaisuutena on muun muassa siihen osallistuvia henkilöitä kuvaava ominaisuus

```
crm:P11_had_participant
```

ja tapahtumapaikkoja kuvaava ominaisuus

```
crm:P7_took_place_at.
```

Henkilöillä on oma ontologiansa, jossa henkilöt ovat luokan

```
crm:E21_Person
```

yksilöitä. Biografiasampo sisältää siltauksen HISTO-ontologian henkilöihin, joten sitä ei tarvinnut erikseen tehdä. Paikkoja kuvataan eri ontologioihin liittyvillä resursseilla, joista osan URI:t ovat jo vanhentuneet. Suuri osa käytetyistä paikka-entiteeteistä liittyy Suomen ajalliseen paikka-ontologiaan (SAPO)[18], joka kuvaa historiallisia paikkoja.

[16] https://seco.cs.aalto.fi/ontologies/histo/
[17] https://agricolaverkko.fi/
[18] https://seco.cs.aalto.fi/ontologies/sapo/

3.3 Fennica

Fennica[19] on Kansalliskirjaston avoin tietokanta, johon on kerätty suomalaisesta julkaisutuotannosta vuodesta 1488 eteenpäin. Tietokannassa on lähes miljoonan julkaisun tiedot. Aineisto on julkaistu avoimesti linkitetyn datan muodossa, ja se on käytettävissä SPARQL-kyselyrajapinnalla ja ladattavissa RDF-muodossa. Tähän työhön Fennicasta on louhittu yhteyksiä, joissa henkilö on merkitty tekijäksi kirjalliseen teokseen, joka kuvaa jotain paikkaa.

Fennica käyttää datan kuvaamiseen schema.org-sanastoa[20]. Kirjallisia töitä kuvataan luovien töiden kuvaamiseen tarkoitetulla luokalla

```
schema:CreaticeWork.
```

Ominaisuuksina tämän luokan yksilöillä on muun muassa

```
schema:author,
```

joka kuvaa työn tekijää, sekä

```
schema:about,
```

joka kuvaa työn aihetta. Fennica käyttää omaa henkilö-ontologiaansa, johan Biografiasammosta on tehty siltaus. Aiheita kuvaavina resursseina käytetään YSO:n eli Yleisen suomalaisen ontologian[21] käsitteitä. Tähän ryhmään kuuluvat kirjan aiheena olevat paikat, joita kuvataan YSO-paikat ontologian käsitteillä. Fennica sisältää myös dataa kirjojen julkaisupaikoista. Tältä osin data on kuitenkin ilmeisesti vielä kesken, ja julkaisupaikkoja kuvataan vain literaaliarvoilla.

3.4 Kirjasampo

Kirjasampo [22, 48, 49, 51, 50] on suomalaista fiktiokirjallisuutta kuvaava linkitetyn datan datajoukko, johon perustuvaa web-palvelua[22] ylläpitävät yleiset kirjastot. Datajoukossa on metadataa esimerkiksi romaanien kirjoittajista ja tapahtumapaikoista. Tässä työssä on käytetty lähteenä ldf.fi-palvelussa julkaistua avointa data-

[19] https://www.kansalliskirjasto.fi/fi/palvelut/metadatan-muunto-ja-valityspalvelut/avoin-data
[20] https://schema.org/
[21] https://finto.fi/yso/fi/
[22] https://www.kirjasampo.fi/

joukkoa[23]. Kirjastojen ylläpitämän Kirjasampo-palvelun URI:t poikkeavat alkuosaltaan avoimen datasetin käyttämistä. Kirja-instanssien URI:t on erikseen muunnettu Kirjasampo-palvelun käyttämään muotoon, jotta sen tarjoamiin kirjojen websivuihin voisi tarjota käyttäjille linkkejä. Tähän työhön datajoukosta on louhittu yhteyksiä, joissa henkilön kirjoittama romaani kuvaa tiettyä paikkaa.

Kirjasampo käyttää datan kuvaamiseen kaunokirjallisuutta varten luotua Kaunokki-ontologiaa [61], jonka nimiavaruuden etuliite määritellään seuraavasti:

```
@prefix kaunokki: <http://www.yso.fi/onto/kaunokki#> .
```

Kirjoja kuvataan Kaunokki-ontologian luokalla

```
kaunokki:Romaani.
```

Romaani-resursseilla voi olla ominaisuuksinaan muun muassa

```
kaunokki:tekija,
```

joka kuvaa romaanin kirjoittajaa, sekä

```
kaunokki:worldPlace,
```

joka kuvaa romaanin aiheena olevaa todellisen maailman paikkaa. Henkilöitä varten Kirjasampolla on oma ontologiansa, johon Biografiasammosta on tehty siltaus. Tosin avoimen datan versiota varten URI:en alkuosat piti vaihtaa. Paikkoja kuvaavien resurssien URI:t ovat vanhentuneita ja liittyvät FINTO-projektiin, mutta Kirjasammon avoin datajoukko sisältää niiden oleelliset tiedot, joiden avulla siltaus on mahdollinen.

3.5 Kansallisgallerian tietokanta

Kansallisgallerian[24] tarjoama avoin data sisältää metatietoja Kansallisgallerian kokelmiin kuuluvista yli 40000:sta teoksesta. Kansallisgalleria käyttää datan kuvaamiseen Dublin Core -metadatamallia. Kansallisgalleria ei tarjoa avointa dataa RDF-muodossa. Tässä työssä on käytetty Kansallisgallerian tarjoamaa JSON-muotoista datapakettia. Tähän työhön on kyseisestä datasta louhittu yhteyksiä, jossa henkilö

[23]http://www.ldf.fi/dataset/kirjasampo/index.html
[24]http://kokoelmat.fng.fi/api/v2support/docs/#/suomeksi

on luonut jotain paikkaa kuvaavan taideteoksen. Tekijäksi merkityt taiteilijat on yhdistetty yksinkertaisesti nimien perusteella Biografiasammon henkilöihin, ja vastaavasti aiheena olevat paikat on yhdistetty nimien perusteella käytettyyn paikkaontologiaan.

3.6 J. V. Snellmanin kootut teokset

Tätä työtä varten saatiin Edita Publishing Oy:ltä käyttöön J. V. Snellmanin koottuja teoksia koskevaa dataa. Datan sisältöä ja muuntamista RDF-muotoon käsitellään tarkemmin seuraavassa luvussa. Tähän työhön kyseisestä datasta käytettiin kirjeiden lähetys- ja vastaanottopaikkoja.

4 J. V. Snellmanin teosten datan muuntaminen

J. V. Snellman on 1800-luvulla vaikuttanut merkittävä suomalainen poliitikko, sanomalehtimies ja filosofi. Snellmanin laaja kirjallinen tuotanto koostuu suuresta määrästä erityyppisiä kirjoituksia, kuten romaaneista, tieteellisistä julkkaisuista, lehtiartikkeleista ja kirjeistä. Kriittinen editio J. V. Snellmanin kootuista teoksista [62] on julkaistu avoimena verkkosivustona[25]. Aineisto sisältää Snellmanin omien kirjoitusten lisäksi muiden, usein historiallisesti merkittävien henkilöiden, Snellmanille lähettämiä kirjeitä. Yhteensä aineistossa on noin 3000 tekstilähdettä, joista noin 1500 on kirjeitä. Erityisesti Snellmanin kirjeet ovat luonteeltaan sellaisia, että ne linkittävät voimakkaasti 1800-luvun Suomen historiallisesti merkittäviä ihmisiä toisiin ihmisiin ja paikkoihin. Siksi aineisto on hyödyllinen tässä työssä rakennetulle yhteyshaulle. Aineiston hyödyntämisen helpottamiseksi aineisto muunnettiin yksinkertaiseksi RDF-graafiksi. Aineistoa on mahdollista käyttää myöhemmin myös muussa tutkimuksessa. Sivuston kustantaja Edita Publishing Oy toimitti otoksen sivuston käyttämästä Drupal-tietokannasta datan RDF-muotoon muuntamista varten. Muunnos tehtiin Python-ohjelmointikielellä RDFLib-kirjastoa[26] käyttäen. Luodun RDF-graafin oletusnimiavaruus on:

```
@prefix : <http://ldf.fi/snellman/> .
```

Jos ei erikseen mainita, niin tässä luvussa esitellyt resurssit liittyvät tähän nimia-

[25] http://snellman.kootutteokset.fi/fi
[26] https://github.com/RDFLib/rdflib

varuuteen. Taulukko 1 sisältää muiden nimiavaruuksien etuliitteitä.

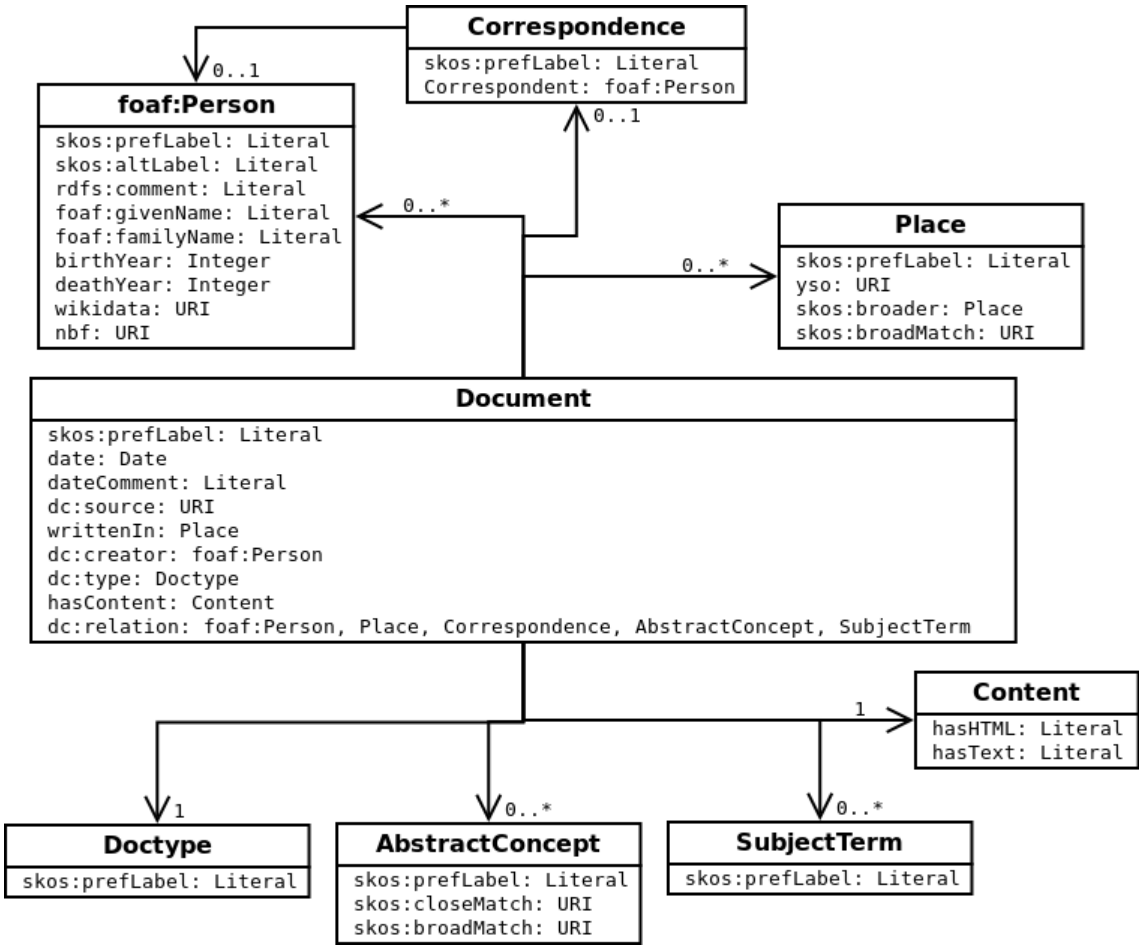
Toimitettu data muodostui noin kahden gigatavun kokoisesta xml-formaatissa olevasta export.xml siirtokopiosta, sekä useista csv-tiedostoista. Export.xml-tiedosto koostui elementeistä, jotka kuvasivat kukin yksittäistä web-sivuston osaa tai J. V. Snellmaniin liittyvää tekstilähdettä, sekä näiden dokumentteja kuvaavien elementtien lapsielementeista ja atribuuteista. Tämä tiedosto sisälsi dokumenttien tekstit HTML-muodossa, sekä muita kuhunkin dokumenttiin liittyviä tietoja. Dokumentin nimen ja päivämäärän kaltaisten tietojen lisäksi dokumentteihin oli merkitty niihin liittyviä henkilöitä, paikkoja, aiheita ja muita asioita käyttäen numerokoodeja. Näiden koodien selitykset löytyivät toimitetuista csv-tiedostoista. Merkinnät näistä teksteihin liittyvistä asioista ovat tehneet asiantuntijat käsin, mikä tekee aineistosta erityisen arvokkaan ja hyödyllisen.

Aineiston RDF-muunnoksen perusajatuksena oli, että export.xml-tiedoston kuvaamat tekstilähteet tulkittiin tarkoitusta varten luodun Document-luokan instansseiksi, joiden ominaisuuksia kuvataan erityisesti dokumenttien kuvaamiseen tarkoitetuilla Dublin Core[27] metadatatermeillä siltä osin kuin se on luontevaa. Jos ominaisuuden resurssi objektiin viitattiin tietokannassa koodilla, luotiin RDF-graafiin viite resurssia kuvaavaan instanssiin. Jos tietokannassa käytettiin literaalia arvoa, myös RDF-graafissa käytettiin suoraan vastaavaa literaalia arvoa. Graafeja luotiin kaksi erillistä, joiden unioni sisältää kaiken tiedon. Content.ttl-graafi sisältää lähinnä tekstilähteiden ja verkkosivujen sisällöt teksti- ja HTML-muodossa. Tämä graafi on kooltaan lähes 100 megatavua, mutta sisältää melko vähän kolmikkoja. Pääasiallinen graafi on nimeltään snellman.ttl ja se sisältää suurimman osan kolmikoista. Jos jokin asia sisältyy content-graafiin, siitä mainitaan erikseen. Koska dokumenttien sisällöt on erotettu omaan graafiinsa, on tämä pääasiallinen graafi käytännöllisen pieni kooltaan. Kuva 3 esittää graafin tietomallia UML-kaaviona.

Lähdemateriaalin csv-tiedostot sisälsivät kukin yhden tyyppisiä resursseja: paikkoja, henkilöitä, Snellmanin kirjeenvaihdon kohteita, dokumenttityyppejä, ”asioiksi” kutsuttuja käsitteitä, ”termejä” ja viitteitä Snellmanin elämäkertaa kuvaaviin kirjoihin sekä Snellmanin elämäkerran kuvauksen lukuihin. Viitteet kirjoihin ja lukuihin liittyivät vain web-sivuston sivuihin. Muut resurssit liittyivät tekstilähteisiin.

Lukumääräisesti merkittävin resurssityyppi olivat henkilöt. Aineisto sisälsi yli 4000 henkilöä. Pieni määrä näistä oli tosin selkeitä virheitä syötteessä. Lisäksi monelle oikein syötetylle henkilölle ei löydy viittauksia tekstilähteistä. Useimmille, lähes 3000

[27] http://dublincore.org/documents/dcmi-terms/



Kuva 3: J. V. Snellmanin tekstien UML-kaavio

henkilölle, kuitenkin löytyy. Tämä henkilötietokanta sisältää lähinnä Julius Caesarin ja Kaarle XII:n kaltaisia historiallisia henkilöitä sekä 1800-luvulla eläneitä ihmisiä, jotka kuuluivat J. V. Snellmanin sosiaaliseen verkostoon, tai joita Snellman muuten kommentoi esimerkiksi lehtikirjoituksissaan. J. V. Snellman itse puuttui aineistosta, mutta hänen kuvaamisekseen luotiin oma resurssinsa, jolle lisättiin vastaavat ominaisuudet kuin muillekin henkilöille. Kuva 4 sisältää esimerkkinä J. V. Snellmania vastaavan henkilö-resurssin kuvauksen.

Henkilöresurssien kuvaamiseen käytettiin luokkaa

```
foaf:Person.
```

Jokaisesta henkilöstä luotiin henkilö-luokan instanssi ja instansseihin liitettiin ominaisuuksia. Aineistossa henkilöistä oli merkitty nimi, yleensä kahdessa muodossa:

```
<http://ldf.fi/snellman/1> a foaf:Person ;
   :birthYear 1806 ;
   :deathYear 1881 ;
   :nbf <http://ldf.fi/nbf/p996> ;
   :wikidata <http://www.wikidata.org/entity/Q127688> ;
   rdfs:comment "1806-1881 Poliitikko, kirjailija, sanomalehtimies,
       valtiomies ja Suomen kansallisfilosofi."@fi ;
   skos:altLabel "J. V. Snellman"@fi ;
   skos:prefLabel "Snellman, Johan Vilhelm"@fi ;
   foaf:familyName "Snellman"@fi ;
   foaf:givenName "Johan Vilhelm"@fi .
```

Kuva 4: Esimerkki henkilö-resurssin RDF-kuvauksesta

suku- ja etunimi ensin, sekä useimmissa tapauksissa muutaman sanan mittainen elämäkerta. Ensimmäinen nimi, joka oli kaikilla henkilöillä, kuvattiin

`skos:prefLabel`

ominaisuuden arvona. Mahdollinen nimen toinen kirjoitusmuoto merkittiin

`skos:altLabel`

ominaisuuden arvoksi. Henkilöiden nimistä poimittiin säännöllisillä lausekkeilla etu- ja sukunimet, joita kuvattiin ominaisuuksilla

`foaf:familyName`

ja

`foaf:givenName.`

Elämäkertaa kuvattiin ominaisuudella

`rdfs:comment.`

Elämäkerran alkuun oli aineistossa yleensä merkitty henkilön syntymä- ja kuolinvuosi. Nämä oli mahdollista poimia helposti aineistosta säännöllisillä lausekkeilla. Näin monille henkilö-instansseille saatiin syntymävuosi

```
:birthYear,
```

ja kuolinvuosi

```
:deathYear.
```

Henkilö-instansseille luotiin siltaukset Wikidataan[28] ja Biografiasammon henkilöihin nimen ja syntymävuoden perusteella. Erityisesti siltaus Biografiasampoon oli keskeistä, koska yhteyshaku-sovellus toteutettiin näille henkilöille. Siltaus tehtiin vertailemalla henkilöiden nimiä ja syntymävuosia. Wikidatasta löytyi vastineet vajaalle tuhannelle henkilölle, kun taas Biografiasammon henkilöihin siltaus onnistui noin 250 tapauksessa. Yksinkertaisen arvion perusteella siltauksen tarkkuus on korkea, mutta saanti voisi olla parempi. Yhtään väärin sillauttua henkilöä ei huomattu. Wikidataan tehdystä siltauksesta puuttuu kuitenkin useita merkittäviä henkilöitä. Biografiasampoon siltaus onnistui useimmille keskeisille henkilöille, mutta parantamisen varaa on naisten kohdalla.

Toiseksi eniten oli ”kirjeenvaihto”-resursseja: 256 kappaletta. Nämä viittaavat Snellmanin kirjeenvaihtokumppaneihin eli tahoihin, joille Snellman on lähettänyt kirjeitä tai joilta hän on vastaanottanut niitä. Henkilöiden lisäksi nämä tahot sisältävät muita toimijoita kuten Porvoon tuomiokapitulin. Vertailemalla kirjeiden nimiä henkilöiden nimiin on näitä resursseja yhdistetty henkilöresursseihin ominaisuudella

```
:correspondent.
```

Kirjeenvaihtoresurssien nimet eivät aina olleet kirjoitettu samassa muodossa kuin henkilöiden nimet ja siksi henkilöresurssien ja kirjeenvaihtoresurssien yhdistäminen ei ole kaikissa tapauksissa onnistunut, ja on mahdollista, että se sisältää virheitä.

Kolmanneksi runsaslukuisin resurssien tyyppi olivat paikat. Nitä oli kuitenkin aineistossa merkittävästi vähemmän kuin henkilöitä: 135 kappaletta. Paikkojen kuvaamiseen luotiin luokka

```
:Place.
```

Aineistossa oli pakoista tietona vain nimi, joka kirjattiin paikka-intanssien

```
skos:prefLabel
```

[28]https://www.wikidata.org/wiki/Wikidata:Main_Page

ominaisuuden arvoksi. Paikat sillattiin nimen perusteella Yleisen suomalaisen ontologian paikkoihin (YSO-paikat). Siltausta varten määriteltiin ominaisuus

```
:yso.
```

joka määritettiin

```
skos:closeMatch
```

ominaisuuden aliominaisuudeksi. Paikoille luotiin sen perusteella hierarkia

```
skos:broader
```

ominaisuuksilla. Osa paikoista piti sillata käsin. Jos vastinetta ei löytynyt lainkaan, luotiin hierarkinen viittaus yleisempään paikkaan.

Neljänneksi eniten, 48 kappaletta, oli resursseja, joita kutsuttiin lähdeaineistossa ”asioiksi”. Parhaiten niitä voisi kuvailla abstrakteiksi konsepteiksi ja niille luotiin oma luokka

```
:AbstractConcept.
```

Esimerkkejä näiden resurssien nimistä ovat ”Yliopisto”, ”Opiskelu” ja ”Filosofia”. Aihe-instanssit sillattiin YSO-ontologiaan.

Aineisto sisälsi myös ”termeiksi” nimettyjä asioita, jotka liittyvät dokumentteihin. Tällaisia ”termejä” olivat esimerkiksi ”Lauantaiseura” ja ”Saima”. Näitä termejä oli kuitenkin vain 13 ja ne liittyivät hyvin pieneen määrään tekstilähteitä. Ne viittaavat ilmeisesti lähinnä julkaisuihin joihin Snellman on kirjoittanut. Niille luotiin oma luokkansa

```
:Term.
```

Resursseille ”Kirja” ja ”Yläluku” luotiin myös omat luokkansa. Ne kuitenkin liittyvät ”Snellmanin kootut teokset”-sivustoon, eivätkä ole tekstilähteiden automaattisen tulkinnan kannalta millään ilmeisellä tavalla mielenkiintoisia.

Tekstilähteiden, joita tässä kutsutaan dokumenteiksi, kuvaamiseksi luotiin oma luokkansa

```
:Document.
```

Luokan ominaisuuksien kuvaamiseen käytettiin suurimmaksi osaksi Dublin Core -termejä. Siirtokopiota export.xml luettiin aina yhtä dokumenttia kuvaava osa kerrallaan. Dokumentille luotiin aina oma instanssinsa RDF-graafiin ja sen ominaisuuksien arvot luettiin tiedostosta joko suoraan tai päätellen. Kuva 5 näyttää erästä tekstilähdettä vastaavan resurssin RDF-kuvauksen.

```
<http://ldf.fi/snellman/3893> a :Document,
        :Material,
        foaf:Document ;
    :hasContent :c3893 ;
    :letterReceiver <http://ldf.fi/snellman/1> ;
    :materialType "tekstilahde" ;
    :relatedCorrespondence <http://ldf.fi/snellman/13226> ;
    :writtenIn <http://ldf.fi/snellman/13225> ;
    dc:creator <http://ldf.fi/snellman/12448> ;
    dc:date "1830-07-25"^^xsd:date ;
    dc:relation <http://ldf.fi/snellman/10024>,
        <http://ldf.fi/snellman/12356>,
        <http://ldf.fi/snellman/12448>,
        <http://ldf.fi/snellman/12450>,
        <http://ldf.fi/snellman/12454>,
        <http://ldf.fi/snellman/13222>,
        <http://ldf.fi/snellman/13225> ;
    dc:source <http://snellman.kootutteokset.fi/fi/dokumentit/carl-august-snellmanilta-7> ;
    dc:type <http://ldf.fi/snellman/13218> ;
    skos:prefLabel "Carl August Snellmanilta" .
```

Kuva 5: Esimerkki tekstilähde-resurssin RDF-kuvauksesta

Aineistossa teksteille ei oltu merkitty kirjoittajaa suoraan. Kirjoittaja pääteltiin tekstin tyypin ja nimen päätteen perusteella. Jos kyseessä oli kirje ja sen otsikossa olevan henkilön nimen päätteenä oli ”-lta” tai ”-ltä”, eli toisin sanoen jos se oli ablatiivissa, pääteltiin, että kirjeen on kirjoittanut joku muu kuin J. V. Snellman. Tarkempi kirjoittaja pääteltiin siten, että kirjeisiin liittyy kirjeenvaihto-resurssi, johon liittyy henkilö. Esimerkiksi ”Johan Ludvig Runebergiltä” nimisen tekstilähteen päätteestä voi päätellä sen olevan jonkun muun Snellmanille lähettämä kirje. Kirjeellä on ominaisuus

```
dc:relation,
```

jonka kohde on Correspondence luokan instanssi nimeltä ”Johan Ludvig Runeberg”, jolla on kirjeenvaihto-ominaisuus, jonka kohde on J. L. Runebergia kuvaava henkilöinstanssi. Tämä merkitään

```
dc:creator
```

ominaisuuden arvoksi. Jos tekstiin ei liittynyt kirjeenvaihto-resurssia tai sen otsikosta ei löytynyt ablatiivin päätettä, tekstin kirjoittajaksi oletettiin J. V. Snellman.

Dokumentin tyyppiä kuvattiin ominaisuudella

```
dc:type.
```

Ominaisuudella

```
dc:source
```

kuvattiin dokumenttia kuvaavan www-sivun osoitetta J. V. Snellman -portaalissa. Ominaisuus

```
dc:date
```

kuvaa dokumentin luomispäivämäärää. Jos päivämäärä ei ole tarkka ja se on jouduttu arvioimaan, liittyi dokumenttiin päivämäärää kuvaava kommentti, yleensä: ”päivämäärä ei tarkka”. Jos tällainen kommentti löytyi, lisättiin dokumentille ominaisuus

```
:dateComment,
```

jonka arvoksi sijoitettiin kommentin teksti.

Dokumenteille merkittyjä yhteyksiä muihin resursseihin kuvattiin pääasiassa

```
dc:relation
```

ominaisuudella. Ominaisuuden kohteiksi merkittiin vastaavat instanssit. Resurssit dokumentteihin liittävien yhteyksien luonnetta ei oltu tarkemmin määritelty aineistossa, ja siksi jouduttiin käyttämään tällaista epämääräistä kuvausta. Kirjeiden lähetyspaikat oli kuitenkin asia, jonka katsottiin olevan mahdollista päätellä. Noin joka toiselle kirjeelle oli merkitty vain yksi siihen liittyvä paikka, ja näissä tapauksissa se lähes poikkeuksetta vaikutti olevan paikka, josta kirje on lähetetty. Kirjeen lähetys- tai kirjoituspaikkaa kuvattiin ominaisuudella

```
:writtenIn.
```

Muita tarkemmin määriteltyjä ominaisuuksia olivat kirjoittaja sekä kirjeen vastaanottaja, jota kuvattiin ominaisuudella

```
:letterReceiver.
```

Dokumentin sisältämään tekstiin viitataan ominaisuudella

```
:hasContent.
```

Tämän arvona on sisältöä kuvaava resurssiin, jonka tiedot on lisätty content-graafiin.

5 Yhteyksiä kuvaavan graafin muodostaminen

5.1 Tietomalli

Yhteyksien kuvaamista varten suunniteltiin oma tietomallinsa. Perusajatuksena on kuvata jokaista semanttista yhteyttä omana yhteyttä kuvaavan luokan instanssinaan. Yhteyksiä edustavien resurssien keskeisiä ominaisuuksia ovat yhteyden päätepisteet sekä luonnollisen kielen kuvaus yhteydestä. Muita mahdollisia ominaisuuksia ovat esimerkiksi yhteyden aika ja lähde. Tässä luvussa käytetty oletusnimiavaruus määritellään seuraavasti

```
@prefix : <http://ldf.fi/relse/> .
```

Tällaisen tietomallin ongelmaksi saattaa muodostua instanssien suuri määrä. Esimerkiksi kahden ihmisen välisen yhteyden ”sama syntymäpaikka” kuvaaminen saattaisi vaatia hyvin suuren määrän instansseja. Esimerkiksi vuonna 2016 Helsingissä asui 20095 ihmistä jotka olivat syntyneet Vantaalla[29]. Näistä teoriassa muodostuvien yhteistä syntymäpaikkaa kuvaavien yhteys-instanssien määrän voi laskea binomikertoimella, jonka yleinen kaava on:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Saatava luku tarkoittaa sitä, kuinka monella eri tavalla n alkiota sisältävästä joukosta voidaan poimia k alkiota sisältävä osajoukko. Koska tietomallissa yhteyden suunnalla on merkitystä, tämä luku pitää vielä kertoa kahdella. Yhteyksien määrä saadaan siis seuraavasti:

$$2 * \binom{20095}{2} = 2 * \frac{20095!}{2!(20095-2)!} = 2 * 201894465 = 403788930$$

[29]https://www.hel.fi/hel2/tietokeskus/julkaisut/pdf/17_06_28_Tilastoja_1_Maki_Vuori.pdf

Pelkästään Helsingissä asuvien Vantaalla syntyneiden henkilöiden saman syntymäpaikan muodostaman yhteyden kuvaamiseen tarvittaisiin siis yli 400 miljoonaa instanssia. Muistin kulutuksen kannalta tällainen määrä vaikuttaa epäkäytännölliseltä. Lisäksi on huomattava, että fasetit laskevat potentiaalisten yhteyksien määriä aina yhden valinnan jälkeen, ja näin suurten vaihtoehtojen määrien läpikäyminen voi olla hidasta.

Toisaalta tässä työssä ollaan kiinnostuneita henkilöiden ja paikkojen välisistä yhteyksistä. Tässä tapauksessa yhteyksien määrät eivät kasva helposti hallitsemattomiksi. Jos halutaan kuvata kaikkien Helsingissä asuvien Vantaalla syntyneiden syntymäpaikkaa tällä tavalla, tarvitaan yksinkertaisesti vain 20095 instanssia, eli Vantaalla syntyneiden määrä. Käytännössä tämä on myös luonteva kuvata vain yksisuuntaisena, joten tätä lukua ei tarvitse kertoa kahdella. Riittää siis ”Henkilö Matti Meikäläinen on syntynyt paikassa Vantaa,” eikä tarvita erikseen yhteyttä ”Paikassa Vantaa on syntynyt henkilö Matti Meikäläinen.” Sopivalla sovellusalalla yhteyksien määrä pysyy siis kohtuullisissa rajoissa, mutta on tärkeä ymmärtää tietomallin mahdolliset rajoitteet.

Tietomallin keskeinen luokka on yhteyttä kuvaava Relation-luokka. Muodostettava graafi koostuu Relation-luokan instansseista ja niihin liittyvistä ominaisuuksista. Nämä ominaisuudet on lueteltu taulukossa 3. Yhteyksien yksinkertaistettua muotoa voi kuvata kolmikkona

```
<henkilö> <liittyy tietyllä tavalla> <paikka>,
```

joka vastaa luonnollisen kielen lausetta. Näin ajatellen subjekti ja objekti kuvaavat semanttisen yhteyden päätepisteitä ja predikaatti kuvaa yhteyden luonnetta. Tietomallissa jokaista näistä vastaa oma ominaisuutensa.

Yhteyden päätepisteitä kuvaavat ominaisuudet

```
:personSubject,
```

joka kuvaa yhteyteen liittyvää henkilöä ja

```
:placeObject,
```

joka kuvaa yhteyteen liittyvää paikkaa. Nämä ovat ominaisuudet ovat yleisemmän yhteyden päätepisteitä kuvaavan ominaisuuden

```
:relationEndpoint
```

aliominaisuuksia. Henkilö on tässä työssä luodussa graafissa aina yhteyden subjekti ja paikka objekti. Tietomalli ei suoraan vaadi sitä, mutta yhteyksien kuvaukset muodostuvat luonnollisemmin sitä kautta. Yleisempi ominaisuus on tarkoitettu tulevaisuutta varten, jos tietomallia halutaan laajentaa ja luoda muunlaisia yhteyksiä.

Ominaisuus	**Kuvaus**	**Arvo**
personSubject	Yhteyteen liittyvä henkilö	URI
placeObject	Yhteyteen liittyvä paikka	URI
date	Yhteyden varhaisin mahdollinen aika (ei välttämätön)	xsd:date
relationType	Yhteyden tyyppi	URI
source	Resurssi johon yhteys liittyy alkuperäisessä lähteessä	URI
sourceLink	Web sivu joka kuvaa yhteyttä alkuperäisessä lähteessä (usein sama kuin source)	URL
sourceName	Ihmisluettava kuvaus lähteestä josta yhteys on louhittu	literaali
skos:prefLabel	Yhteyden ihmisluettava kuvaus	literaali

Taulukko 3: Yhteyksiä kuvaavien resurssien ominaisuudet

Yhteyden tyyppiin viitataan ominaisuudella

```
:relationType.
```

Tämän ominaisuuden arvoksi tulee jokin yhteyden tyyppiä kuvaavan luokan

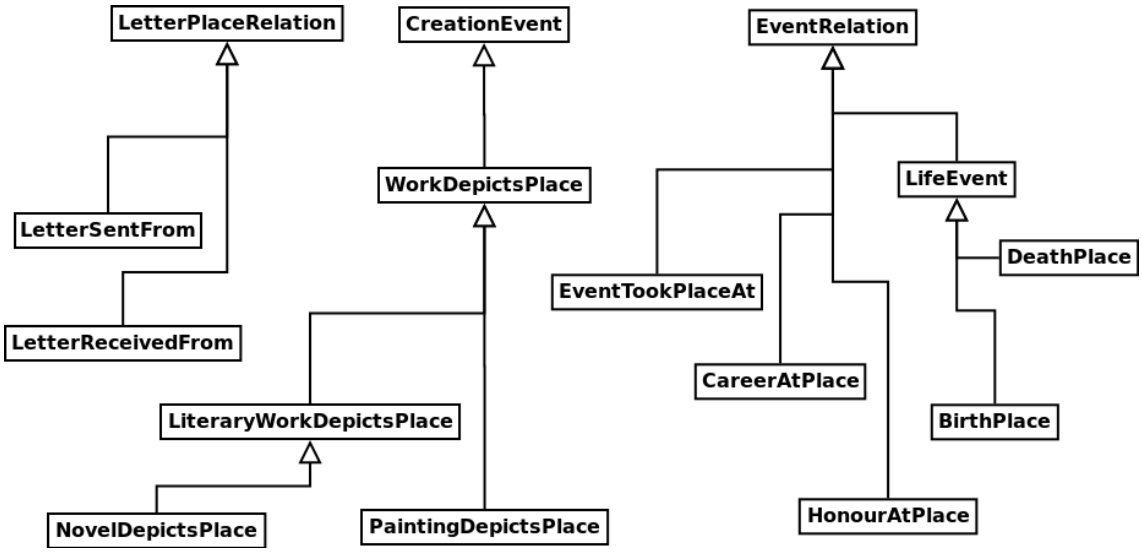
```
:RelationClass
```

instanssi. Yhteyden luokkaa kuvaavilla rerusseilla on oma hierarkiansa jossa saman tyyppiset luokat ovat sukua toisilleen. Kuvassa 6 on esitetty näiden luokkien hierarkia.

Muita yhteys-entiteettien ominaisuuksia ovat esimerkiksi

```
:date,
```

joka kuvaa yhteyden aikaisinta mahdollista alkamispäivämäärää. Lähdemateriaalin luonteesta johtuen päivämäärä on joskus epämääräinen ja puuttuu joskus kokonaan. RDF-resurssiin, jonka kautta yhteys muodostuu, viitataan ominaisuudella



Kuva 6: Yhteyksien hierarkia

```
:source.
```

Ominaisuuden

```
:sourceLink
```

arvo on web-sivu joka sisältää ihmisluettavaa tietoa liittyen yhteyden lähteeseen. Ominaisuuden

```
:sourceName
```

arvo on literaali ihmisluettava kuvaus lähteestä josta yhteys on löydetty. Lopulta ominaisuuden

```
skos:prefLabel
```

arvona on yhteyden ihmisluettava kuvaus. Tässä työssä yhteyksien kuvaukset on toteutettu vain suomeksi, mutta graafiin mahdollista liittää helposti selityksiä myös muilla kielillä. Kuva 7 näyttää esimerkin eräästä yhteyttä kuvaavasta resurssista Turtle-muodossa.

```
<http://ldf.fi/relse/http%3A%2F%2Fldf.fi%2Fnbf%2Fp1233http%3A%2F%2Fldf.fi%
2Frelse%2Fp1935http%3A%2F%2Fwww.yso.fi%2Fonto%2Fkaunokki%
23ateos_27905novel_depicts>
        a                    relse:Relation ;
        relse:date           "1970-01-01Z"^^xsd:date ;
        relse:personSubject  nbf:p1233 ;
        relse:placeObject    relse:p1935 ;
        relse:relationType   relse:novelDepictsPlace ;
        relse:source         kaunokki:ateos_27905 ;
        relse:sourceLink     <https://www.kirjasampo.fi/fi/kulsa/kauno%
253Aateos_27905> ;
        relse:sourceName     "Kirjan tiedot Kirjasammossa"@fi ;
        skos:prefLabel       "Hella Wuolijoki on kirjoittanut romaanin
'Työmiehen perhe' joka kuvaa paikkaa Helsinki." .
```

Kuva 7: Esimerkki Yhteys-resurssista

5.2 Käytetyt henkilö- ja paikka-ontologiat

Alkuperäisenä tarkoituksena oli käyttää suoraan Biografiasammon ontologioita henkilöiden ja paikkojen kuvaamiseen. Käytännössä tämä osoittautui vaikeaksi. Osin tämä johtuu siitä, että Biografiasammon kehitys oli osin kesken tätä työtä toteutettaessa. Toisaalta Biografiasampo sisältää monessa tapauksessa useita resursseja, jotka kuvaavat samaa henkilöä tai paikkaa, mikä on ongelmallista. Kolmanneksi Biografiasampo sisältää myös eläviä henkilöitä, ja heidän tietojensa julkaisemiseen saattaa lain mukaan liittyä rajoitteita. Lisäksi paikkoja oli joissain tapauksissa tarpeen tunnistaa pelkän literaaliarvon perusteella. Rajatumpi paikka-ontologia auttaa välttämään virheitä.

Edellä mainituista syistä lopullisissa yhteyksissä esiintyvät henkilöt ovat vain osajoukko Biografiasammon henkilöistä ja paikoista. Paikkoja varten luotiin omat resurssinsa, jotka vastasivat YSO-paikat ontologian paikkoja. Biografiasammon paikat on sillattu YSO-paikat ontologiaan, joten ylimääräistä siltausta ei tarvinnut tehdä. Yhteykset muodostavien muotojen toiminnan vuoksi on kuitenkin tarpeellista, että yhtä paikkaa vastaa vain yksi resurssi. Siksi siltaus rajoitettiin vain yhteen Biografiasammon paikka-resurssiin jokaista paikkaa kohden.

Biografiasampo sisältää suuren määrän henkilö-instansseja, mutta tässä työssä haluttiin käyttää vain datan ytimen muodostamia noin 13000 henkilöä, joilla on SKS:n Biografiakeskuksen kirjoittama elämäkerta. Biografiasampo sisältää henkilöistä paljon sellaista tietoa, joka on turhaa tämän työn kannalta. Tämän vuoksi luotiin erilli-

nen Turtle-tiedosto, joka sisältää Biografiasammon ydinhenkilöt tämän työn kannalta oleellisine ominaisuuksineen. Ongelmia aiheuttaa, että samalla henkilöllä saattaa joissain harvoissa tapauksissa olla elämäkerta useammassa kuin yhdessä tietokannassa. Tämä katsottiin kuitenkin niin harvinaiseksi, että asiaan ei erikseen puututtu. Lisäksi osa elämäkerroista kuvaa useampaa kuin yhtä henkilöä, esimerkiksi pariskuntia tai sisaruksia Tämä tuottaa yhteyshaun kannalta hankalia tuloksia, joten ne pyrittiin rajaamaan pois tuloksista. Lisäksi tietoturvasyistä elävät henkilöt haluttiin rajata julkisen sovelluksen ulkopuolelle. Tämä rajaus tehtiin lopulliseen dataan yhteyksien luomisen jälkeen.

5.3 Yhteyksien louhinta

Yhteys-instanssit muodostettiin SPARQL CONSTRUCT -kyselyillä lukuunottamatta Kansallisgallerian dataa, jota ei ole julkaistu RDF-muodossa. Kansallsigallerian data luettiin JSON-tiedostosta, ja yhteydet lisättiin graafiin Pythonin RDFLib-kirjaston avulla. Yhteys-instanssit muodostettiin kolmessa osassa. Ensimmäisessä vaiheessa käytettiin yleiskäyttöisiksi tarkoitettuja CONSTRUCT-hahmoja. Toisessa vaiheessa käytettiin erikoistuneita CONSTRUCT-hahmoja, joilla sijoitettiin henkilö- ja paikka-instansseiksi tämän sovelluksen käyttämien ontologioiden resurssit ja luotiin URI:t yhteys-instansseille. Lisäksi tässä vaiheessa luotiin luonnollisen kielen selitykset yhteyksille. Kolmannessa vaiheessa lopulliseen graafin tulevia yhteyksiä rajattiin vielä yksinkertaisella hahmolla.

Yhteyksien louhimista varten potentiaalisesti kiinnostavien avointen aineistojen tunnistamisen jälkeen oli tarpeen tunnistaa niiden sisältämät kiinnostavat yhteystyypit ja ontologiat, joilla ne on ilmaistu. Erityisesti oltiin kiinnostuneita luovan työn tuloksista. Tällaisia yhteyksiä, joissa jokin luova työ kuvaa paikkaa, löytyikin useista lähteistä. Nämä yhteydet oli myös yleensä kuvattu karkeasti ottaen samanlaisella mallilla, jota voi kutsua ”teos, jolla on tekijä ja aihe”-malliksi. Koska Biografiasampo ja HISTO-ontologia käyttävät molemmat tapahtumiin perustuvaa tietomallia, niiden sisältämät yhteydet on kuvattu tavalla jota voi karkeasti kutsua ”tapahtuma, johon liittyy henkilö ja paikka”-malliksi.

Kun URI:t luotiin CONSTRUCT lauseilla, URI muodostettiin yhdistämällä henkilön, paikan ja lähteen URI:t sekä lisäämällä yhteystyyppiä kuvaava koodi. URI:n tarkoituksena ei ole sisältää informaatiota. Se on muodostettu näin jotta URI:t voitaisiin helposti muodostaa deterministisesti. URI:t olisi mahdollista muodostaa lyhyemmin, mutta luettavuuden kustannuksella.

Yhteyksien luonnollisen kielen kuvaukset muodostettiin mahdollisimman yksinkertaisella tavalla, eli sijoittamalla muuttujat, kuten hekilön ja paikan nimet, valmiiseen lausekehikkoon. Esimerkki tällaisesta kehikosta on:

> ”Henkilö *<henkilön nimi>* on vuonna *<vuosi>* maalannut taulun nimeltä *<taulun nimi>*, joka kuvaa paikkaa *<paikan nimi>*.”

Suomen kielessä on suhteellisen monimutkaiset säännöt sanojen taivuttamiseen. Lauseita voi kuitenkin muodostaa melko vapaasti. Siksi kuvaukset oli mahdollista muodostaa siten, että muuttujia ei tarvinnut taivuttaa. Lopulliset kuvaukset ovat ymmärrettävää suomea, mutta joissain tapauksissa voivat tuntua keinotekoisilta.

5.3.1 Biografiasammon yhteydet

Yhteyksien pääasiallisena lähteenä oli Biografiasampo, josta selkeästi suurin osa yhteyksistä on louhittu. Koska henkilö- ja paikkaontologiat olivat osajoukkoja Biografiasammon ontologioista oli niiden yhdistäminen yksinkertaista. Ensimmäisessä versiossa Biografiasammon yhteydet haettiin yhdellä SPARQL muodolla ilman erillistä henkilöiden ja paikkojen yhdistämisen vaihetta. Biografiasammosta louhitut yhteydet koskevat henkilöiden syntymä- ja kuolinpaikkoja, henkilöiden uraan tietyssä paikassa liittyviä tapahtumia ja kunnianosoituksia, jotka liittyvät johonkin paikkaan. Biografiasampo sisältää myös teosten luomistapahtumia, jotka liittyvät paikkaan, mutta niiden kuvaukset olivat usein niin epämääräisiä, että niistä olisi harvoin muodostunut järkevä luonnollisen kielen kuvaus yhteydestä. Siksi niitä yhteyksiä ei otettu mukaan lopulliseen työhön. Biografiasampoa toteutettiin samaan aikaan tämän työn kanssa, joten ymmärrettävästi siihen tuli muutoksia ja dokumentointi oli vielä keskeneräistä.

Syntymä ja kuolema oli ilmaistu käytännössä samalla tavalla. Molempia yhteystyyppejä varten luotiin oma muotonsa. Biografiasammossa syntymällä tai kuolemalla on oma kyseistä tapahtumaa kuvaava resurssinsa. Tapahtumaa koskevaan henkilöön viitataan syntymissä CIDOC CRM -ominaisuudella

```
crm:P98_brought_into_life
```

ja kuolemissa ominaisuudella

```
crm:P100_was_death_of.
```

Tapahtuman paikkaan viittaava ominaisuus on kuitenkin vain Biografiasammon oma ominaisuus, eikä se ole minkään yleisemmän ominaisuuden alaluokka. Tämä vaikeuttaa muodon mahdollista yleisempää soveltamista, mutta CIDOC CRM -mallin ominaisuuksista voi tunnistaa syntymää ja kuolemaa koskevat tapahtumat, ja näihin perustuen muotoa voisi ehkä pienin muokkauksin soveltaa myös muualla. Lopulliseen graafiin tuli 7182 syntymäpaikkaa ja 7349 kuolinpaikkaa kuvaavaa yhteyttä.

Biografiasampo sisältää erilaisia tapahtumia, joiden luokka on joko ”Ura”, ”Kunnianosoitus” tai ”Tuote”. Tapahtumia kuvaava luokka on luokan

```
schema:Event
```

alaluokka. Tapahtumat, joiden luokka oli ”Tuote” jätettiin pois, koska niiden perusteella ei yleensä onnistuttu luomaan järkevää ihmisluettavaa kuvausta. Lisäksi tapahtumat, joiden paikkana oli ”Suomi”, ”Suomen leijona” tai ”Pohjola” suodatettiin pois. ”Suomi” oli tässä yleensä liian epäkiinnostava paikka. ”Suomen leijona” ja ”Pohjola” taas sisälsivät suuren määrän virheitä. Lopulliseen graafiin tuli 20536 uraa kuvaavaa yhteyttä ja 2528 kunnianosoituksiin liittyvää yhteyttä.

5.3.2 Luovan työn aiheita kuvaavat yhteydet

Yhteyksiä, joissa henkilö on luonut teoksen joka kuvaa paikkaa, muodostettiin Fennicasta, Kirjasammosta sekä Kansallisgallerian tietokannasta. Kaikki näistä kuvasivat tätä yhteyttä karkeasti saman tyyppisellä mallilla: teoksella on oma instanssinsa, johon liittyy jollain ominaisuudella henkilö, joka on teoksen tekijä, sekä paikka, joka on teoksen aiheena. Periaatteessa pitäisi olla mahdollista luoda yleinen hahmo, jolla voi hakea luovan työn aiheisiin liittyvät yhteydet näistä kaikista lähteistä. Kanallisgallerian datasta tulisi kuitenkin ensin tehdä muunnos RDF-muotoon, mitä tässä työssä ei tehty. Lisäksi Fennican ja Kirjasammon sisältämille erityyppisille teoksille haluttiin tarkkuuden vuoksi luoda omat hahmonsa.

Kansallisgallerian datasta yhteydet louhittiin JSON-muotoisesta tiedostosta käyttämällä yksinkertaista Python-kielellä kirjoitettua ohjelmaa. Ohjelma kävi läpi kaikki tietokannan taideteokset ja yhdisti taideteoksen tekijän Biografiasammon henkilöön, jos näillä oli sama nimi ja jos Biografiasammon henkilön ammatiksi oli merkitty taidemaalari. Paikat tunnistettiin literaaleista aiheiden nimistä vertaamalla paikkojen nimikkeisiin. Tämä aiheutti jonkin verran ongelmia. Aihe ”Johannes” jouduttiin rajaamaan pois. Suomessa on ollut kunta nimeltä Johannes, johon tämä aihe olisi

muuten yhdistänyt. Epäilemättä lähes kaikissa tapauksissa aihe ”Johannes” tarkoittaa taulussa kuitenkin apostoli Johannesta, eikä paikkaa. Osasta tauluista puuttuu aika, jolloin ne on luotu, ja siten myöskään niistä louhituille yhteyksille ei ole annettu päivämäärää. Taideteoksen kuvaamiin paikkoihin liittyviä yhteyksiä tuli lopulliseen graafiin 1091 kappaletta.

Fennica eli kansalliskirjaston tietokanta käyttää suureksi osaksi schema.org-metadatamallia. Kirjojen aiheita kuvataan YSO-ontologian avulla. Fennica noudattaa siis melko hyvin standardeja, ja sitä varten luodun hahmon voi kuvitella soveltuvan myös muualle pienin muutoksin. Fennicassa ongelmallisia ovat henkilöt. Samaa henkilöä saattaa esimerkiksi kuvata useampi resurssi. Biografiasampoon oli tehty siltaus Fennican henkilöihin, mutta tätä työtä varten tehtiiin erillinen siltaus yksinkertaisella tavalla nimiä ja syntymäaikoja vertaamalla, jotta saatiin parempi tarkkuus. Fennican yhteyksille ei poimittu päivämääriä. Fennicasta saatiin lopulliseen graafiin 881 yhteyttä, jotka kuvaavat paikkaa kirjallisen teoksen aiheena.

Kirjasammosta luohitut yhteydet, jotka kuvaavat romaanien aiheina olevia paikkoja, ovat tässä tulkittu omaksi yhteyksien luokakseen, joka on kirjallisten töiden alaluokka. Kirjasammon henkilöihin oli Biografiasammossa luotu siltaus, mutta koska se oli suljettuun datajoukkoon piti URI:en alkuosat muuttaa avoimen datajoukon käyttöä varten. Paikoille luotiin oma siltauksensa. Koska käyttäjille haluttiin tarjota linkkejä Kirjasammon sivuille, piti avoimen datasetin kirjojen URI:en alkuosat vielä muuttaa sopiviksi. Lopulliseen graafiin tuli 290 Kirjasammosta louhittua romaanin aiheena olevaa paikkaa kuvaavaa yhteyttä.

5.3.3 Muut yhteydet

HISTO-ontologiasta oli suhteellisen yksinkertaista muodostaa yhteydet, jotka kuvaavat henkilön osallistumista historialliseen tapahtumaan tietyssä paikassa. HISTO-ontologia käyttää CIDOC CRM -metadatamallia, ja sen vuoksi sitä varten muodostettu hahmo sopii luultavasti pienin muutoksin myös muihin vastaaviin tietokantoihin. Historiallisten tapahtumien perusteella muodostui 345 yhteyttä.

Snellman-aineistosta muodostettiin yhteyksiä, jotka liittyvät kirjeiden lähetys- ja vastaanottopaikkoihin. Molempia yhteystyyppejä varten muodostettiin omat hahmonsa. Hahmojen muodostaminen oli helppoa, koska Snellman-aineistosta oli aikaisemmassa vaiheessa muodostettu RDF-graafi erityisesti Biografiasampoa ja yhteyshakua silmällä pitäen. Kirjeiden lähetyspaikkaan liittyviä yhteyksiä syntyi 575

kappaletta ja vastaanottopaikkaan liittyviä 124 kappaletta. Suurin osa näistä koskee J. V. Snellmania, mutta näitä yhteyksiä löytyi myös monille muille henkilöille.

Yhteyksiä muodostettiin yhteensä noin 40000 kappaletta. Taulukko 4 näyttää yhteyksien määrät niiden tyyppien mukaan.

Yhteyden tyyppi	**Yhteyksiä**
Historiallinen tapahtuma paikassa	345
Kirje lähetetty paikasta	575
Kirje vastaanotettu paikasta	124
Kirjoitus kuvaa paikkaa	881
Kunnianosoitus liittyy paikkaan	2528
Kuollut paikassa	7349
Maalaus kuvaa paikkaa	1091
Romaani kuvaa paikkaa	290
Syntynyt paikassa	7182
Ura tai opiskelu liittyy paikkaan	20536
Yhteensä	40901

Taulukko 4: Yhteyksien määrä tyypin mukaan

6 Yhteyshaun demonstraattori

Tässä työssä kehitetystä yhteyshausta on julkaistu demonstraattori[30] osana Biografiasampo-portaalia[31]. Web-sovellus perustuu Mikko Kohon ja muiden toteuttamaan JavaScript-työkaluun nimeltä SPARQL Faceter[32] [37]. Sovelluksen arkkitehtuuri perustuu AngularJS-sovelluskehyksellä toteutettuun selainpuoleen, joka tekee SPARQL kyselyjä kolmikkotietokantaan. Yhteyshakudemonstraattorin ohjelmoinnissa on käytetty mallina SPARQL Faceter työkalun toimintaa esitteleviä demoja[33][34]. Sovelluksen ulkonäkö on pyritty saamaan vastaamaan Biografiasammon yleisempää ilmettä muun muassa käyttämällä samoja tyylitiedostoja. Sovelluksen pystyttämisestä palvelimelle ja sen ohjaamisesta oikeaan osoitteeseen vastasivat Semanttisen laskennan tutkimusryhmän Jouni Tuominen ja Esko Ikkala.

[30] http://biografiasampo.fi/yhteyshaku
[31] http://biografiasampo.fi/
[32] https://github.com/SemanticComputing/angular-semantic-faceted-search
[33] https://github.com/SemanticComputing/sparql-faceter-dbpedia-demo
[34] https://github.com/SemanticComputing/sualt-fha-finds-faceter

Yhteyshaku-sovellus ei käytä suoraan Biografiasammon kolmikkotietokantaa. Yhteyshakua varten on luotu oma kolmikkotietokanta, joka sisältää omat rajatut versiot Biografiasammon henkilö- ja paikkaontologioista. Lisäksi tietokantaan on ladattu Biografiasammon paikkaontologian sisältävä Turtle-tiedosto paikkojen hierarkian muodostamiseksi, sekä henkilöiden ammatteja kuvaava Turtle-tiedosto.

Sovelluksen käyttöliittymä tarjoaa jokaisen yhteyden kohdalla linkin yhteyteen liittyvän henkilön ja paikan sivuille Biografiasammossa. Linkit on muodostettu yhdistämällä henkilön tai paikan URI:n loppuosa sivutyypin alkuosaan.

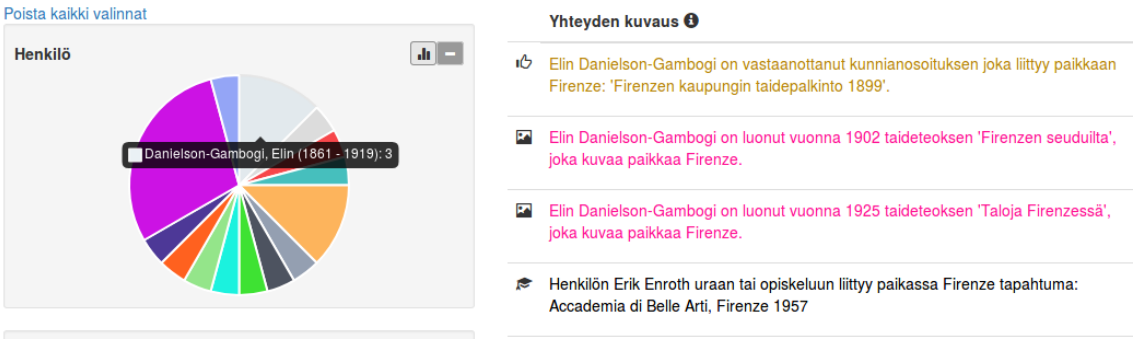
6.1 Sovelluksen käyttöliittymä

Web-sovelluksen ajatuksena on näyttää yhteydet listana ja tarjota käyttäjälle mahdollisuus rajata yhteyksiä sen tyypin sekä yhteyksiin littyvien henkilöiden ja paikkojen perusteella. Kuva 8 näyttää yleiskuvan sovelluksen käyttöliittymästä. Fasetit näkyvät vasemmalla ja yhteydet niiden oikealla puolella.

Kuva 8: Yleisnäkymä sovellukseen

Yhteyksien rajaamisessa käytetään fasetteja eli suodattimia. Suodattimet on toteutettu SPARQL Faceter -kirjastolla. Yhdestä suodattimesta voi valita vain yhden vaihtoehdon kerrallaan, mutta osa suodattimista on hierarkisia ja niissä hierarkiassa

ylempänä oleva valinta sisältää hierarkiassa alempana olevat valinnat. Suodattimet päivittyvät automaattisesti, kun yhdessä tehdään valinta. Jokainen suodatin laskee silloin uudestaan rajauksiin sopivat vaihtoehdot ja niiden lukumäärät. Suodattimet tarjoavat valittavaksi vain rajauksia, joihin liittyy sisältöä. Jokainen suodatin sisältää myös napin, josta kyseisen fasetin esityksen voi vaihtaa piirakkakaavioksi, joka esittää graafisesti mahdollisten yhteyksien suhteellisen määrän voimassa olevilla valinnoilla. Esimerkiksi kuvassa 9 näkyy fasetti piirakkakaaviona josta voi vertailla eri taiteilijoiden Firenzeen liittyvien yhteyksien määriä. Valinnat on järjestetty faseteissa aakkosjärjestykseen. Toinen ilmeinen vaihtoehto olisi järjestää valinnat yhteyksien määrän mukaan. Molemmilla ratkaisuilla on etunsa. Aakkosjärjestys auttaa löytämään tietyn valinnan helpommin, kun taas suuruusjärjestys auttaa löytämään merkittäviä valintoja. Tässä on ajateltu, että piirakkakaavion käyttäminen tarjoaa riittävän avun määrällisesti merkittävimpien valintojen löytämiseen.

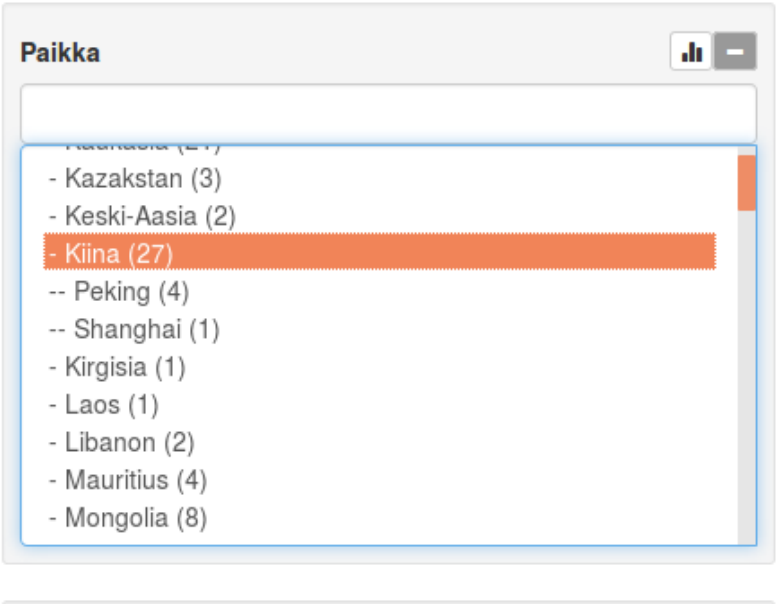


Kuva 9: Esimekki fasetista piirakkakaaviona

Käyttäjälle tarjotut suodattimet ovat ”Henkilö”, ”Arvo, ammatti tai toiminta”, ”Paikka” ja ”Yhteyden tyyppi”. Henkilö-suodatin näyttää henkilön nimen ja rajaa yhteyksiä yksittäisen henkilön mukaan. Arvo, ammatti tai toiminta -suodatin rajaa yhteyksiä, niihin liittyvien henkilöiden ammatin perusteella. Ammatti-fasetista on jätetty julkisessa versiossa hierarkia pois koska useat hierarkiset fasetit saattavat aiheuttaa merkittävää hidastumista. Paikka-suodatin on hierarkinen, ja rajaa yhteyksiä niihin liittyvän paikan perusteella. Yhteyden tyyppi -suodatin rajaa yhteyksiä niiden typin mukaan. Myös tästä suodattimesta on jätetty hierarkia pois, ainakin ensimmäisestä julkisesta versiosta, liiallisen hidastumisen välttämiseksi.

Fasetit mahdollistavat sen, että käyttäjä voi hakea yhteyksiä laajemmille ryhmille, eikä vain yksittäisille entiteeteille. Valitsemalla Italian paikka-fasetista saa yhteyk-

siä, jotka liittyvät muun muassa Firenzeen ja Roomaan. Vastaavasti kuvassa 10 on esitetty paikka-fasetti, jossa on valittuna Kiina ja sen kautta Peking ja Shanghai. Valitsemalla ammatin ”kirjailija” saa yhteyksiä, jotka liittyvät muun muassa Aleksis Kiveen ja Paavo Haavikkoon. Tämä ryhmittely perustuu henkilö- ja paikkaontologioihin. Eli esimerkiksi Firenzeen liittyvään yhteys-instanssiin ei suoraan liity tietoa siitä, että kyseinen yhteys koskee myös Italia. Tämä tieto syntyy paikka-ontologian hierarkian perusteella.



Kuva 10: Esimerkki hierarkisesta fasetista

Haun tulokset esitetään listana jossa on viisi saraketta. Ensimmäinen sarake näyttää yhteyden selityksen luonnollisella kielellä, esimekiksi: ”Ida Aalberg on syntynyt paikassa Janakkala vuonna 1857.” Erityyppiset yhteydet on kirjoitettu eri väreillä. Esimerkiksi syntymä-tyyppiä olevien yhteyksien teksti on vihreää. Lisäksi jokaisella yhteystyypillä on oma ikoninsa, joka näytetään ennen tekstiä. Toinen sarake näyttää hyperlinkin yhteyden henkilöön. Teksti näyttää henkilön nimen sekä syntymä- ja kuolinvuoden, esimerkiksi: ”Aalberg, Ida (1857 – 1915)”. Linkistä käyttäjä voi halutessaan siirtyä kyseisen henkilön omalle sivulle Biografiasammossa, josta voi löytää laajemmin tietoa tästä henkilöstä. Seuraava sarake näyttää hyperlinkin, jonka tekstinä on yhteyteen liittyvä paikka. Linkistä käyttäjä pääsee Biografiasammon kyseistä paikkaa kuvaavalle sivulle. Neljäs sarake sisältää hyperlinkin, jonka tekstinä on sen lähteen kuvaus, josta kyseinen yhteys on löydetty. Linkistä käyttäjä pääsee

halutessaan alkuperäisen lähteen ihmisluettavalle kotisivulle. Tällainen voi olla esimerkiksi taideteoksen kotisivu Kansallisgallerian sivustolla, josta löytyy taulun kuva ja tietoa taulusta. Viimeisenä sarakkeena on yhteyden tyyppi, johon on kirjoitettu yhteyden tyypin kuvaus yhteyden tyypin tunnusvärillä.

Yhteyksiä näytetään kerrallaan maksimissaan 20 kappaletta yhdellä sivulla. Sivun alaosassa on painike, jolla voi selata yhteyksiä, jos niitä on enemmän kuin 20. Yhteydet on järjestetty aakkosjärjestykseen ensisijaisesti yhteyteen liittyvän henkilön sukunimen perusteella ja toissijaisesti yhteyteen liittyvän paikan nimen mukaan.

Sivun yläosassa on palkki, joka on yhteinen kaikille Biografiasammon sovelluksille. Palkki on luotu erikseen tälle ja muille Biografiasammon sovelluksille. Vaikka yhteistä mallia on käytetty lähtökohtana, voi pieniä eroja olla. Palkin oikeassa yläkulmassa on Google Kääntäjä -elementti[35], jonka avulla sivun voi kääntää eri kielille. Käännöset ovat Googlen kääntäjän automaattisesti luomia eivätkä yleensä vastaa laadultaan ihmisen tekemiä.

6.2 Esimerkkejä sovelluksen käytöstä

Kalle Päätalo (1919–2000) oli tuottelias kirjailija, joka on tunnettu erityisesti omaelämäkerrallisista romaaneistaan. Siksi hän sopii hyvin esimerkiksi demonstraattorin käytöstä. Valitsemalla Kalle Päätalon henkilö-suodattimesta, käyttäjälle näytetään lista, jossa on 39 yhteyttä Päätalon ja jonkin paikan välillä. 32 näistä on ovat tyypiltään ”Romaani kuvaa paikkaa”, eli Päätalo on kirjoittanut romaanin, joka kuvaa tiettyä paikkaa. Tässä on hyvä huomata, että yksi romaani voi luoda useita yhteyksiä, jos se kuvaa useaa paikkaa. Kaksi paikkaa, joita Päätalo on kuvannut sovelluksen mukaan eniten, ovat Tampere ja Taivalkoski. Tämän voi todeta joko selaamalla yhteyksien selityksiä tai käyttämällä hyväksi fasetteja ja mahdollisesti niiden piirakkakaavio-toimintoa. Muun tyyppiset yhteydet liittyvät Päätalon syntymään, kuolemaan, uraan ja kunnianosoituksiin. Näiden yhteyksien selityksistä ja tyypeistä voi helposti nähdä, että Tampere ja Taivalkoski liittyvät vahvasti Päätalon elämään. Tämän perusteella voi siis selittää, miksi Kalle Päätalo on kuvannut paikkoja teoksissaan: selvästi hän on kuvannut paikkoja, jotka liittyvät hänen omaan elämäänsä. Tämä on niin ilmeinen ja Suomessa laajalti tunnettu esimerkki, että se ei luultavasti tule kenellekään yllätyksenä, mutta se toimii hyvänä esimerkkinä tällaisen yhteyshaun toiminnasta ja mahdollisuuksista.

[35]https://translate.google.com/manager/website/

Toisen esimerkin voi ottaa taiteen alalta. Valitsemalla yhteyden tyypiksi ”Maalaus kuvaa paikkaa” voi henkilö-fasetista nähdä, että tämän tyyppisiä yhteyksiä on selkeästi eniten Werner Holmbergillä. Hän vaikuttaisi siis pelkästään tämän perusteella olevan taiteilija, joka on kuvannut paikkoja töissään erityisen paljon. Werner Holmbergin henkilölinkistä käyttäjä voi siirtyä Holmbergin omalle sivulle[36] Biografiasammossa, jossa Holmbergin elämäkerta kuvaa häntä Suomen merkittävimmäksi maisemamaalariksi, mikä on linjassa yhteyshausta saadun informaation kanssa. Toisaalta valitsemalla ammatti-fasetista ammatiksi taidemaalarin, voi etsiä yleisesti taiteilijoille löydettyjä yhteyksiä. Selaamalla paikka-fasetista yhteyksien määriä eri paikoille, voi huomata yhteyksien keskittyvän ulkomailla tietyille kaupungeille. Erityisen paljon taiteiljoilla on yhteyksiä Pariisin, kuten saattaa odottaakin. Myös Tukholmaan ja Pietariin liittyy paljon yhteyksiä, mitä saattaa odottaa, koska ne ovat läheisiä metropoleja ja valtakunnan pääkaupunkeja Suomen historian eri vaiheissa. Muita erottuvia kaupunkeja yli kahdellakymmenellä yhteydellä ovat Firenze ja Düsseldorf. Firenze on tietenkin tunnettu taiteen kaupunki, mutta Düsseldorf voi tuntua yllättävältä käyttäjälle, joka ei tunne taidehistoriaa. Düsseldorf oli kuitenkin 1900-luvulla merkittävä opiskelupaikka suomalaisille taiteilijoille. Esimerkiksi yllämainittu Werner Holmberg opiskeli ja myös kuoli siellä. Firenze ja Düsseldorf näyttäytyvät ylipäätään silmiinpistävän samanlaisina tässä haussa. Molempiin näiden paikkojen ja suomalaisten välillä näyttää olevan karkeasti saman verran yhteyksiä, ja ne ovat osin samanlaisia. Esimerkiksi molemmissa on kuollut yhteyshaun mukaan kaksi suomalaista taiteilijaa. Voi herätä kysymys vihjaako tämä esimerkiksi siihen, että Düsseldorfilla ja Firenzellä on suomalaisten taiteilijoiden elämässä jotenkin samanlainen asema vaikkakin eri aikoina. Tällaiseen arvioitiin ei sovellus kuitenkaan suoraan tarjoa keinoja.

Se, että fasetit näyttävät tietyn valinnan tarjoamien yhteyksien määrän ja piilottavat valinnat joille ei löytyisi osumia, on keskeinen sovelluksen ominaisuus. Näistä luvuista voi saada paljon tietoa erityisesti käyttämällä piirakkakaavio-ominaisuutta. Esimerkiksi valitsemalla tyypiksi kunnianosoitukset paikaksi Saksan voi nähdä, että Saksaan liittyy tässä 234 yhteyttä, joka on melko paljon verrattuna moniin muihin maihin. Ulkomaista vain Ruotsiin liittyy enemmän tämän tyyppisiä yhteyksiä. Eniten tällaisia yhteyksiä (8) liityy, kuten ehkä saattaa odottaa, Carl Gustaf Emil Mannerheimiin. Yhteyksien selityksistä näkee, että Mannerheim on saanut saksalaisia mitaleita ensimmäisen ja toisen maailmansodan aikaan. Lisäksi hän on saanut sotien välillä mitaleita Saksan Punaiselta Ristiltä.

[36] http://biografiasampo.fi/henkilo/p265

7 Yhteenveto

7.1 Arviointi

Tässä työssä rakennetulle sovellukselle on tehty vain epämuodollinen arviointi. Sen perusteella sovellus toimii yleisesti ottaen toivotulla tavalla ja antaa järkevän tuntuisia tuloksia. Vaikuttaa myös sieltä, että ainakin maallikko voi oppia uusia asioita kokeilemalla järjestelmää. Esimerkiksi taidehistoriaa tuntematon voi oppia, että Düsseldorf oli 1800-luvulla merkittävä kaupunki suomalaisille taiteilijoille. Epäselväksi jää, voisiko tätä järjestelmää käyttää myös kulttuurihistorian tutkimukseen. Sen arviointi vaatisi yhteistyötä historian tutkijoiden kanssa.

Sovellus on julkisesti kokeiltavissa, mutta kävijämäärät eivät ole ainakaan vielä suuria. Marraskuussa 2018 Biografiasammon yhteyshaku-näkymällä oli Google Analytics -palvelun mukaan 70 katselua ja 49 yksilöityä katselua. Keskimääräinen sivulla käytetty aika oli minuutti ja 24 sekuntia. Kävijämäärät ovat niin pieniä, että on syytä varoa tekemästä kävijöiden käyttäytymisestä liian jyrkkiä päätelmiä keskiarvojen perusteella. Vaikuttaa kuitenkin siltä, että ihmiset eivät käytä sovellusta kovin pitkiä aikoja kerrallaan. Muutaman minuutin tai sen alle oleva käyttöaika tuntuu odotetun mukaiselta. ”Kävijän kulku” -analytiikan mukaan käyttäjä siirtyy yhteyshausta kohtuullisen usein Biografiasammon henkilöä kuvaavalle sivulle. Tämä on odotettua. Jokaisen yhteyden kohdalla tarjotaan linkki tätä henkilöä käsittelevään sivuun. Yllättäen paikkaa käsittelevälle sivulle siirtyminen vaikuttaa olevan hyvin harvinaista, vaikka myös yhteyksiin liittyviin paikkoihin tarjotaan aina linkki. Ei ole selvää, mistä tämä ero johtuu. Ehkä kyse on vain siitä, että elämäkertoja käsittelevän sivun käyttäjälle on luontevaa olla kiinnostuneempi henkilöistä kuin paikoista.

Sovelluksessa tiedon lähde on aina selkeästi ja helposti saatavilla, mikä mahdollistaa lähdekritiikin. Oleellista on kuitenkin ymmärtää, että jonkin yhteyden puutumiseen voi olla valtava määrä eri syitä. Tässä työssä käytetyt lähteet on valittu lähinnä sillä perusteella, mikä sattuu olemaan helposti saatavilla. Esimerkiksi Kansallisgallerian tietokanta ei sisällä kaikkia suomalaisia maalauksia, vaan vain Kansallisgallerian omistamia. Ei välttämättä ole ilmeistä, miksi jokin asia puuttuu tietystä datasta, ja miksi jokin toinen asia on siellä. Lisäksi esimerkiksi henkilöiden tai paikkojen siltaus on voinut epäonnistua joissain tapauksissa, mikä vääristää tuloksia. Demonstraattorin esittämästä yksittäisestä yhteydestä voi tehdä päätelmiä, koska siihen voi yleensä luottaa ja sen luotettavuuden arvioiminen on helppoa. Yhteyden puut-

tumisesta ei kuitenkaan voi päätellä mitään varmuudella. Siten myös yhteyksien suhteellisten määrien visualisoimisen hyödyllisyys jää epäselväksi. Parhaimmillaan niistä voi ehkä saada vihjeitä olemassa olevista lainalaisuuksista tarkempaa tutkimusta varten. Pahimmillaan ne voivat johtaa harhaan, jos käyttäjä ei ymmärrä järjestelmän rajoitteita. Rajoitteita on yritetty kommunikoida käyttäjälle ohjeissa, mutta jää epäselväksi, kuinka helppo tällaisia ohjeita on ymmärtää. Yhteyksien visualisoiminen fasettihaulla vaikuttaa kuitenkin yleisesti ottaen mielenkiintoiselta konseptilta. Esimerkiksi yhteen selkeään aineistoon rajatussa sovelluksessa voisi tällaisista suhteellisten määrien visualisoinneista saada paljonkin irti myös historian tutkimuksen kannalta.

Käyttöjärjestelmässä on kokeilujen perusteella parannettavaa. Tällä hetkellä yhteydet järjestyvät vain yhdellä tavalla, eli aakkosjärjestykseen henkilön ja paikan nimen perusteella. Tämä ei välttämättä ole intuitiivisin tapa järjestää yhteyksiä. Käyttäjä saattaisi esimerkiksi odottaa, että yhteydet järjestyisivät henkilön perusteella ja aikajärjestyksessä saman henkilön yhteyksien joukossa, jolloin henkilön syntymä olisi ensimmäisenä ja kuolema viimeisenä. Luultavasti olisi tarpeen antaa käyttäjälle mahdollisuus järjestää yhteydet haluamallaan tavalla. Käyttäjä saattaa myös kyllästyä lukemaan suurta määrää yhteyksien selityksiä, jotka ovat kaikki luotu samalla kaavalla. Yhteyksien tyyppien visualisointi väreillä ja kuvakkeilla auttaa tätä ongelmaa. Käyttäjä pystyy yhdellä silmäyksellä näkemään paljon asioita, eikä hänen tarvitse lukea jokaista selitystä erikseen, ellei hän ole kiinnostunut tarkemmista tiedoista.

Tässä työssä keskeistä on yhteyksien fasettihaku eli se, että fasettihaulla seulotaan yhteyksiä, eikä esimerkiksi henkilöitä tai paikkoja. Tämä saattaa olla käyttäjille totuttelua vaativa ajatus, koska se poikkeaa hienovaraisesti sellaisesta fasettihausta, johon käyttäjä saattaa olla tottunut. Esimerkiksi jos yhteyden tyypiksi on valittuna ”Romaani kuvaa paikkaa”, saattaisi ammatti-fasetti näyttää numeron 48 ammatin ”professori” perässä. Käyttäjä saattaa olettaa tämän tarkoittavan, että kyseinen valinta rajaisi haun 48:aan professoriin. Näin ei kuitenkaan ole, vaan kyseinen valinta rajaa haun 48:aan yhteyteen. Eli teoriassa on mahdollista, että yksi henkilö jolla on ammatti ”professori”, on kirjoittanut yhden kirjan, joka kuvaa 48:aa eri paikkaa, tai kaksi professoria on kirjoittanut 24 kirjaa kumpikin, joista jokainen kuvaa yhtä paikkaa. Molemmat skenaariot tuottaisivat 48 uniikkia yhteyttä henkilön ja paikan välillä.

Yhteyksien louhiminen tietokannoista on suhteellisen nopeaa. Nopeus riippuu tieto-

kannasta ja yhteyden tyypistä. Eri tietokantojen monimutkaisuudessa ja koossa on suuria eroja. Lisäksi muodostetut hahmot sisältävät eri määriä ylimääräisiä asioita, joiden tarkoitus on helpottaa hahmojen soveltamista eri metadatamalleihin ja tietokantoihin. Tällaiset lisäykset saattavat kuitenkin hidastaa ajoa. Useimmissa tapauksissa hahmojen louhiminen kestää muutamia sekunteja. Joissain tapauksissa useita minuutteja. Hahmojen luomisessa ei ole huolehdittu nopeudesta, joten luultavasti niitä olisi helposti mahdollista optimoida ja nopeuttaa. Taulukko 5 sisältää karkeita arvioita eri tyyppisten hahmojen ajamisen kestosta. Ajat eivät ole suoraan vertailukelpoisia. Osa kyselyistä on ajettu tehokkaalla PC:llä itse ja osa on kohdistettu ulkoiseen tietokantaan, jonka resurssit ja kuormitus voivat vaihdella. Datan muoto ja määrä vaihtelevat myös paljon. Suurin ero liittyy kuitenkin luultavasti hahmoihin vaihtelevasti lisättyihin yleisyyttävä parantaviin piirteisiin. Nämä saattavat hidastaa suoritusta merkittävästi, koska tällöin ei esimerkiksi suoraan sanota, että halutaan juuri tietty ominaisuus, vaan kyseisen ominaisuuden aliominaisuudet kelpaavat myös. Tämän hidastumisen vaikutuksen näkee hyvin siinä miten Biografiasammon tapahtumiin liittyviä yhteyksiä hakevan hahmon, joka sisältää juuri tällaisia muotoiluja, ajon kesto on noin kymmenen minuuttia muiden hahmojen ajoaikojen pysyessä muutamassa sekunnissa. Koska yhteys-instansseja ei tässä luoda reaaliaikaisesti, ei suoritusajalla ole suurta merkitystä. Vaikuttaa kuitenkin siltä, että jos yhteyksiä louhittaisiin reaaliaikaisesti, tulisi hahmojen olla mahdollisimman erikoistuneita ja optimoituja.

Hahmo	**kesto**
BS tapahtumat	590 s
BS kuolemat	11 s
BS syntymät	10 s
Fennica	10 s
Kirjasampo	7 s
HISTO	4 s
Kirjeet	alle 1 s

Taulukko 5: CONSTRUCT-hahmojen ajamiseen kuluva aika

Datan esikäsittely lisää nopeutta käyttötilanteessa, mutta tarkoittaa kasvavaa muistivaatimusta palvelimelle. Tässä sovelluksessa muistin kulutus pysyi kohtuullisena. Yhteysgraafin koko on noin 28,5 megatavua, mikä ei aiheuta ongelmia edes tavalliselle kotitietokoneelle. Voi kuitenkin kuvitella, että joissain sovelluksissa muistivaatimus kasvaisi liian suureksi. Riippuen sovelluksesta tällaisen haun voi tehdä myös

dynaamisesti, jolloin tarvetta esikäsittelyyn ei ole. Käytännössä tällaisessa tapauksessa, jossa tietoa haetaan useista ulkoisista lähteistä, on esikäsittely on helpompaa.

Vaikka esikäsitellyt yhteydet on erittäin nopea hakea, tiettyjä suorituskykyyn liityviä ongelmia säilyy. Jokainen fasetti tekee erikseen kyselyn, joka laskee mahdolliset yhteydet jokaisella vaihtoehdolla. Tämä on hidasta erityisesti kun hierarkisia fasetteja on käytössä, koska ne muodostavat monimutkaisia kyselyjä. Käyttäjä saa valitsemansa yhteydet heti, mutta voi joutua odottamaan useita sekunteja, että jokin fasetti saa laskettua vaihtoehdot ja tulee taas käytettäväksi. Jos sovelluksella olisi useita käyttäjiä samaan aikaan, jotka tekevät raskaita kyselyitä, voisi järjestelmä helposti tukkeutua.

Yhteysien graafin luominen ei ollut erityisen vaativaa. Tarvittavat hahmot on suhteellisen helppo kirjoittaa. Ne piti kuitenkin kirjoittaa jokaista tapausta varten uudelleen, ja jokaisen tietokannan käyttämään yksilölliseen skeemaan piti erikseen tutustua. Lisäksi henkilö- ja paikkaontologioiden käyttö on vaihtelevaa ja riippuu tietokannasta. Ihannetilanteessa, jossa kaikki semanttinen tieto olisi tallennettu yhtenäisten standardien mukaan, olisi kerran luotuja hahmoja helppo soveltaa suoraan kaikiin muihinkin tietokantoihin. Käytännössä näin ei ole. Pelkästään Suomen sisällä kulttuurihistoriaan liittyvää aineistoa on tallennettu käyttäen hyvin erilaisia skeemoja ja ontologioita. Ei vaikuta todennäköiseltä, että olisi helppo luoda hahmoja, joita pystyisi täysin automaattisesti soveltamaan yleisesti. On kuitenkin syytä uskoa, että valmiita hahmoja voi soveltaa pienin muutoksin eri tilanteissa. Olisi ehkä mahdollista jopa luoda jokin järjestelmä, jolla puoliautomaattisesti täydennetään hahmon ominaisuuksia, jolloin hahmo muuttuisi koko ajan yleisemmäksi. Silti vaikka tieto olisi ilmaistu samoilla ominaisuuksilla, voi todellisessa merkityksessä olla hienovaraisia eroja. Esimerkiksi Fennica ilmaisee kirjan tekijää ominaisuudella author eli tekijä. Tuntui aluksi luontevalta antaa tällaisella yhteydelle selitykseksi, että joku henkilö on kirjoittanut kirjan. Tämä tuotti kuitenkin omituisia tuloksia, koska usein tekijä saattoi tarkoittaa esimerkiksi taiteilijaa, joka on kuollut jo kauan ennen kirjan julkaistu, mutta kirja sisältää kuvia hänen taiteestaan. Tässä tapauksessa olisi outoa kuvata taiteilijaa kirjan kirjoittajaksi.

Hahmojen luomisen lisäksi haasteellista on muodostaa sopiva henkilö- ja paikkaontologia. Rajattu määrä henkilöitä pitää yhteyksien määrän kohtuullisena. Liian suuri määrä yhteyksiä johtaa suuriin tiedostoihin ja erityisesti hitaaseen fasettien käyttöön. Oleellista olisi myös henkilöiden valinta siten, että he ovat sopivalla tavalla kiinnostavia käyttökontekstin näkökulmasta. Tässä työssä muodostettiin yh-

teyksiä periaatteessa kaikille Biografiasammon henkilöille, mutta paikoista jätettiin osa pois. Parempi lopputulos saattaisi tulla rajaamalla henkilöt yleisesti kiinnostavimman datajoukon eli kansallisbiografian, henkilöihin. Esimerkiksi pappien datajoukot sisältävät suurimmaksi osaksi henkilöitä, jotka ovat maallikon näkökulmasta liian tuntemattomia ollakseen kiinnostavia. Toisaalta paikkojen rajaamiseen ei ole vastaavaa tarvetta. Harvinaisimpiin paikkoihin muodostuu hyvin vähän yhteyksiä, joten ne eivät oleellisesti hidasta sovelluksen toimintaa, mutta saattavat joissain erikoistapauksissa olla hyvin kiinnostavia. Ongelmalliseksi laajassa paikka-ontologiassa muodostuu paikkojen tunnistaminen pelkkien nimien perusteella, mutta ihannetapauksessa sitä ei tarvitse tehdä.

J. V. Snellmanin koottujen teosten aineiston muuntaminen linkitetyn datan tehtiin melko yksinkertaisella tavalla. Jo yksinkertaisella muunnoksella saatiin kuitenkin selkeitä etuja, ja se mahdollistaa aineiston käytön helposti erilaisissa sovelluksissa. Alkuperäisen web-sivun aineisto oli luotu tietystä näkökulmasta ja tarkoitettu ihmisten selattavaksi. Aineistoon oli kohtuullisen helppo lisätä esimerkiksi kirjoitusten kirjoittajat, mutta jotkut asiat olivat vaikeita. Aineiston luomisessa käytetyn rajoitetun tekniikan vuoksi aineiston luojat ovat tienneet teksteistä paljon enemmän asioita kuin aineistoon lopulta on merkitty. Kirjeiden lähetys- ja vastaanottopaikat saatiin pääteltyä kohtuullisella tarkkuudella noin puolessa tapauksista, mutta on valitettavaa, että koneymmärrettävyyttä ei ole otettu heti huomioon. Kirjeiden lähetys- ja vastaanottopaikkojen mielenkiintoisuus ei ole aivan ilmeistä. Kuitenkin esimerkiksi J. L. Runebergin kohdalle ne tuntuvat hyvin täydentävän kuvaa henkilön elämästä. Sovellus näyttää Runebergille kolme vastaanotettua kirjettä Helsingistä ja kolme lähetettyä kirjettä Porvoosta. Runebergiin liittyy sovelluksen mukaan muuten vain kaksi tapahtumaa Helsinkiin ja kaksi Porvooseen. Kirjeiden paikat korostavat Helsingin ja Porvoon merkitystä Runebergin elämään tavalla, joka tuntuu oikealta.

Tunnettuja puutteita sovelluksessa liittyy sekä käyttöliittymään että dataan. Kuten yllä on todettu, käyttöliittymässä olisi varaa parantaa. Käytettävyyttä myös parantaisi ontologioiden parantaminen. Esimerkiksi ammattiontologia on tällä hetkellä epäselvä, ja sen yksinkertaistaminen auttaisi yhteyksien suodattamista ammatin perusteella. Fasetteja voisi ehkä myös lisätä. Esimerkiksi sukupuolta ja aikaa kuvaavien fasettien lisääminen saattaisi parantaa sovellusta. Päivämäärän kohdalla pitäisi päättää mitä tehdään sillon kuin yhteydellä ei suoraan ole päivämäärää. Saman henkilön esiintyminen kahdesti henkilöontologiassa aiheuttaa joissain tapauksissa ongelmia. Tämä vaikuttaa harvinaiselta, mutta varmasti paremman lopputu-

loksen saisi luomalla henkilöontologian, jossa jokaista henkilöä vastaa yksiselitteisesti yksi entiteetti. Lisäksi tällä hetkellä myös samanlaisen yhteyden löytyminen eri lähteistä tuottaa useampia yhteys-entiteettejä, mikä ei välttämättä ole toivottavaa. Esimerkiksi sama kirja voi löytyä sekä Fennican, että Kirjasammon aineistosta. Nyt niistä syntyy kaksi eri yhteyttä, joiden oleellinen ero on niiden lähde. Tämä on harvinaista, mutta näkyy esimerkiksi joidenkin Mika Waltarin teosten kohdalla. Toivottavimpi käyttäytyminen olisi ehkä, että syntyisi yksi yhteys, mutta tällä yhteydellä olisi useampi lähde.

7.2 Lopuksi

Tässä työssä esitettiin tietämykseen perustuva metodi kulttuurihistoriaan liittyvien henkilöiden ja paikkojen välisten yhteyksien louhimiseen. Lisäksi esitettiin fasettihakuun perustuva menetelmä tällaisten yhteyksien hakemiseen. Avoimesta datasta luotiin SPARQL CONSTRUCT -hahmojen avulla yhteyksiä kuvaava graafi, ja näiden yhteyksien hakemiseen luotiin fasettihakua käyttävä web-sovellus.

Ensimmäinen tutkimuskysymys koski yhteyksien louhimiseen käytettyjen hahmojen luomisen helppoutta ja niiden yleistettävyyttä. Voidaan todeta, että yhteyksiä kuvaavien muotojen luominen on suhteellisen yksinkertaista, mutta hahmojen soveltaminen laajemmin erilaisiin aineistoihin vaatisi jonkin verran työtä. Kulttuurihistorian aineistot eivät yleensä noudata täysin samoja standardeja. Siksi hahmoja luultavasti joutuu muokkaamaan tapauskohtaisesti. Jokaista yhteystyyppiä varten täytyy myös luoda erikseen oma hahmonsa, mikä tekee erityisen monimutkaisten yhteyksien louhimisen työlääksi. Koska monimutkaisten yhteyksien louhiminen on työlästä, monet erityisen yllättävät ja kiinnostavat yhteydet jäävät helposti löytämättä. Toisaalta tämän metodin etu on se, että yhteyksiä löytyy vain rajallinen määrä, ja ne ovat keskimäärin mielenkiintoisempia kuin satunnaiset yhteydet. Kiinnostavia yhteyksiä on siten helpompi löytää. Etuna on lisäksi se, että yhteydelle voidaan antaa luonnollisen kielen selitys.

Toinen tutkimuskysymys koski fasettihaun soveltamista yhteyshakuun. Fasettihaun soveltaminen yhteyksiin tarjoaa uuden ja kiinnostavan näkökulman aineistoon. Fasettihaku ei ole pelkkä metodi kyselyn muodostamiseen, vaan se tarjoaa tietoa haettavien asioiden lukumääristä tietyillä vaihtoehdoilla, ja näiden lukumäärien avulla käyttäjä voi suunnata hakuaan. Suhteelliset lukumäärät tietyillä valinnoilla voivat myös itsessään olla mielenkiintoisia. Tässä työssä toteutettu sovellus tarjoaa mahdollisuuden hakea yhteyksiä fasettihaun avulla. Yhteyksiä on siksi mahdollista et-

siä myös eri tavoin määritellyille laajemmille kokonaisuuksille, eikä vain yksilöille. Tällaisessa fasettihaussa voi tarkastella myös yhteyksien suhteellisia määriä tietyillä valinnoilla. Tällaiset vertailut voivat tarjota yllättävän mielenkiintoisia näkökulmia aineistoon. Tällaista yhteyshakua ei ilmeisesti aiemmin ole toteutettu. Tämä johtuu luultavasti siitä, että yhteyksiä on aineistossa helposti niin paljon, että fasettihaun toteuttaminen niiden hakemiseen ilman rajoituksia vaatisi liikaa resursseja. Tässä työssä käytetty metodi pitää yhteyksien määrän rajallisena, ja siten mahdollistaa fasettihaun ilman käyttäjän kannalta kohtuuttomia viiveitä.

Vaatisi lisää tutkimusta selvittää, kuinka helposti tällainen yhteyksien fasettihaku olisi toteutettavissa muissa tapauksissa. Esimerkiksi henkilöiden välisten yhteyksien yhteyshaku saattaisi tuottaa merkittävästi suuremman määrän yhteyksiä. Yhteyksien suuri määrä lisää palvelimen muistivaatimusta, mutta ei vaikuta mahdottomalta toteuttaa tällaista hakua sopivassa tapauksessa dynaamisesti. Suurempi ongelma saattaa olla se, että fasettihaku hidastuu liikaa suurilla yhteyksien määrillä.

Jatkotutkimukselle on tarvetta niin fasettihaun soveltamisessa kuin yhteyksien louhimisen tavan yleistämisessä. Tällä hetkellä eri tietokannat harvoin noudattavat yhtenäisiä standardeja. Siksi yleistäminen on vaikeaa. Saattaisi kuitenkin olla mahdollista kehittää jonkinlainen helppokäyttöinen menetelmä hahmojen täydentämiseen. Vielä kunnianhimoisempi tavoite voisi olla käyttää neuroverkkoihin perustuvaa oppivaa järjestelmää, joka opetettaisiin valmiilla hahmoilla, ja joka niiden perusteella voisi automaattisesti tunnistaa vastaavien asioiden kuvauksen tapoja aineistoista. Myös yhteyksien suhteellisten määrien visualisoiminen erilaisilla tilastoilla saattaisi tuottaa tutkimuksen kannalta mielenkiintoisia tuloksia. Metodin soveltaminen muunlaisiin yhteyksiin vaatisi lisää tutkimusta.

Lähteet

1 K. Athukorala, D. Głowacka, G. Jacucci, A. Oulasvirta, and J. Vreeken. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651, February 2016.

2 E. Berndtson. *Politiikka tieteenä - Johdatus valtio-opilliseen ajatteluun*. Valtion painatuskeskus, Helsinki, 1992.

3 T. Berners-Lee. Linked data. https://www.w3.org/DesignIssues/LinkedData.html, 2006.

4 T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

5 C. Bizer, T. Heath, and T. Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

6 G. Carothers and E. Prud'hommeaux. RDF 1.1 Turtle. W3C recommendation, W3C, Feb. 2014. http://www.w3.org/TR/2014/REC-turtle-20140225/.

7 G. Cheng, F. Shao, and Y. Qu. An empirical evaluation of techniques for ranking semantic associations. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):2388–2401, 2017.

8 G. Cheng, Y. Zhang, and Y. Qu. Explass: exploring associations between entities via top-k ontological patterns and facets. In *International Semantic Web Conference*, pages 422–437. Springer, 2014.

9 M. Doerr. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.

10 M. Doerr. Ontologies for cultural heritage. In *Handbook on Ontologies*, pages 463–486. Springer, 2009.

11 J. English, M. Hearst, R. Sinha, K. Swearingen, and K. Lee. Flexible search and navigation using faceted metadata. Technical report, University of Berkeley, School of Information Management and Systems, 2003.

12 L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: explaining relationships between entity pairs. *Proceedings of the VLDB Endowment*, 5(3):241–252, 2011.

13 V. Fionda and G. Pirro. Explaining and querying knowledge graphs by relatedness. *Proceedings of the VLDB Endowment*, 10(12):1913–1916, 2017.

14 W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57, 1992.

15 A. Gangemi. Ontology design patterns for semantic web content. In *International semantic web conference, The Semantic Web – ISWC 2005*, pages 262–276. Springer, 2005.

16 A. Gangemi and V. Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

17 L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

18 N. Guarino, D. Oberle, and S. Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

19 M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, 2002.

20 P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. RelFinder: Revealing relationships in RDF knowledge bases. In *International Conference on Semantic and Digital Media Technologies. SAMT 2009.*, pages 182–187. Springer, 2009.

21 P. Heim, S. Lohmann, and T. Stegemann. Interactive relationship discovery via the semantic web. In *The Semantic Web: Research and Applications: 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 –June 3, 2010, Proceedings, Part I*, volume 6088 of *Lecture Notes in Computer Science*, pages 303–317. Springer, 2010.

22 K. Hypén and E. Mäkelä. An ideal model for an information system for fiction and its application: Kirjasampo and Semantic Web. *Library Review*, 60(4), April 2011.

23 E. Hyvönen. *Publishing and using cultural heritage linked data on the semantic web.* Morgan & Claypool Publishers, 2012.

24 E. Hyvönen. *Semanttinen web.* Gaudeamus, Helsinki, 2018.

25 E. Hyvönen, O. Alm, and H. Kuittinen. Using an ontology of historical events in semantic portals for cultural heritage. In *Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007)*, November 2007. `https://seco.cs.aalto.fi/publications/2007/hyvonen-et-al-history-2007.pdf`.

26 E. Hyvönen, M. Alonen, E. Ikkala, and E. Mäkelä. Life stories as event-based linked data: Case semantic national biography. In *Proceedings of ISWC 2014 Posters & Demonstrations Track*, pages 1–4. CEUR Workshop Proceedings, October 2014.

27 E. Hyvönen, P. Leskinen, M. Tamper, J. Tuominen, and K. Keravuori. Semantic national biography of finland. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, pages 372–385. CEUR Workshop Proceedings, March 2018.

28 E. Hyvönen, T. Lindquist, J. Törnroos, and E. Mäkelä. History on the semantic web as linked data-an event gazetteer and timeline for World War I. In *Proceedings of CIDOC 2012 - Enriching Cultural Heritage, Helsinki, Finland*. CIDOC, June 2012. `https://seco.cs.aalto.fi/publications/2007/hyvonen-et-al-history-2007.pdf`.

29 E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, et al. CultureSampo–finnish culture on the semantic web 2.0. thematic perspectives for the end-user. In *Museums and the Web 2009. Selected Papers from an international conference*, April 2009. `https://www.museumsandtheweb.com/mw2009/papers/hyvonen/hyvonen.html`.

30 E. Hyvönen, T. Ruotsalo, T. Häggström, M. Salminen, M. Junnila, M. Virkkilä, M. Haaramo, E. Mäkelä, T. Kauppinen, and K. Viljanen. Culturesampo–finnish culture on the semantic web: The vision and first results. In *Developments in Artificial Intelligence and the Semantic Web - Proceedings of the 12th Finnish AI Conference STeP 2006*, volume 2, pages 63–72. Finnish Artificial Intelligence Society, October 2006.

31 E. Hyvönen, S. Saarela, and K. Viljanen. Application of ontology techniques to view-based semantic search and browsing. In *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, pages 92–106. Springer, 2004.

32 E. Hyvönen, J. Tuominen, M. Alonen, and E. Mäkelä. Linked data finland: A 7-star model and platform for publishing and re-using linked datasets. In *The Semantic Web: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 226–230. Springer, May 2014.

33 E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä. Building a national semantic web ontology and ontology service infrastructure–the FinnONTO approach. In *The Semantic Web: Research and Applications. ESWC 2008.*, volume 5021 of *Lecture Notes in Computer Science*, pages 95–109. Springer, June 2008.

34 M. Kaminskas, I. Fernández-Tobías, F. Ricci, and I. Cantador. Knowledge-based music retrieval for places of interest. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 19–24. ACM, November 2012.

35 T. Kauppinen, R. Henriksson, R. Sinkkilä, R. Lindroos, J. Väätäinen, and E. Hyvönen. Ontology-based disambiguation of spatiotemporal locations. In *Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*. CEUR Workshop Proceedings, June 2008. `https://seco.cs.aalto.fi/publications/2008/kauppinen-et-al-spatiotemporal.pdf`.

36 K. J. Kochut and M. Janik. SPARQLeR: Extended SPARQL for semantic association discovery. In *The Semantic Web: Research and Applications. European Semantic Web Conference 2007.*, pages 145–159. Springer, 2007.

37 M. Koho, E. Heino, and E. Hyvönen. SPARQL Faceter - client-side faceted search based on SPARQL. In *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop*. CEUR Workshop Proceedings, May 2016. `http://ceur-ws.org/Vol-1615/semdevPaper5.pdf`.

38 B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 313–322. ACM, June 2009.

39 J. Kurki and E. Hyvönen. Relational semantic search: Searching social paths on the semantic web. *Poster Proceedings of the International Semantic Web Conference (ISWC 2007), Busan, Korea*, November 2007. `https://seco.cs.aalto.fi/publications/2007/kurki-hyvonen-relational-2007.pdf`.

40 O. Lassila, R. R. Swick, et al. Resource Description Framework (RDF) model and syntax specification. `https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`, 1999.

41 J. Lehmann, J. Schüppel, and S. Auer. Discovering unknown connections - the DBpedia relationship finder. In *Proceedings of the 1st Conference on Social Semantic Web (CSSW 2007)*, pages 99–110, September 2007.

42 P. Leskinen, E. Hyvönen, and J. Tuominen. Analyzing and visualizing prosopographical linked data based on biographies. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, volume 2119,

pages 39–44. CEUR Workshop Proceedings, July 2018. http://ceur-ws.org/Vol-2119/paper7.pdf.

43 P. Leskinen, J. Tuominen, E. Heino, and E. Hyvönen. An ontology and data infrastructure for publishing and using biographical linked data. In *Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017)*. CEUR Workshop Proceedings, October 2017. https://seco.cs.aalto.fi/publications/2017/leskinen-et-al-biographies-2017.pdf.

44 B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems and their Applications*, 15(5):47–55, 2000.

45 S. Lohmann, P. Heim, and P. Díaz. Exploiting the semantic web for interactive relationship discovery in technology enhanced learning. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pages 302–306. IEEE, 2010.

46 S. Lohmann, P. Heim, T. Stegemann, and J. Ziegler. The RelFinder user interface: interactive exploration of relationships between objects of interest. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 421–422. ACM, February 2010.

47 E. Mäkelä. Survey of semantic search research. In *Proceedings of the seminar on knowledge management on the Semantic Web*. Department of Computer Science, University of Helsinki, Helsinki, 2005. https://seco.cs.aalto.fi/publications/2005/makela-semantic-search-2005.pdf.

48 E. Mäkelä, K. Hypén, and E. Hyvönen. BookSampo - lessons learned in creating a semantic portal for fiction literature. In *The Semantic Web – ISWC 2011*, volume 7032 of *Lecture Notes in Computer Science*, pages 173–188. Springer, October 2011.

49 E. Mäkelä, K. Hypén, and E. Hyvönen. Improving fiction literature access by linked open data-based collaborative knowledge storage-the BookSampo project. In *78th IFLA General Conference and Assembly*, August 2012.

50 E. Mäkelä, K. Hypén, and E. Hyvönen. Fiction literature as linked open data - the booksampo dataset. *Semantic Web*, 4(3):299–306, 2013.

51 E. Mäkelä, E. Hyvönen, and T. Ruotsalo. How to deal with massively heterogeneous cultural heritage data–lessons learned in CultureSampo. *Semantic Web*, 3(1):85–109, 2012.

52 G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, October 2006.

53 N. Marie and F. Gandon. Survey of linked data based exploration systems. In *Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014)*, October 2014.

54 K. McGarry. A survey of interestingness measures for knowledge discovery. *The knowledge engineering review*, 20(1):39–61, March 2005.

55 Y. Miao, J. Qin, and W. Wang. Graph summarization for entity relatedness visualization. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164. ACM, August 2017.

56 E. Miller and F. Manola. RDF primer. W3C recommendation, W3C, February 2004. `"http://www.w3.org/TR/2004/REC-rdf-primer-20040210/"`.

57 K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

58 G. Pirrò. Explaining and suggesting relatedness in knowledge graphs. In *The Semantic Web - ISWC 2015, 14th International Semantic Web Conference*, pages 622–639. Springer, October 2015.

59 G. Pirrò and A. Cuzzocrea. RECAP: building relatedness explanations on the web. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 235–238. International World Wide Web Conferences Steering Committee, April 2016.

60 E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. W3C recommendation, W3C, January 2008. `http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`.

61 J. Saarti and K. Hypén. From thesaurus to ontology: the development of the kaunokki finnish fiction thesaurus. *The Indexer*, 28(2):50–58, 2010.

62 R. Savolainen, editor. *J. V. Snellman: Kootut teokset 1–24*. Opetus- ja kulttuuriministeriö, Helsinki, 2002–2004.

63 D. Seo, H. K. Koo, S. Lee, P. Kim, H. Jung, and W.-K. Sung. Efficient finding relationship between individuals in a mass ontology database. In *International Conference on U-and E-Service, Science and Technology*, pages 281–286. Springer, December 2011.

64 A. Sheth, B. Aleman-Meza, I. B. Arpinar, C. Bertram, Y. Warke, C. Ramakrishanan, C. Halaschek, K. Anyanwu, D. Avant, F. S. Arpinar, et al. Semantic association identification and knowledge discovery for national security applications. *Journal of Database Management (JDM)*, 16(1):33–53, 2005.

65 A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, volume 95, pages 275–281. AAAI Press, August 1995.

66 G. Tartari and A. Hogan. WiSP: Weighted shortest paths for RDF graphs. In *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA 2018)*. CEUR Workshop Proceedings, vol. 2187, October 2018.

67 D. Tunkelang. *Faceted search*. Morgan & Claypool Publishers, 2009.

68 J. Tuominen, E. Hyvönen, and P. Leskinen. Bio CRM: A data model for representing biographical data for prosopographical research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, pages 59–66. CEUR Workshop Proceedings, November 2017. `http://ceur-ws.org/Vol-2119/paper10.pdf`.

69 N. Voskarides, E. Meij, M. Tsagkias, M. De Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 564–574. Association for Computational Linguistics, July 2015.

70 R. Wilén and M. Holopainen. "Älä jätä käyttämättä sattuman tarjoamia mahdollisuuksia" –serendipisyys tiedonhaun ilmiönä. *Informaatiotutkimus*, 36(2), 2017. https://doi.org/10.23978/inf.65195.

71 Y. Zhang, G. Cheng, and Y. Qu. Towards exploratory relationship search: A clustering-based approach. In *Semantic Technology, Third Joint International Semantic Technology Conference, JIST 2013*, pages 277–293. Springer, November 2013.