Sotahistorian kuvaaminen ja rikastaminen linkitettynä datana

Erkki Heino

Pro gradu -tutkielma
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 25. kesäkuuta 2017

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Erkki Heino

Työn nimi — Arbetets titel — Title

Sotahistorian kuvaaminen ja rikastaminen linkitettynä datana

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	25. kesäkuuta 2017	56

Tiivistelmä — Referat — Abstract

Linkitetty data mahdollistaa erillisten aineistojen yhdistämisen, mistä syntyvä kokonaisuus mahdollistaa aineistojen tietojen paremman ymmärtämisen. Aineistojen välisten linkkien avulla voidaan päätellä uutta tietoa helpommin kuin tarkastelemalla aineistoja erikseen.

Tutkielmassa käsitellään sotahistoriallisten aineistojen mallintamista ja julkaisua linkitettynä avoimena datana sekä aineistojen automaattista rikastamista muiden aineistojen avulla. Työn tavoitteena oli selvittää miten tällaisia aineistoja kannattaa mallintaa linkitettynä datana, miten niitä kannattaa yhdistää muihin aineistoihin, mitä lisäarvoa tästä saadaan ja miten aineistot kannattaa visualisoida.

Aineistoina käytettiin tietokirjoista digitoituja tapahtumia sekä Sotamuseon SA-kuvapalvelun valokuvien metatietoja. Aineistot mallinnettiin käyttäen CIDOC CRM -standardia ja niitä rikastettiin linkittämällä niiden sisältämiä resursseja automaattisesti henkilö-, joukko-osasto- ja paikkaontologioiden avulla. CIDOC CRM:n määrittämä tapahtumakeskeinen mallinnustapa mahdollistaa aineistojen yhteentoimivuuden paitsi toistensa myös muiden historiallisten aineistojen kanssa.

Automaattiseen rikastamiseen liittyi monia haasteita, sillä viittaukset toisiin aineistoihin oli poimittava suurelta osin tekstimuotoisista kuvauksista, jolloin ongelmaksi nousee nimettyjen entiteettien kuten henkilöiden ja paikkojen tunnistaminen ja yksilöinti tekstistä. Työssä käsitellään kyseisiä haasteita, esitellään käytetyt ratkaisut ja arvioidaan näiden toimivuutta.

Aineistoja visualisoimaan toteutettiin myös JavaScript-sovellukset. Aineistot ja sovellukset on julkaistu osana Sotasampo-portaalia, joka muodostaa yhteenlinkitetyn kokonaisuuden erilaisista aineistoista liittyen toiseen maailmansotaan Suomessa. Portaali palvelee paitsi Suomen historiasta ja sodissa taistelleiden omaistensa liikkeistä kiinnostuneita kansalaisia, myös historian tutkijoita tarjoamalla aineistot vapaasti kyseltävässä rakenteisessa muodossa.

ACM Computing Classification System (CCS):
Information systems → Resource Description Framework (RDF)
Theory of computation → Data modeling
Information systems → Entity resolution

Avainsanat — Nyckelord — Keywords
linkitetty data, RDF, SPARQL, CIDOC CRM, nimettyjen entiteettien linkitys, visualisointi

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Toisesta maailmansodasta löytyy paljon tietoa World Wide Webistä (WWW). Tämä tieto on kuitenkin pääasiassa dokumenttipohjaista ihmisten luettavaksi tarkoitettua sisältöä. Siinä missä ensimmäisestä maailmansodasta on useita koneluettavia datajulkaisuja – esimerkiksi Europeana Collections 1914–1918[1], 1914–1918 Online[2], WW1 Discovery[3], Out of the Trenches[4], CENDARI[5], Muninn[6] ja WW1LOD [MTLH16] – koneluettavassa muodossa olevaa dataa toisesta maailmansodasta ei WWW:stä löydy juuri lainkaan. Koneluettava data mahdollistaisi erilaisten sovellusten toteuttamisen sitä käyttäen ja erilaisten työkalujen käytön datan analysoinnissa.

Dataa voidaan julkaista koneluettavasti web-ohjelmointirajapintojen (Web API) avulla. Näiden käyttö on kuitenkin rajattu niihin toimintoihin, jotka rajapinnan toteuttaja on määrittänyt. Lisäksi tällaiset rajapinnat ovat usein suljettuja järjestelmiä, ja niiden sisältämien resurssien tunnisteet ovat järjestelmän sisäisiä eivätkä siis yleismaailmallisesti yksilöiviä. Yleiskäyttöisemmin dataa voidaan julkaista linkitettynä datana [HB11]. Tällöin dataan voidaan kohdistaa kyselyitä standardoidun SPARQL-kyselykielen avulla, ja kullekin resurssille voidaan antaa tunniste, joka yksilöi sen globaalisti. Näin järjestelmän resursseihin voidaan viitata sen ulkopuolelta.

Sotahistoriallinen tieto on lupaava käyttötapaus linkitetylle datalle: se on heterogeenistä, hajautunut eri maihin ja kirjoitettu eri kielillä [HHL+16]. Esimerkiksi tietoa toisen maailmansodan tapahtumista voi löytää tietokirjoista, tietoa sodassa taistelleista Kansallisarkistosta tai Wikipediasta ja valokuvia SA-kuva-arkiston verkkopalvelusta[7]. Eri aineistojen tuominen yhteen linkitettynä datana muodostaa resurssien välisten linkkien kautta kontekstin, joka mahdollistaa uuden tiedon muodostamisen [NDO05].

Tämän pro gradu -työn tarkoituksena on tuottaa koneluettavaa dataa toisen maailmansodasta mallintamalla ja julkaisemalla asiaankuuluvaa dataa linkitettynä datana, rikastaa tätä muiden aineistojen avulla ja visualisoida tulos loppukäyttäjille. Tällaisista aineistoista ja visualisoinneista voivat olla kiinnostuneita paitsi historian tutkijat myös esimerkiksi kansalaiset, joilla on sodassa taistelleita sukulaisia. Tutkimuskysymykset ovat seuraavat:

1. Miten sotahistoriallisia aineistoja kannattaa mallintaa linkitettynä datana?

2. Miten käytössä olevat aineistot kannattaa yhdistää muihin aineistoihin?

[1] `http://www.europeana-collections-1914-1918.eu`
[2] `http://www.1914-1918-online.net`
[3] `http://ww1.discovery.ac.uk`
[4] `http://www.canadiana.ca/en/pcdhn-lod/`
[5] `http://www.cendari.eu/research/first-world-war-studies/`
[6] `http://blog.muninn-project.org`
[7] `http://sa-kuva.fi`

3. Mitä lisäarvoa aineistojen yhdistämisestä muihin aineistoihin saavutetaan?

4. Miten mallinnetut aineistot kannattaa visualisoida?

Tutkimusmenetelmä perustuu suunnittelutieteeseen (design science), jossa tutkimuskysymyksiä pyritään ratkaisemaan toteuttamalla ongelma-alueessa hyödyllisiä artefakteja ja arvioimalla näitä [HMPR04]. Näiden artefaktien, kuten mallien, prosessien tai järjestelmien, hyödyllisyys määrittää tutkimuksen onnistumisen. Tutkimuskysymysten vastaamiseksi toteutettiin sotahistoriallisen tiedon datajulkaisu linkitettynä datana ja sovellus visualisoimaan tuotettua dataa. Työ koostuu siis neljästä kokonaisuudesta: 1) sotahistoriallisten aineistojen mallintaminen linkitetyn datan periaatteiden mukaisesti, 2) aineistojen rikastaminen linkittämällä ne muihin aineistoihin, 3) linkitysten evaluointi ja 4) aineistoja visualisoivan sovelluksen toteuttaminen ja arviointi.

Lopputuloksena syntyvälle järjestelmälle voidaan erottaa kaksi pääasiallista käyttäjäryhmää: Suomen historiasta kiinnostuneet kansalaiset ja historian tutkijat. Näistä tutkijoita voidaan palvella avaamalla historiallista aineistoa monipuolisesti kyseltävässä muodossa. Historiasta ylipäänsä kiinnostuneita käyttäjiä ajatellen helppokäyttöinen aineistoja visualisoiva sovellus on tarpeellinen.

Aineistojen mallinnusta ja sovelluksen kehittämistä ohjaamaan erotettiin järjestelmälle seuraavat käyttötapaukset:

1. Käyttäjä haluaa löytää tietoa Suomen vaiheista toisen maailmansodan aikana.

2. Käyttäjä on kiinnostunut sodanajan valokuvista ja haluaa hakea niistä häntä kiinnostavia kuvia.

3. Käyttäjä haluaa tutkia yksittäisen henkilön – esimerkiksi sukulaisensa – tarinaa sotien aikana.

4. Tutkija haluaa muodostaa uutta tietoa yhdistämällä tietoa eri lähteistä.

Työ toteutettiin osana Sotasampo-projektia [HHL+16, HTM+15]. Sotasampo on portaali, jossa julkaistaan tietoa toisesta maailmansodasta avoimena linkitettynä datana ja tarjotaan sovelluksia tämän pohjalta. Työssä suunniteltiin ja toteutettiin talvi- ja jatkosodan tapahtumien sekä valokuvien ontologiat ja sovellukset näiden visualisoimiseksi. Aineistoja rikastettiin linkittämällä ne Sotasammon toimija- ja paikkaontologioihin.

Luvussa 2 esitellään yleisesti linkitetty data ja sen teknologioita. Luku 3 käsittelee tapahtumaperustaista mallinnusta ja tapahtumien kuvaamiseen tarkoitettuja metadatamalleja. Luvussa 4 käsitellään entiteettien yksilöimistä tekstistä. Luvussa 5 esitellään työssä käytetyt aineistot ja luvussa 6 näiden

mallintaminen linkitettynä datana. Luku 7 esittää aineistojen muunnokset alkuperäisestä muodosta linkitetyn datan tietomalliin ja luku 8 aineistojen rikastamisen toisiin aineistoihin linkittämällä. Luvussa 9 esitellään aineistojen julkaisu ja luvussa 10 aineistoja visualisoivat verkkosovellukset. Tulokset arvioidaan suhteessa tutkimuskysymyksiin luvussa 11 ja yhteenveto on esitetty luvussa 12.

2 Linkitetty data

Perinteisesti dataa on säilötty tiettyä käyttötarkoitusta varten erillisissä siiloissa, jotka on toteutettu yleensä relaatiotietokantoina. Siiloissa käytetään usein resurssien yksilöintiin sisäisiä tunnisteita, jotka eivät ole universaalisti uniikkeja. Tästä syystä eri siilojen resurssien välille ei yleensä ole mahdollista määrittää suoria yhteyksiä. Koska eri lähteet tarjoavat erilaisia rajapintoja tietojen noutamiseksi, on ohjelmistot suunniteltava käytettävien rajapintojen mukaan [Biz09].

Linkitetty data on joukko käytäntöjä datan julkaisemiseksi ja yhdistämiseksi WWW:ssä [BL06, BHBL09]. Sen tarkoituksena on ehkäistä datan pirstaloitumista ja parantaa datan käytettävyyttä. Linkitetyn datan neljä periaatetta ovat seuraavat [BL06]:

1. Käytä URI-tunnisteita asioiden nimeämiseen.
2. Käytä HTTP-URI-tunnisteita, jotta nimetyistä asioista voi löytää lisätietoa.
3. Tarjoa hyödyllistä tietoa RDF- ja SPARQL-standardeja käyttäen, kun tunnisteen perusteella haetaan lisätietoa.
4. Sisällytä dataan linkkejä toisiin URI-tunnisteisiin, jotta näiden kautta voidaan löytää muita resursseja.

Tim Berners-Lee [BL06] johtaa linkitetyn datan periaatteista viiden tähden mallin avoimen datan julkaisun laadun arvioimiseksi. Mallissa datajulkaisua voidaan arvioida antamalla sille tähtiä sen mukaan, kuinka helposti se on käytettävissä – esimerkiksi kuinka avoimessa tiedostomuodossa data on saatavilla:

★	Data on julkaistu avoimella lisenssillä missä tahansa muodossa.
★★	Data on julkaistu strukturoidussa muodossa (esimerkiksi Microsoft Excel -taulukkona).
★★★	Data on julkaistu avoimessa strukturoidussa muodossa (esimerkiksi CSV-muodossa).

★★★★ Datan resurssit on yksilöity URI-tunnisteilla.

★★★★★ Data on linkitetty muihin aineistoihin kontekstin luomiseksi.

Hyvönen ja kumppanit [HTAM14] täydentävät mallia kahdella lisätähdellä:

★★★★★★ Datan malli on eksplisiittisesti määritelty ja dokumentoitu.

★★★★★★★ Datan ja tämän mallin vastaavuus on varmistettu.

Linkitetyn datan palveluista kenties kenties suurimmat ovat DBpedia[8] ja Wikidata[9]. DBpedia tarjoaa Wikipedian sisällön linkitettynä datana [HL15], ja Wikidata on muun muassa Wikipedian käyttämä kaikkien muokattavissa oleva tietokanta [VK14].

Linkitetyn datan periaatteet mahdollistavat resurssien yksilöimisen globaalisti, jolloin niihin voidaan viitata ja tällä tavoin laajentaa niihin liittyvää tietoa alkuperäisen järjestelmän ulkopuolelta. Linkitetty data muodostaa näin maailmanlaajuisen dataverkon, semanttisen webin [HB11].

2.1 Semanttinen web

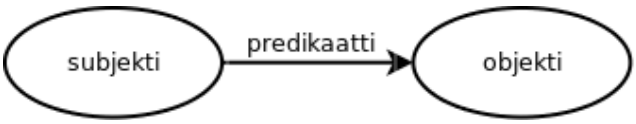
Siinä missä World Wide Web (WWW) koostuu pääasiassa ihmisten luettavaksi tarkoitetuista dokumenteista, semanttinen web koostuu koneluettavasta tiedosta [SHBL06, HB11]. Perinteisesti WWW:ssä URI-tunnisteet viittaavat dokumentteihin, kun taas semanttisessa webissä mihin tahansa asiaan voi viitata tunnisteella. Semanttisen webin tarkoituksena on parantaa tiedon löydettävyyttä ja käytettävyyttä. Ihminen tunnistaa helposti perinteisestä dokumentista sen sisältämän tiedon, esimerkiksi verkkosivulla olevan ruokareseptin. Reseptien tunnistaminen tietokoneen avulla automaattisesti on kuitenkin monimutkaisempaa, koska resepti voidaan visuaalisesti esittää monella tavalla. Jos resepti ja sen ainekset merkitään eksplisiittisin tunnistein, on reseptin sisällön erottaminen automaattisesti dokumentin muusta sisällöstä helppoa. Tällöin siis tietosisältö ja sen merkitys (semantiikka) erotetaan esitystavasta.

2.2 Resource Description Framework

Resource Description Framework (RDF) on malli tiedon esittämiseen World Wide Webissä [WLC14].

RDF:n perusyksikkö on kolmikko, joka koostuu subjektista, predikaatista ja objektista. RDF-kolmikot muodostavat RDF-graafeja [WLC14] — kuvassa 1 on visualisoitu graafi, joka koostuu kahdesta solmusta (subjekti ja objekti) ja niistä yhdistävästä predikaatista eli RDF-kolmikosta.

[8] `http://wiki.dbpedia.org/`
[9] `https://www.wikidata.org/`



Kuva 1: RDF-kolmikko [WLC14].

RDF-graafien sarjallistamiseen on olemassa useita syntakseja, kuten XML-pohjainen RDF/XML[10] ja helpommin ihmisten luettavissa oleva Turtle[11].

2.3 SPARQL

SPARQL on RDF-muotoiseen dataan kohdistuvien kyselyiden muodostamiseen tarkoitettu kieli ja protokolla [CFTW13]. Sen avulla RDF-tietokannasta voidaan hakea tietoa joustavasti. SPARQL-protokolla perustuu HTTP:n käyttöön [CFTW13], joten SPARQL-palvelupistettä voidaan käyttää suoraan internetin välityksellä. Aineiston julkaiseminen SPARQL-palvelupisteen avulla mahdollistaakin sen monipuolisen käytön.

SPARQL:n mahdollistamat mielivaltaiset kyselyt RDF-tietokantaan ovat paitsi mahdollisuus myös ongelma. Suuren joukon tuloksia palauttavat tai laskennallisesti monimutkaiset kyselyt voivat vaatia palvelupisteeltä paljon resursseja. Tällaisilla kyselyillä voidaan toteuttaa (tai tahattomasti aiheuttaa) palvelunestohyökkäys, jos palvelupiste ei rajoita kyselyiden suorittamista [CFTW13]. Esimerkiksi DBpedian julkinen SPARQL-palvelupiste rajoittaa palautettujen tulosten määrää eikä suorita kyselyitä, jotka arvioidaan liian raskaiksi [Wil11].

3 Tapahtumaperustainen mallinnus

Mallit jäsentävät käsiteltävää sovellusalaa ja auttavat ymmärtämään mallinnuksen alaista dataa. Jotta aineistoja voidaan hyödyntää tehokkaasti tietoteknisesti, on ne kuvattava formaalisti eli mallinnettava. Tapahtumat ovat keskeisiä historian ja kulttuuriperintöaineistojen kuvaamisessa [vMS+11], joten niiden mallintaminen on tämän työn kannalta tärkeässä asemassa. Tällaisia aineistoja on haastavaa käsitellä tarkasti yhtenäisellä tavalla, koska ne ovat hyvin kirjavia, ja historiallinen tieto on usein epätäydellistä ja mahdollisesti ristiriitaista. Yleiset sanastot kuten Dublin Core[12] kuvaavat aineistoja liian yleisellä tasolla, jolloin tietoa katoaa. Toisaalta erilaisten aineistojen yhteensovittaminen on vaikeaa, jos yhteistä sanastoa tai rakennetta ei ole [Doe03].

[10] https://www.w3.org/TR/rdf-syntax-grammar/
[11] https://www.w3.org/TR/turtle/
[12] `http://dublincore.org/`

Ontologia on filosofiassa tutkimusala, joka tutkii olemisen perimmäistä olemusta. Tietojenkäsittelytieteessä termi tarkoittaa mallia, joka formaalisti määrittää jonkin joukon entiteettejä ja niiden välisiä suhteita yhteisesti sovitulla tavalla [GOS09].

Tapahtumien mallintamiseen on olemassa useita metatietomalleja, kuten LODE [Sha10], Event Ontology [RA07], Simple Event Model (SEM) [vMS+09] ja CIDOC Conceptual Reference Model (CIDOC CRM) [Doe03].

LODE [Sha10] on yksinkertainen ontologia tapahtumien julkaisuun linkitettynä datana. Se koostuu vain yhdestä luokasta (*Event*, tapahtuma) ja seitsemästä predikaatista.

Event Ontology [RA07] on musiikkitapahtumien kontekstissa kehitetty ontologia tapahtumien kuvaamiseen, joskaan sitä ei ole rajattu mihinkään sovellusalaan. Event Ontology määrittelee kolme luokkaa ja seitsemän predikaattia. Malli asettaa vähän rajoituksia ja käyttää muita ontologioita ominaisuuksien arvojen määrittelyssä.

SEM [vMS+11] pyrkii määrittelemään tapahtumien kuvaamiseen mallin, joka asettaa mahdollisimman vähän semanttisia rajoituksia. Tällä tavoin se pyrkii ottamaan huomioon WWW:n monimuotoiset aineistot. Malli ei esimerkiksi määritä pakollisia ominaisuuksia tai ominaisuuksien kardinaliteetteja. SEM mahdollistaa myös periaatteessa minkä tahansa tyyppijärjestelmän käytön aineistoja mallinnettaessa.

CIDOC CRM [Doe03] on kulttuuriperintöaineistojen mallintamiseen tarkoitettu malli. Se on muita yllä esiteltyjä malleja huomattavasti monimutkaisempi sisältäen 94 luokkaa ja 168 ominaisuutta. CIDOC CRM on ISO-standardi (21127:2014)[13], ja esimerkiksi The British Museum käyttää sitä kokoelmajulkaisussaan [The11]. CIDOC CRM:n tavoitteena on olla yleinen kulttuuriperintöaineista harmonisoiva malli, eikä se siis ole vain tapahtumia varten, mutta tapahtumat ovat siinä kuitenkin keskeisessä roolissa. Esimerkiksi henkilön syntymä mallinnetaan syntymätapahtumana (mallissa luokka *E67 Birth*), jonka osallisena henkilö on, eikä esimerkiksi henkilöilmentymän syntymäaika- ja -paikkaominaisuuksina. Mallia ei ole kuitenkaan suunniteltu linkitetyn datan lähtökohdasta, joten kaikki sen määrittelyt eivät ole suoraan yhteensopivia RDF-tietomallin kanssa. Erityisesti malli määrittää ominaisuuksien ominaisuuksia, kuten tapahtuman osallistujan rooli, eikä tällaisten mallintaminen RDF:n avulla ole suoraviivaista.

4 Nimettyjen entiteettien linkitys

Mallintamisen lisäksi toinen tärkeä kokonaisuus tämän työn kannalta on aineistojen rikastaminen luomalla linkkejä muihin ontologioihin. Esimerkiksi tapahtumien ja valokuvien kuvausteksteissä mainitaan henkilöitä, paikkoja

[13] `http://www.iso.org/iso/catalogue_detail?csnumber=57832`

ja joukko-osastoja. Näiden yksilöiminen ja yhdistäminen vastaaviin ontologioihin mahdollistaa uuden tiedon löytämisen ja esimerkiksi henkilöiden sodanajan vaiheiden automaattisen koostamisen.

Aineistojen suuren koon vuoksi tekstien läpikäynti käsin ei ollut käytännössä mahdollista. Työ oli siis automatisoitava, ja luonnollisen kielen käsittelyyn tarvittiin kieliteknologisia tekniikoita. Luonnollisessa kielessä esiintyvien nimimainintojen yksilöimistä ja yhdistämistä tietämyskannassa oleviin tietueisiin kutsutaan nimettyjen entiteettien linkitykseksi (named entity linking, entity linking, named entity disambiguation) [HRN+13, LME12, Cuc07, LWH+13, BP06, HSZ11].

Nimetty entiteetti tarkoittaa entiteettiä, jolle on olemassa yksi tai useampi *kiinteä nimittäjä* (rigid designator), joka poimii kyseisen entiteetin kaikista mahdollisista maailmoista [NS07]. Näin esimerkiksi lauseessa "Mannerheim on tasavallan presidentti" "Mannerheim" on nimetty entiteetti mutta "tasavallan presidentti" ei.

Hachey ja kumppanit [HRN+13] jakavat nimettyjen entiteettien linkittämisen kolmeen osaan tai komponenttiin: 1) erotin (extractor), 2) etsijä (searcher) ja 3) yksilöijä (disambiguator). *Erotin* erottaa linkitettävät entiteetit tekstistä, *etsijä* hakee tietämyskannasta mahdolliset linkityskohteet eli kandidaatit ja *yksilöijä* valitsee kandidaateista oikean tai hylkää ne, jos oikeaa kohdetta ei löytynyt.

Entiteettien erottaminen tekstistä ei ole täysin suoraviivaista erityisesti suomen kaltaisen synteettisen kielen tapauksessa [Mä14]. Naiivi tapa olisi yksinkertaisesti muodostaa tekstistä eri mittaisia katkelmia eli n-grammeja ja käyttää näitä *etsijän* syötteenä. Suomen kielen taivutusmuotojen takia tämä tapa ei kuitenkaan tuottaisi hyvää tulosta, sillä tekstissä mainitut entiteettien nimet eivät näin useinkaan olisi perusmuodossa. Viitattavissa ontologioissa entiteettien nimet sen sijaan ovat useimmiten perusmuodossa, joten myös tekstin viittaukset tarvitaan tässä muodossa. Entiteettien tunnistamiseksi teksti on siis muutettava perusmuotoon, minkä jälkeen tekstin osia on verrattava linkitettävään käsitteistöön.

Perusmuotoistaminen ei sekään ole aivan yksinkertaista. Jokaisen sanan perusmuotoistaminen ei aina tuota haluttua tulosta, sillä esimerkiksi joukko-osastoon viittaavan maininnan "Kannaksen armeijan" naiivi perusmuotoistaminen tuottaa sanat "Kannas armeija", vaikka haluttu muoto onkin sanaliitto "Kannaksen armeija".

Han ja kumppanit [HS11] erottavat kaksi entiteettien linkittämiseen liittyvää ongelmaa: nimien vaihtelevaisuus (name variation) ja nimien monitulkintaisuus (name ambiguity). Nimien vaihtelevaisuudella tarkoitetaan sitä, että samasta entiteetistä voidaan käyttää useampaa eri nimeä. Esimerkiksi Suomen marsalkka Carl Gustaf Emil Mannerheimiin voidaan viitata esimerkiksi käsitteillä "sotamarsalkka Mannerheim" ja "Marski". Nimien monitulkintaisuus taas tarkoittaa sitä, että samalla nimellä voidaan viitata useampaan eri entiteettiin, jolloin konteksti määrittää viittauksen kohteen.

Esimerkiksi Haapamäki-nimisiä paikkoja on Maanmittauslaitoksen Paikannimirekisterissä 242 kappaletta.

Entiteettien linkittämisessä on yksilöinnissä usein käytetty Wikipediaa [HRN+13]. Esimerkiksi Bunescu ja Paşca [BP06] käyttivät Wikipedian uudelleenohjaus- ja yksikäsitteistämissivuja sekä kategorioita apuna linkityksessä.

5 Käytetyt aineistot

Koska valmista linkitettyä dataa toisesta maailmansodasta ei juuri ole saatavilla, oli data luotava ja mallinnettava itse. Tätä varten oli löydettävä sopivat lähteet linkitetyn tietokannan luomiseksi.

Tapahtumien tietokirjalähteiksi valikoituivat tietokirjat Talvisodan pikkujättiläinen [LJ06] ja Jatkosodan pikkujättiläinen [LJ05]. Näiden lisäksi Aalto-yliopiston semanttisen laskennan tutkimusryhmässä poimittiin Kansallisarkiston digitoimista Suomen puolustusvoimien sodanajan organisaatiokortistosta 642 taistelua [HHL+16, Les16]. Käytettävissä oli myös huomattava määrä metatietoa SA-kuva-arkiston sodanaikaisista valokuvista. Taulukossa 1 on esitetty lähtöaineistojen koot.

Aineisto	**Koko**	**Lähde**
Sotatapahtumat	1045 tapahtumaa	Tietokirjat [LJ06, LJ05]
Taistelut	642 taistelua	Kansallisarkisto
Valokuvat	163 783 valokuvaa	SA-kuva-arkisto

Taulukko 1: Aineistojen koot.

Seuraavissa kappaleissa esitellään käytetyt aineistot tarkemmin.

5.1 Sotatapahtumat

Talvi- ja jatkosodan tapahtumista ei ollut saatavilla strukturoidussa muodossa olevaa sähköistä aineistoa, joten se oli tuotettava ei-strukturoiduista lähteistä. Talvisodan osalta kirjassa Talvisodan pikkujättiläinen [LJ06] on listattu sotien tapahtumia luokiteltuna poliittisiksi, sotilaallisiksi, kotirintamaan ja pommituksiin liittyviksi sekä muiksi tapahtumiksi. Vastaavasti Jatkosodan pikkujättiläisessä [LJ05] on listattu jatkosodan aikaisia tapahtumia luokiteltuna poliittisiin tapahtumiin, sotatapahtumiin Suomessa, sotatapahtumiin maailmalla ja muihin tapahtumiin.

Tietokirjatapahtumien lisäksi Suomen Puolustusvoimien organisaatiokortistosta eristettiin joukko-osastojen taisteluita Sotasammon toimijaontologian

kokoamisen yhteydessä [Les16]. Taistelutapahtumia luotiin tällä tavoin 642 kappaletta.

5.2 SA-kuva-arkisto

SA-kuva-arkisto on Puolustusvoimien Kuvakeskuksen tuottama kokoelma rintamamiesten ottamia kuvia vuosilta 1939–1945 [Puo]. Valokuvia on yli 160 000, ja ne ovat rintamamiesten ja Puolustusvoimien tiedotuskomppanian kuvaajien ottamia. Valokuvat dokumentoivat taistelutilanteita sekä sotahistoriallista ja kansantieteellistä aineistoa.

Valokuvien metatietoja oli saatu Puolustusvoimien kuvapalvelulta taulukkomuodossa, minkä lisäksi tietoja oli saatavilla SA-kuva-arkiston verkkorajapinnan kautta. Kuviin liittyviä historiallisesti mielenkiintoisia tietoja olivat kuvaaja, ajankohta, paikka, kuvia luokitteleva kuvaselostusnumero, sanallinen kuvaus ja lisäkommentit. Näiden tietojen lisäksi taulukoissa oli tietoa valokuvien digitoinnista, kuten ajankohta ja käytetty laitteisto.

Taulukkomuotoisissa tiedoissa oli saatavilla enemmän tietoa kuin rajapinnasta, mutta ainoastaan rajapinnan tiedoissa oli valokuvien URL-osoitteet, minkä lisäksi tiedot osasta valokuvia – erityisesti värilliset kuvat – oli saatavilla ainoastaan rajapinnan kautta.

6 Aineistojen mallinnus

Jotta käsiteltävänä olevat tapahtuma- ja valokuva-aineistot voitiin julkaista linkitettynä datana, oli ensin päätettävä, miten ne mallinnettaisiin. Mallille erotettiin seuraavat vaatimukset korostaen mallinnettujen aineistojen hyödyllisyyttä ja yleiskäyttöisyyttä:

1. Aineistojen mallin tulee olla sellainen, että sitä hyödyntäen aineisto on helposti käytettävissä sovelluskehityksessä.
2. Mallin tulee tukea aineiston rikastamista muita aineistoja käyttäen.
3. Mallin tulee olla sellainen, että aineiston tietoja on helppo kysellä sen avulla.

Aineistojen mallit voitaisiin johtaa suoraviivaisesti lähtöaineistojen rakenteesta. Luokat ja ominaisuudet voitaisiin määrittää mallin sisäisesti ja riippumattomasti. Tällainen aineistosidonnainen malli vaikeuttaisi kuitenkin sen yhdistämistä muihin malleihin. Tarkoituksena on mallintaa aineistot paitsi Sotasammon sisäisesti yhteensopivasti myös yleistä yhteiskäyttöisyyttä silmällä pitäen.

CIDOC CRM ei ole rajoitettu ainoastaan tapahtumiin, vaan sitä voidaan käyttää muidenkin kulttuuriperintöaineistojen mallintamiseen. Tämän vuoksi mallia voidaan käyttää muidenkin liittyvien aineistojen mallinnukseen, jolloin

kokonaisuus voidaan pitää mahdollisimman yhtenäisenä. Sotasammmon aineistot onkin mallinnettu CIDOC CRM:n mukaisesti [HHL+16], ja näistä syistä tapahtumien mallintamiseen valittiin CIDOC CRM.

CIDOC CRM soveltuu myös valokuvien mallintamiseen, joten yhtenäisyyden vuoksi sen käyttäminen on perusteltua. Tällainen malli on suhteellisen monimutkainen, koska jokaiseen kuvaan liittyy esimerkiksi tapahtuma kuvan ottamisesta. Yksinkertaisemmin valokuvien metatiedot voitaisiin mallintaa käyttämällä yleisiä sanastoja kuten Dublin Corea, jolloin jokainen valokuva voidaan kuvata yhtenä resurssina.

Seuraavassa käydään läpi kunkin aineiston mallinnus ja perusteet mallien valintaan.

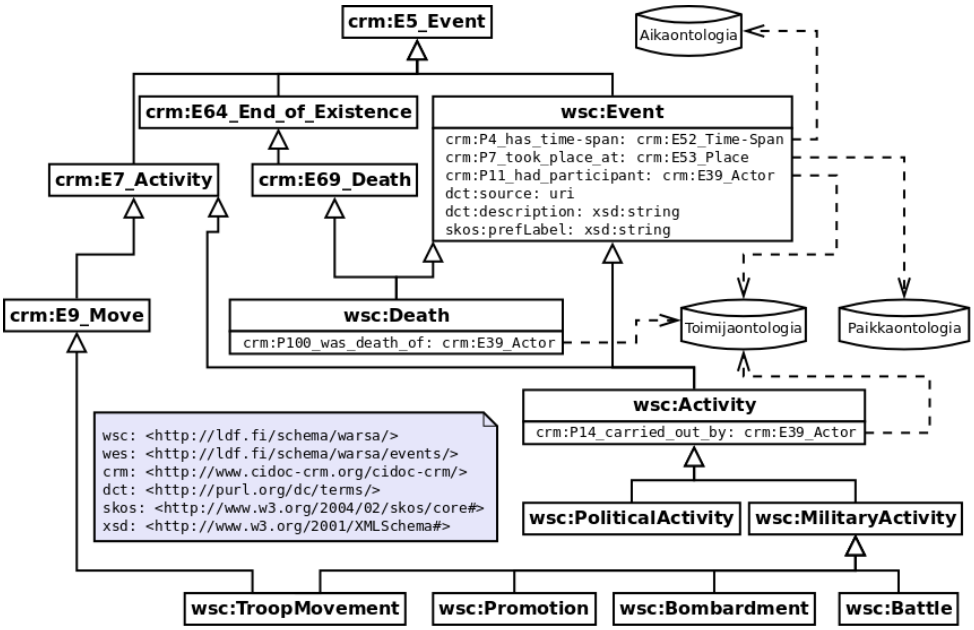
6.1 Sotatapahtumien malli

Sotatapahtumat mallinnettiin linkitettynä datana käyttäen CIDOC CRM -mallia [Doe03]. Tapahtumien luokittelussa käytettiin CIDOC CRM:iin sisältyville luokille *E5 Event*, *E7 Activity* ja *E69 Death* määriteltyjä alaluokkia. Nämä alaluokat ovat *Tapahtuma (Event)*, *Toiminta (Activity)* ja *Kuolema (Death)*. *Toiminta*-luokalle määriteltiin lisäksi alaluokat *Poliittinen toiminta (PoliticalActivity)* ja *Sotatoimi (MilitaryActivity)*, joista jälkimmäiselle määriteltiin myös neljä alaluokkaa tarkempaa luokittelua varten: *Ylennys (Promotion)*, *Pommitus (Bombardment)*, *Taistelu (Battle)* ja *Joukkojen liikutus (TroopMovement)*. Käytetty malli on esitetty kuvassa 2. Kuvassa yläluokkien ominaisuuksia ei ole esitetty alaluokissa, mutta kaikki yläluokkien ominaisuudet periytyvät myös alaluokille.

Tapahtumien mallinnettavat, suoraan lähtöaineistosta saatavat tiedot olivat luokan lisäksi päivämäärä, tyyppi ja lähde. Taulukossa 2 on esitetty nämä tiedot ja niiden kiinnitys RDF-predikaateiksi. Taulukossa esiintyvä nimiavaruuslyhenne *dct* viittaa Dublin Core -sanastoon ja *skos* Simple Knowledge Organization System (SKOS) -sanastoon[14]. Näitä standardeja käytetään yleisen yhteensopivuuden vuoksi ominaisuuksissa kuten resurssien ihmisluettavissa nimissä (*skos:prefLabel*). Tapahtumille ei ollut muuta kuvaavaa nimeä kuin itse tapahtuman kuvausteksti, joten *skos:prefLabel* ja resurssia kuvaileva ominaisuus *dct:description* saivat saman arvon.

Edellä mainittujen tietojen lisäksi mallinnuksessa otettiin huomioon tapahtumiin liittyvät toimijat ja paikat, jotka myöhemmässä vaiheessa tunnistettiin tapahtumakuvauksista. Tapahtuman toimijat kuvataan CIDOC CRM:n määrittämillä ominaisuuksilla *P11 had participant* ja *P14 carried out by*. Näistä ensimmäinen kuvaa tapahtumassa jollain tavalla osallisena olevaa toimijaa, ja jälkimmäinen tapahtumaan vaikuttavaa aktiivista toimijaa. *P14 carried out by* on ominaisuuden *P11 had participant* alaominaisuus, joten ensin mainittu ominaisuus implikoi jälkimmäisen [CDG+15] — toisin

[14] `http://www.w3.org/2004/02/skos/core#`



Kuva 2: Tapahtumien malli.

Tieto	**RDF-predikaatti**
Tapahtuman tyyppi	rdf:type
Tapahtuman päivämäärä	crm:P4_has_time-span
Tapahtuman kuvaus	dct:description, skos:prefLabel
Tapahtuman lähde	dct:source

Taulukko 2: Tapahtumien lähtöaineiston tiedot.

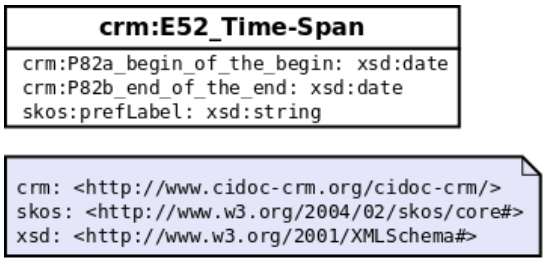
sanoen toiminnan suorittaja on loogisesti myös osallisena kyseisessä toiminnassa. Ominaisuuksien arvona on toimijaontologian resurssiin viittaava URI-tunniste. Näiden lisäksi kuolematapahtumassa kuollut henkilö kuvataan ominaisuudella *P100 was death of*. Tapahtuman paikka kuvataan ominaisuudella *P7 took place at*, jonka arvo on viittaus paikkaontologian resurssiin. Tapahtumien ajankohdat mallinnettiin CIDOC CRM:n mukaisesti, jolloin tapahtuman ajankohtaa määrittävän ominaisuuden *P4 has time-span* arvo on aikaväliä kuvaavan resurssin URI. Esimerkki mallinnetusta tapahtumasta sarjallistettuna Turtle-muodossa on esitetty kuvassa 3.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wac: <http://ldf.fi/warsa/actors/> .
@prefix wev: <http://ldf.fi/warsa/events/> .
@prefix wsc: <http://ldf.fi/schema/warsa/> .
@prefix wso: <http://ldf.fi/warsa/sources/> .
@prefix wti: <http://ldf.fi/warsa/events/times/> .

wev:event_1039 a wsc:Death ;
    dct:source wso:source12 ; # Jatkosodan pikkujättiläinen
    crm:P11_had_participant wac:actor_12455 ,  # 8. Divisioona
        wac:person_69 ; # Claes Winell
    crm:P100_was_death_of wac:person_69 ; # Claes Winell
    crm:P4_has_time-span wti:time_1943-01-09-1943-01-09 ; # 9.1.1943
    crm:P7_took_place_at <http://ldf.fi/warsa/places/municipalities/m_place_509> ; # Helsinki
    dct:description """8. Divisioonan komentaja, kenraalimajuri Claes Winell kuoli Helsingissä
        sydänkohtaukseen."""@fi ;
    skos:prefLabel """8. Divisioonan komentaja, kenraalimajuri Claes Winell kuoli Helsingissä
        sydänkohtaukseen."""@fi .
```

Kuva 3: Mallinnettu tapahtuma.

Tapahtumien ajankohdat eivät aina ole esitettävissä yhtenä päivämääränä, ja toisaalta niiden tarkka ajankohta ei välttämättä ole tiedossa. Tästä syystä ajankohdat ovat CIDOC CRM -mallissa oma luokkansa (*E52 Time-Span*), jolloin aikavälit ja epätäydellinen tieto voidaan esittää formaalisti. Ajankohdat eivät ole yksinkertaisia päivämääräarvoja tapahtuman ominaisuutena, vaan kukin ajankohta on oma resurssinsa. Jokainen ajankohtaa kuvaava resurssi on CIDOC CRM -mallissa määritellyn *E52 Time-Span* -luokan ilmentymä. Itse aika-arvot määritellään kyseisen luokan ominaisuuksilla *P81 ongoing throughout* ja *P82 at some time within*, joista ensimmäinen kuvaa aikavälin vähimmäis- ja jälkimmäinen enimmäiskestoa. Koska tapahtumien ja valokuvien ajat tiedetään suurimmalta osin päivän tarkkuudella, käytettiin ominaisuutta *P82 at some time within*. Aikavälien kuvaaminen yhden RDF-predikaatin avulla ei kuitenkaan ole tyydyttävä ratkaisu, sikäli kuin aikavälejä halutaan voida verrata toisiin esimerkiksi SPARQL-kyselyissä [CID11]. Resurssilla on siis CIDOC CRM:n suosituksen [CID11] mukaisesti aikamääreen aikaisinta alkamisaikaa merkitsevä *P82a begin of the begin* ja viimeistä mahdollista päättymisaikaa merkitsevä *P82b end of the end* -predikaatti. Näiden arvo on tyyppiä *xsd:date* eli päivämäärä. Lisäksi resurssilla on aikavälin ihmisluettavasti kuvaava merkkijono ominaisuuden *skos:prefLabel* arvona. Aikavälit muodostavat oman ontologiansa, joiden yk-

silöihin tapahtumat viittaavat. Aikavälien malli on esitetty kuvassa 4, ja kuvassa 5 on esimerkki mallinnetusta ajasta (10.12.1939) Turtle-muodossa.



Kuva 4: Aikojen malli.

```
@prefix : <http://ldf.fi/warsa/events/times/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

:time_1939-10-12-1939-10-12 a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1939-10-12"^^xsd:date ;
    crm:P82b_end_of_the_end "1939-10-12"^^xsd:date ;
    skos:prefLabel "1939-10-12" .
```

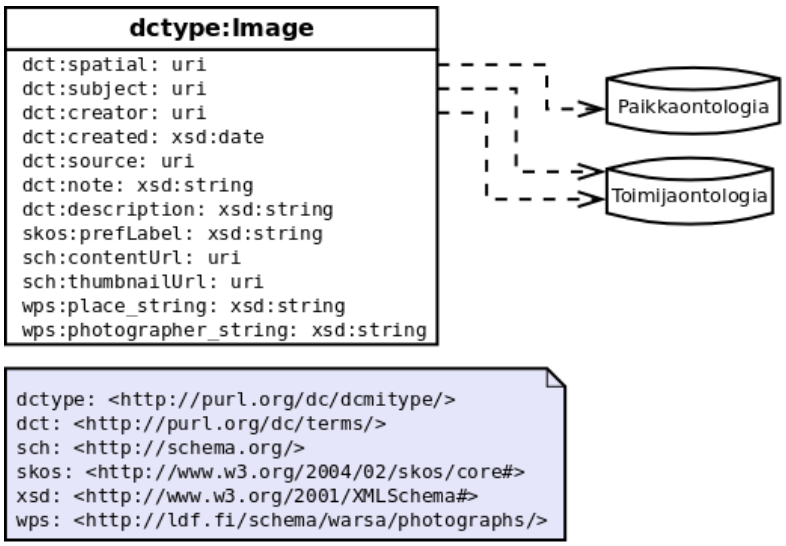
Kuva 5: Mallinnettu aika.

6.2 Valokuvien metatietojen malli

Käytetyn valokuva-aineiston tiedot koskevat digitoituja valokuvia. Tästä syystä aineisto voitaisiin kuvata eri tarkkuuden tasoilla: kaikkein tarkin ja mahdollisimman paljon tietoa sisältävä malli huomioisi digitoimisen ja pitäisi esimerkiksi konkreettiset valokuvat erillisinä olioina digitoiduista kuvista. Toisaalta käyttötapauksen kannalta tällainen erottelu ei välttämättä ole tarpeellinen ja mielekäs.

Valokuvien metatietojen mallintamisen perustana voitaisiin käyttää yleisiä sanastoja, kuten Dublin Corea, jolloin yhden valokuvan metatiedot voitaisiin kuvata yhtenä resurssina ja malli olisi mahdollisimman yksinkertainen. Tällöin lähtöaineisto voitaisiin yksinkertaisesti muuntaa RDF-muotoon sellaisenaan vaihtaen ainoastaan tietueiden nimet sopiviksi sanastoista löytyviksi predikaateiksi. Tällainen tapa mallintaa valokuvat on esitetty kuvassa 6. Mallissa valokuvaaja määritellään ominaisuudella *dct:creator* ja valokuvassa esiintyvät toimijat ominaisuudella *dct:subject*. Valokuvauspaikkaa kuvaa ominaisuus *dct:spatial* ja valokuvan päiväystä *dct:created*. Kuvausteksti kuvataan ominaisuudella *dct:description*, mahdollinen kommentti ominaisuudella *skos:note* ja tietojen lähde ominaisuudella *dct:source*. Digitoidun valokuvan ja sen esikatselukuvan verkko-osoitteet määritellään Schema.org-sanaston ominaisuuksien *sch:contentUrl* ja *sch:thumbnailUrl* avulla. Alkuperäiset merk-

kijonomuotoiset tiedot valokuvaajasta ja paikasta kuvataan mallin sisäisillä predikaateilla *wps:photographer_string* ja *wps:place_string*.



Kuva 6: Valokuvien metatietojen mallinnus yleisiä sanastoja käyttäen.

CIDOC CRM:n käyttö tuottaa yhtenäisen mallin tapahtuma-aineiston kanssa ja mahdollistaa aineiston hyvin yksityiskohtaisen kuvaamisen. Alkuperäinen valokuva (tai sen negatiivi) voidaan mallintaa syntyneeksi valokuvanottamistapahtumasta. Digitoiminen voitaisiin kuvata omana tapahtumanaan, joka synnyttää digitoidun valokuvan, joka taas on konkreettisen valokuvan esitysmuoto. Aineiston kuvaaminen näin yksityiskohtaisesti tekisi mallista kuitenkin hyvin monimutkaisen. Koska käyttötarkoituksena on esittää tietoa historiallisista tapahtumista ja valokuvista, ei digitoimista kuvattu mallissa. Valokuvat päätettiin kuvata CIDOC CRM:n avulla. Malli on huomattavasti monimutkaisempi kuin edellä esitetty; valokuvan ottaminen kuvataan tapahtumana, joka synnyttää valokuvailmentymän. Vaikka CIDOC CRM:n mukainen malli onkin monimutkaisempi, on se myös rikkaampi, ja siitä voidaan saada lisäarvoa: esimerkiksi, koska jokaisesta kuvasta luodaan valokuvaustapahtuma, voidaan henkilöiden toimista muodostaa aikajana yksinkertaisesti hakemalla kaikki henkilöön liittyvät tapahtumat. Sen sijaan erilaisten mallien käytöstä seuraisi, että henkilön aikajanaa luotaessa olisi kunkin mallin mukaiset tiedot muodostettava erikseen.

Malli koostuu kahdesta luokasta: *Photograph* ja *Photography*. *Photograph* kuvaa valokuvaa ja on CIDOC CRM:n luokkien *E38 Image* ja *E31 Document* alaluokka. *Photography* kuvaa valokuvaustapahtumaa ja on CIDOC CRM -luokan *E65 Creation* sekä tapahtumien mallinnuksessa määritellyn toimintaa kuvaavan *Activity*-luokan alaluokka. Se on samanlainen edellisessä aliluvussa esitettyjen tapahtumaluokkien kanssa: ominaisuus *P7 took place*

at määrittää valokuvauspaikan, *P4 has time-span* ajankohdan ja *P11 had participant* valokuvaustapahtumaan osallistuneet toimijat. Valokuvaaja on valokuvauksessa aktiivinen toimija, joten hänet kuvataan ominaisuudella *P14 carried out by*. Valokuvaustapahtumalla on ominaisuus *P94 has created*, joka viittaa itse valokuvaresurssiin.

Photograph-luokka kuvaa valokuvaa ja sisältää linkit digitoidun kuvan ja tämän pienemmän esikatseluversion verkko-osoitteisiin (*sch:contentUrl* ja *sch:thumbnailUrl*). Valokuvassa esiintyvät toimijat kuvataan ominaisuuden *P138 represents* avulla.

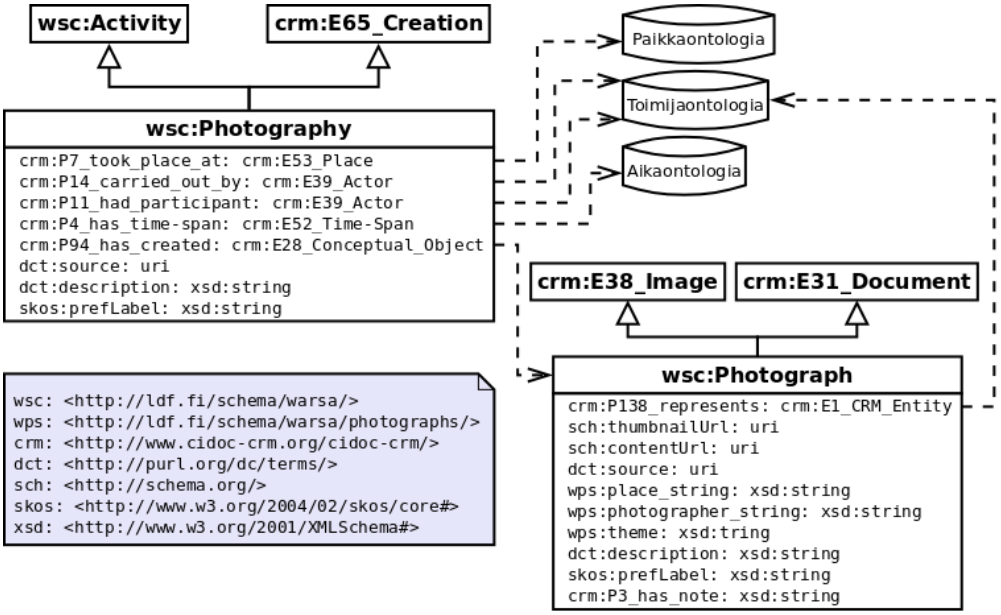
Tieto	**RDF-predikaatti**
Kuvan URL	sch:contentUrl
Esikatselukuvan URL	sch:thumbnailUrl
Valokuvan ottaja	wps:photographer_string (crm:P14_carried_out_by)
Päivämäärä, jolloin valokuva on otettu	crm:P4_has_time-span
Paikka, jossa valokuva on otettu	wps:place_string (crm:P7_took_place_at)
Valokuvan kuvaus	dct:description, skos:prefLabel
Kuvaselostusnumero	wps:theme
Kommentteja valokuvasta tai sen negatiivista	crm:P3_has_note
Valokuvan lähde (SA-kuva)	dct:source

Taulukko 3: Valokuvien lähtöaineiston tiedot.

Lähtöaineiston tiedot ja niitä vastaavat RDF-predikaatit on esitetty taulukossa 3. Sulut predikaatin ympärillä taulukossa tarkoittavat, että kyseisen predikaatin arvoa ei saada suoraan lähtöaineiston tiedosta, vaan johdetaan tämän avulla linkitysvaiheessa. Esimerkiksi valokuvien tiedoissa valokuvaajan nimi on merkkijono, joka myöhemmässä vaiheessa yksilöitiin viittaukseksi toimijaontologian henkilöön. Taulukon predikaatit, joiden nimiavaruuslyhenne on *wps*, ovat mallin sisäisiä, lähtöaineiston alkuperäisiä merkkijonomuotoisia arvoja säilöviä predikaatteja. Taulukossa esiintyvä *kuvaselostusnumero* on SA-kuva-arkiston sisäinen valokuvia luokitteleva tunnus. Valokuvaustapahtuman sanallinen kuvaus (*dct:description*) on sama kuin itse valokuvan. Valokuvan ja valokuvaustapahtuman ihmisluettavana nimenä (*skos:prefLabel*) käytetään myös molempien luokkien tapauksessa kuvaustekstiä, koska muuta kuvaavaa nimeä ei valokuville ole.

Valokuva-aineiston mallinnus on esitetty kuvassa 8. Esimerkki SA-kuva-arkiston valokuvasta on esitetty kuvassa 7 ja kyseiseen valokuvaan liittyvät mallinnetut tiedot sarjallistettuna Turtle-muodossa kuvassa 9. Tapahtumat ja valokuvat muodostavat yhteensopivan kokonaisuuden, joka on esitetty kuvassa 10.

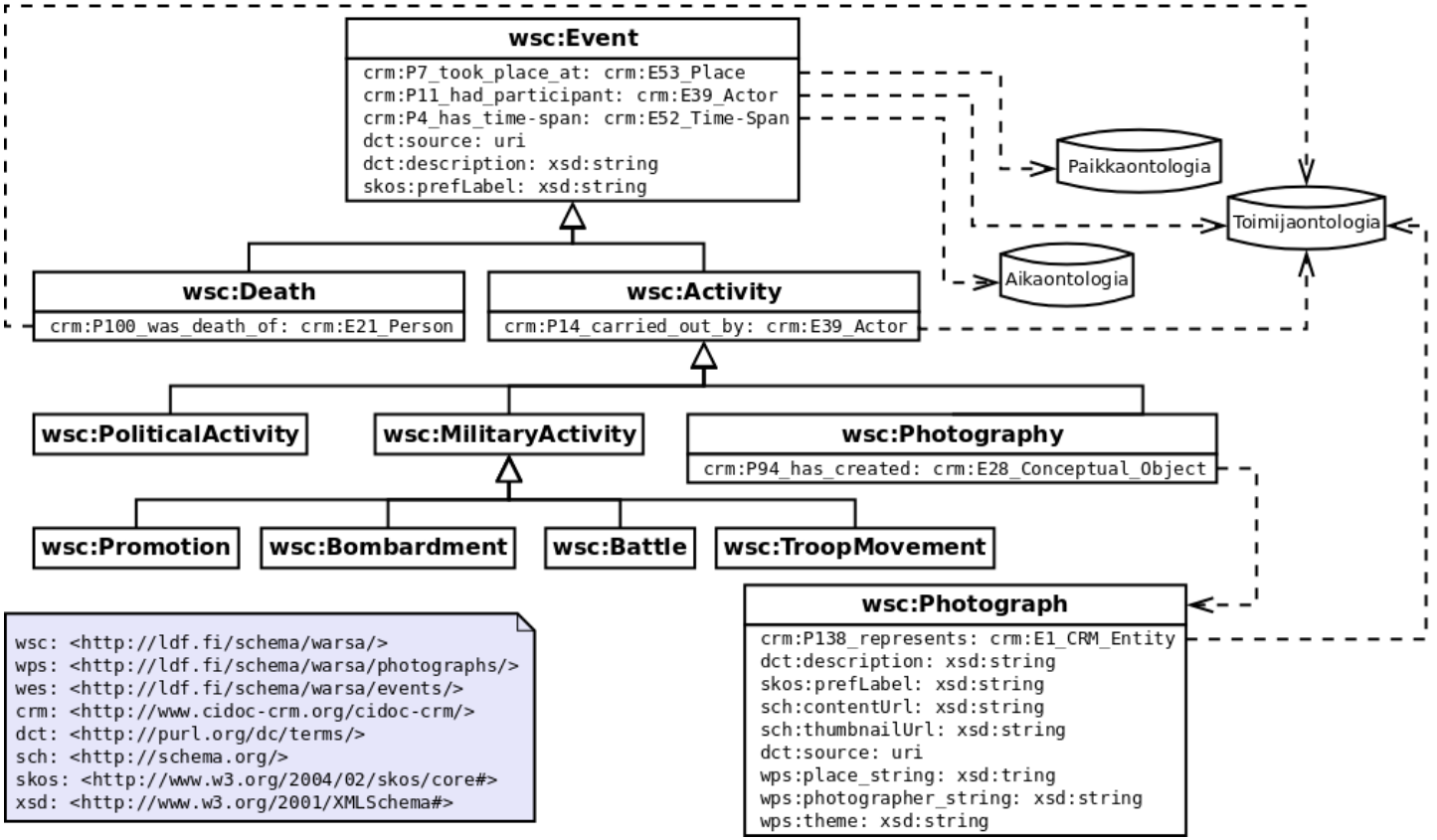
Kuva 7: SA-kuva-arkiston valokuva [Puo].



Kuva 8: Valokuvien metatietojen mallinnus CIDOC CRM -mallia käyttäen.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix sch: <http://schema.org/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wac: <http://ldf.fi/warsa/actors/> .
@prefix wev: <http://ldf.fi/warsa/events/> .
@prefix wph: <http://ldf.fi/warsa/photographs/> .
@prefix wphs: <http://ldf.fi/schema/warsa/photographs/> .
@prefix wsc: <http://ldf.fi/schema/warsa/> .
@prefix wso: <http://ldf.fi/warsa/sources/> .
@prefix wti: <http://ldf.fi/warsa/events/times/> .

wev:sakuva_99632 a wsc:Photography ;
    dct:source wso:source13 ; # SA-kuva
    crm:P11_had_participant wac:person_2587, # Eduard Dietl
        wac:person_52 ; # Hjalmar Siilasvuo
    crm:P14_carried_out_by wac:person_3920 ; # Uuno Laukka
    crm:P7_took_place_at <http://ldf.fi/warsa/places/municipalities/m_place_625> ; # Kiestinki
    crm:P94_has_created wph:sakuva_99632 ;
    dct:description """Erään linnoituskomppanian työsaralta otettuja kuvia: Kenraali Dietl saapuu
        lentokoneella tapaamaan kenraalimajuri Siilasvuota."""@fi ;
    skos:prefLabel """Erään linnoituskomppanian työsaralta otettuja kuvia: Kenraali Dietl saapuu
        lentokoneella tapaamaan kenraalimajuri Siilasvuota."""@fi .

wph:sakuva_99632 a wsc:Photograph ;
    wphs:photographer_string "Uuno Laukka" ;
    wphs:place_string "Kiestingin suunta" ;
    dct:source wso:source13 ; # SA-kuva
    sch:contentUrl <http://static.sotasampo.fi/photographs/r500/sakuva_99632.jpg> ;
    sch:thumbnailUrl <http://static.sotasampo.fi/photographs/r100/sakuva_99632.jpg> ;
    crm:P138_represents wac:person_2587, # Eduard Dietl
        wac:person_52 ; # Hjalmar Siilasvuo
    crm:P3_has_note "Kone Fieseler Storch."@fi ;
    dct:description """Erään linnoituskomppanian työsaralta otettuja kuvia: Kenraali Dietl saapuu
        lentokoneella tapaamaan kenraalimajuri Siilasvuota."""@fi ;
    skos:prefLabel """Erään linnoituskomppanian työsaralta otettuja kuvia: Kenraali Dietl saapuu
        lentokoneella tapaamaan kenraalimajuri Siilasvuota."""@fi .
```

Kuva 9: Mallinnettu valokuva.



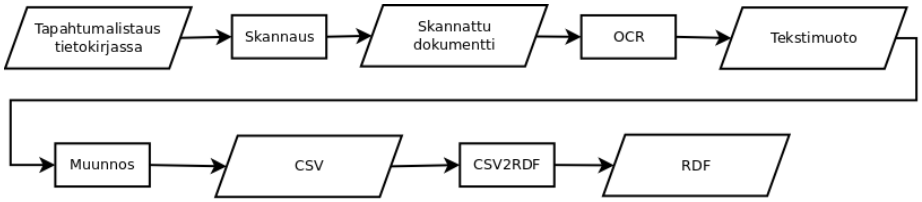
Kuva 10: Valokuvien ja tapahtumien yhdessä muodostama malli.

7 Aineistojen muunnos

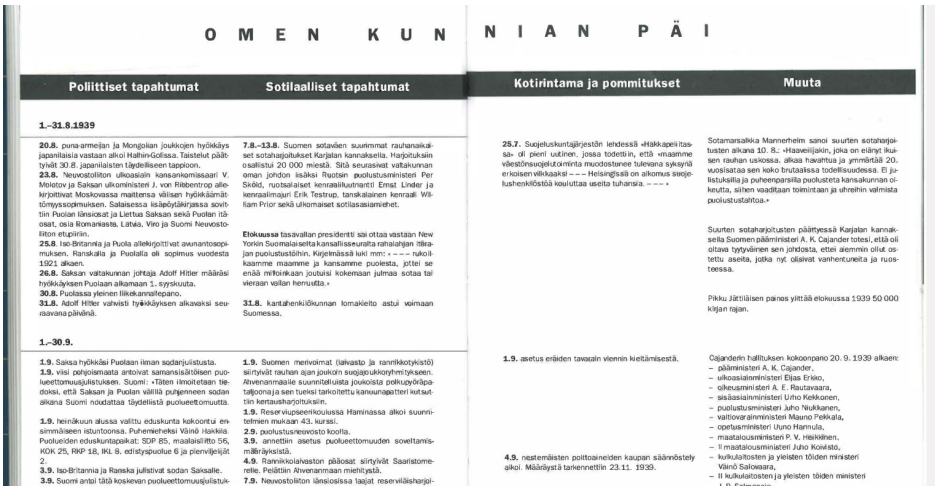
Käytetyt aineistot eivät olleet lähtökohtaisesti linkitettyyn dataan soveltuvassa muodossa, joten ne oli muunnettava RDF-muotoon. Seuraavassa käydään läpi aineistojen muunnosprosessit.

7.1 Tapahtumat

Tapahtumien osalta aineistoja oli kaksi: tietokirjat ja Puolustusvoimien organisaatiokortistosta eristetyt taistelut. Tapahtumat on esitetty järjestelmällisessä muodossa käytetyissä tietokirjoissa [LJ06, LJ05], joten listat päätettiin skannata ja muuntaa tekstintunnistuksen (OCR) avulla sähköiseen muotoon. Kuvassa 12 on yksi skannattu aukeama Talvisodan pikkujättiläisestä [LJ06]. Koko muunnosprosessi on esitetty kuvassa 11.



Kuva 11: Tapahtuma-aineiston muunnos RDF-muotoon.


O M E N K U N N I A N P Ä I

Poliittiset tapahtumat	Sotilaalliset tapahtumat	Kotirintama ja pommitukset	Muuta
1.–31.8.1939			
20.8. puna-armeijan ja Mongolian joukkojen hyökkäys japanilaisia vastaan alkoi Halhin-Golissa. Taistelut päättyivät 30.8. japanilaisten täydelliseen tappioon. **23.8.** Neuvostoliiton ulkoasiain kansankomissaari V. Molotov ja Saksan ulkoministeri J. von Ribbentrop allekirjoittivat Moskovassa maittensa välisen hyökkäämättömyyssopimuksen. Salaisessa lisäpöytäkirjassa sovittiin Puolan länsiosat ja Liettua Saksan sekä Puolan itäosat, osia Romaniasta, Latvia, Viro ja Suomi Neuvostoliiton etupiiriin. **25.8.** Iso-Britannia ja Puola allekirjoittivat avunantosopimuksen. Ranskalla ja Puolalla oli sopimus vuodesta 1921 alkaen. **26.8.** Saksan valtakunnan johtaja Adolf Hitler määräsi hyökkäyksen Puolaan alkamaan 1. syyskuuta. **30.8.** Puolassa yleinen liikekannallepano. **31.8.** Adolf Hitler vahvisti hyökkäyksen alkavaksi seuraavana päivänä.	**7.8.–13.8.** Suomen sotaväen suurimmat rauhanaikaiset sotaharjoitukset Karjalan kannaksella. Harjoituksiin osallistui 20 000 miestä. Sitä seurasivat valtakunnan oman johdon isäksi Ruotsin puolustusministeri Per Sköld, ruotsalaiset kenraaliluutnantit Ernst Linder ja kenraalimajuri Erik Testrup, tanskalainen kenraali William Prior sekä ulkomaiset sotilasasiamiehet. **Elokuussa** tasavallan presidentti sai ottaa vastaan New Yorkin Suomalaiselta kansalaisseuralta rahalahjan liiraajan puolustustöihin. Kirjelmässä luki mm: » – – – rukoilkaamme maamme ja kansamme puolesta, jottei se enää milloinkaan joutuisi kokemaan julmaa sotaa tai vieraan vallan herruutta.« **31.8.** kantahenkilökunnan lomakielto astui voimaan Suomessa.	**25.7.** Suojeluskuntajärjestön lehdessä »Hakkapeliitassa« oli pieni uutinen, jossa todettiin, että »maamme väestönsuojelutoiminta muodostanee tulevana syksynä erikoisen vilkkaaksi – – – Helsingissä on alkomus suojelushenkilöstöä kouluttaa useita tuhansia. – – – « **1.9.** asetus eräiden tavaroiden viennin kieltämisestä. **4.9.** nestemäisten polttoaineiden kaupan säännöstely alkoi. Määräystä tarkennettiin 23.11. 1939.	Sotamarsalkka Mannerheim sanoi suurten sotaharjoitusten alkana 10. 8.: »Haaveilijakin, joka on elänyt ikuisen rauhan uskossa, alkaa havahtua ja ymmärtää 20. vuosisataa sen koko brutaalissa todellisuudessa. Ei julistuksilla ja puheenparsilla puolustusta kansakunnan oikeutta, siihen vaaditaan toimintaan ja uhreihin valmista puolustustahtoa.« Suurten sotaharjoitusten päättyessä Karjalan kannaksella Suomen pääministeri A. K. Cajander totesi, että oli ottava tyytyväinen sen johdosta, ettei alemmin oliut ostettu aseita, jotka nyt olisivat vanhentuneita ja ruosteessa. Pikku Jättiläisen painos ylittää elokuussa 1939 50 000 kirjan rajan. Cajanderin hallituksen kokoonpano 20. 9. 1939 alkaen: – pääministeri A. K. Cajander, – ulkoasiainministeri Eljas Erkko, – oikeusministeri A. E. Rautavaara, – sisäasiainministeri Urho Kekkonen, – puolustusministeri Juho Niukkanen, – valtiovarainministeri Mauno Pekkala, – opetusministeri Uuno Hannula, – maatalousministeri P. V. Heikkinen, – II maatalousministeri Juho Koivisto, – kulkulaitosten ja yleisten töiden ministeri Väinö Salovaara, – II kulkulaitosten ja yleisten töiden ministeri

Kuva 12: Skannattu tapahtumalistaus [LJ06].

Tekstintunnistuksessa käytettiin Adobe Acrobat -ohjelmistoa[15]. Taulukkomuotoisen datan käsittelyyn ja muuntamiseen RDF-muotoon on olemassa

[15] `https://acrobat.adobe.com/fi/fi/acrobat.html`

hyödyllisiä työkaluja. Tästä syystä data muunnettiin ensin puoliautomaattisesti CSV-muotoon. Kuvassa 12 näkyvät sarakkeet muunnettiin ensin yksitellen tekstiksi pdftotext-työkalulla[16], minkä jälkeen teksti muunnettiin CSV-muotoon käyttäen säännöllisiä lausekkeita ja siistimällä lopputulos käsin. Kuvassa 13 on esitetty kuvan 12 vasemman puoleisen sarakkeen tapahtumat CSV-muodossa.

Tapahtumien luokka pääteltiin tietokirjatapahtumien osalta sarakkeesta, jossa tapahtuma oli. Tällä tavoin osa tapahtumista voitiin luokitella sotatoimiksi (*MilitaryActivity*) ja poliittisiksi tapahtumiksi (*PoliticalActivity*). Tämän lisäksi pommitukset luokiteltiin yksinkertaisesti hakemalla tapahtumien kuvauksista merkkijonoa "pommit". Loput tapahtumat luokiteltiin yleisempään tapahtumaluokkaan *Event*. Ylennyksiä, taisteluita ja joukkojen liikutuksia ei pyritty tunnistamaan automaattisesti tietokirjatapahtumista.

CSV-muunnoksen jälkeen data voitiin muuntaa RDF-muotoon csv2rdf-ohjelmalla[17]. Tämä on nimensä mukaisesti ohjelma CSV-muotoisten tiedostojen muuntamiseksi RDF-muotoon. Samalla tapahtumien ajankohdista muodostettiin aikaväliresurssit, joihin tapahtumat viittaavat. Kuvassa 14 on esitetty kuvien 12 ja 13 tapahtumia RDF-muodossa. Muunnoksen jälkeen tapahtumien kuvaustekstien kirjoitusasua korjattiin, kirjoitettiin osin uudestaan ja täydennettiin muiden lähteiden perusteella.

```
PoliticalActivity;1939-08-20;1939-08-30;Puna-armeijan ja Mongolian joukkojen hyökkäys japanilaisia vastaan Halhin-Golissa. Taistelut päättyivät
PoliticalActivity;1939-08-23;1939-08-23;Neuvostoliiton ulkoasiain kansankomissaari V. Molotov Ja Saksan ulkoministeri J. von Ribbentrop allekir
PoliticalActivity;1939-08-25;1939-08-25;Iso-Britannia ja Puola allekirjoittivat avunantosopimuksen. Ranskalla ja Puolalla oli sopimus vuodesta
PoliticalActivity;1939-08-26;1939-08-26;Saksan valtakunnan johtaja Adolf Hitler määräsi hyökkäyksen Puolaan alkamaan 1. syyskuuta.
PoliticalActivity;1939-08-30;1939-08-30;Puolassa yleinen liikekannallepano.
PoliticalActivity;1939-08-31;1939-08-31;Adolf Hitler vahvisti hyökkäyksen alkavaksi seuraavana päivänä.
PoliticalActivity;1939-09-01;1939-09-01;Saksa hyökkäsi Puolaan ilman sodanjulistusta.
PoliticalActivity;1939-09-01;1939-09-01;"Viisi pohjoismaata antoivat samansisältöisen puolueettomuusjulistuksen. Suomi: ""Täten ilmoitetaan tie
PoliticalActivity;1939-09-01;1939-09-01;heinäkuun alussa valittu eduskunta kokoontui ensimmäiseen istuntoonsa. Puhemieheksi Väinö Hakkila. Puol
PoliticalActivity;1939-09-03;1939-09-03;Iso-Britannia ja Ranska julistivat sodan Saksalle.
PoliticalActivity;1939-09-03;1939-09-03;Suomi antoi Iso-Britannian ja Ranskan sodan julistusta Saksalle koskevan puolueettomuusjulistuksen.
```

Kuva 13: Tapahtumalistaus CSV-muodossa.

Taistelutapahtumat eristettiin ja muunnettiin osana Sotasammon toimijaontologian koostamista [Les16], joten tämän työn kannalta niiden osalta muunnosta ei tarvittu.

7.2 Valokuvien metatiedot

Metadataa sodanajan valokuvista oli saatu Puolustusvoimien SA-kuva-arkistolta taulukkomuodossa, ja sitä oli saatavilla myös verkkorajapinnan kautta. Taulukkomuotoisessa metadatassa on enemmän tietoja kuvista verrattuna rajapinnan kautta saatuihin tietoihin, mutta ainoastaan rajapinnan tiedoissa on digitoidun valokuvan verkko-osoite. Kuitenkaan suoraa tapaa yhdistää rajapinnasta saatu data taulukkomuotoiseen dataan ei ollut, sillä eri lähteistä saaduilla tiedoilla ei ollut yhteistä tunnistetta, eivätkä samaan kuvaan liittyvät tiedot olleet aina täysin samoja. Tiedot olivat kuitenkin osittain

[16] Osa Poppler-kirjastoa (`https://poppler.freedesktop.org/`)

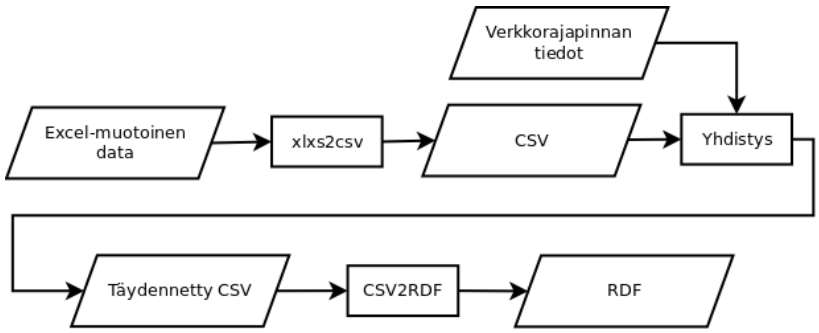
[17] `https://github.com/clarkparsia/csv2rdf`

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wev: <http://ldf.fi/warsa/events/> .
@prefix wsc: <http://ldf.fi/schema/warsa/> .
@prefix wti: <http://ldf.fi/warsa/events/times/> .

wev:event_42 a wsc:PoliticalActivity ;
    crm:P4_has_time-span wti:time_1939-08-20-1939-08-30 ;
    dct:description "Puna-armeijan ja Mongolian joukkojen hyökkäys japanilaisua vastaan Halhin-Golissa. [...]"@fi ;
    skos:prefLabel "Puna-armeijan ja Mongolian joukkojen hyökkäys japanilaisua vastaan Halhin-Golissa. [...]"@fi .

wev:event_43 a wsc:PoliticalActivity ;
    crm:P4_has_time-span wti:time_1939-08-23-1939-08-23 ;
    dct:description "Neuvostoliiton ulkoasiain kansankomissaari V. Molotov ja Saksan ulkoministeri [...]"@fi ;
    skos:prefLabel "Neuvostoliiton ulkoasiain kansankomissaari V. Molotov ja Saksan ulkoministeri [...]"@fi .

wev:event_44 a wsc:PoliticalActivity ;
    crm:P4_has_time-span wti:time_1939-08-26-1939-08-26 ;
    dct:description "Saksan valtakunnan johtaja Adolf Hitler määräsi hyökkäyksen Puolaan alkamaan [...]"@fi ;
    skos:prefLabel "Saksan valtakunnan johtaja Adolf Hitler määräsi hyökkäyksen Puolaan alkamaan [...]"@fi .

wev:event_45 a wsc:PoliticalActivity ;
    crm:P4_has_time-span wti:time_1939-08-30-1939-08-30 ;
    dct:description "Puolassa yleinen liikekannallepano."@fi ;
    skos:prefLabel "Puolassa yleinen liikekannallepano."@fi .
```

Kuva 14: Tapahtumia RDF-muodossa.

samassa järjestyksessä molemmissa lähteissä, joten tiedot saatiin yhdistettyä järjestykseen nojautuen kuvausten vertailun avustuksella.

Kuvassa 15 on esitetty valokuva-aineiston muunnos RDF-muotoon. Taulukkomuotoinen data muunnettiin ensin xlsx2csv-ohjelmalla[18] CSV-muotoon, minkä jälkeen siihen lisättiin verkkorajapinnasta saatu valokuvatiedoston URL-osoite, sekä mahdollinen lisätietoa kuvasta antava kommentti. Metatietotaulukoissa ei ollut kuitenkaan saatavilla kaikkia rajapinnasta löytyviä kuvia, joten rajapinnan tiedoista muodostettiin myös kuvat sikäli kuin niille ei löytynyt vastinetta taulukoista. Valitettavasti tällaisten kuvien osalta ei ollut saatavissa eksplisiittistä tietoa kuvaajasta ja paikasta, ainoastaan kuvateksti ja mahdollinen kommentti.



Kuva 15: Valokuvien metatietojen muunnos RDF-muotoon.

Tämän jälkeen CSV-muotoinen data muunnettiin RDF-muotoon csv2rdf-ohjelmalla. Tässä vaiheessa data pidettiin vielä yksinkertaisessa muodossa,

[18] `https://github.com/dilshod/xlsx2csv`

joka ei ollut CIDOC CRM:n mukainen, jotta aineisto olisi helpommin käsiteltävissä linkityksessä. Aineisto sai lopullisen muotonsa vasta seuraavassa luvussa (8) kuvatun linkityksen yhteydessä. Kun linkitys oli tehty, aineisto muutettiin suunnitellun mallin mukaiseksi säännöllisillä lausekkeilla ja SPARQL-kyselyillä. Tässä yhteydessä valokuvien päivämääristä luotiin mallin mukaiset aikaväliresurssit.

8 Aineistojen linkitys

Työssä käsiteltyjen tapahtuma- ja valokuva-aineistojen kuvausteksteissä ja tiedoissa esiintyy muun muassa henkilöitä, joukko-osastoja ja paikkoja. Kullekin entiteettityypille luotiin oma tunnistusheuristiikka käyttäen apuna SeCo-ryhmän Lexical Analysis Services -palvelua[19] sekä ARPA-järjestelmää [Mä14].

Haasteena työssä käytettyjen aineistojen linkittämisessä muihin aineistoihin on, että linkitettävät entiteetit on suurimmalta osin tunnistettava luonnollisen kielen kuvausteksteistä. Tästä seuraa myös peruste linkkien luomisen hyödyllisyydelle: esimerkiksi johonkin henkilöön liittyviä tapahtumia tai valokuvia olisi vaikea löytää ilman yksitulkintaista yhteyttä. Haasteen ratkaisemiseksi on käytettävä luonnollisen kielen käsittelyn tekniikoita kuten tekstin perusmuotoistamista ja nimettyjen entiteettien linkitystä.

Aineistoissa esiintyviä entiteettejä olivat erityisesti paikat, henkilöt ja joukko-osastot. Tavoitteena oli tunnistaa eri entiteetit ja linkittää nämä sopiviin ontologioihin. Sotasampoon kuuluu toisen maailmansodan aikaisten paikkojen ontologia ja toimijaontologia, joka sisältää talvi- ja jatkosotaan liittyviä joukko-osastoja ja henkilöitä.

8.1 Joukko-osastot

Tapahtumien ja valokuvien kuvauksista haettiin joukko-osastoja käyttäen hyväksi SeCo:ssa toteutettua sotatoimijaontologiaa [HHL+16]. Ontologia sisältää noin 16 000 suomalaista joukko-osastoa toisen maailmansodan ajalta.

Joukko-osastojen tunnistamisessa tekstistä haasteena on erilaiset viittaustavat: osastoon voidaan viitata useammanlaisella lyhenteellä, virallisella nimellä tai lempinimellä. Toimijaontologia kuitenkin sisältää joukko-osastojen nimien lisäksi näiden lyhenteet ja osastojen lempinimiä. Erityisesti lyhenteiden tapauksessa kirjoitusasu saattaa kuitenkin poiketa tekstistä löytyvästä. Ontologiasta löytyvät lyhenteet voivat sisältää pisteitä ja välilyöntejä, mutta tekstissä näitä ei ole käytetty välttämättä samalla tavalla. Esimerkiksi ensimmäisen jalkaväkirykmentin lyhenne “JR 1” voi olla kirjoitettu yhteen (JR1) tai siinä voi olla käytetty roomalaista numeroa (JR I).

Koska joukko-osastotyyppejä ei ole erityisen monta, voitiin lyhenteet normalisoida tekstin esikäsittelyvaiheessa tyyppikohtaisesti. Esimerkiksi “JR1”

[19] `http://demo.seco.tkk.fi/las/`

ja “JR. I” muutettiin esikäsittelyssä joukko-osasto-ontologiassa käytettyyn muotoon “JR 1” käyttäen säännöllisiä lausekkeita. Samankaltainen muunnos tehtiin prikaateille, divisioonille, pataljoonille ja armeijakunnille.

Linkitys toteutettiin muuntamalla teksti perusmuotoon ARPA-järjestelmän avulla, ja vertaamalla tekstistä muodostettuja n-grammeja toimijaontologian joukko-osastojen nimiin SPARQL-kyselyn avulla. Joukko-osastoista käytetyissä nimissä on sanaliittoja, joiden kaikki sanat eivät ole perusmuodossa, kuten “Kannaksen armeija” ja “Haminan ryhmä”. ARPA-järjestelmän ominaisuus perusmuotoistaa vain sanaliittojen viimeinen sana on hyödyllinen tällaisten nimien tunnistamisessa.

Nimiä verrattiin jättämällä pisteet, pilkut, kauttaviivat, välilyönnit ja isot alkukirjaimet huomiotta. Tällä pyrittiin parantamaan saantia tapauksissa, joissa välimerkkejä on käytetty eri tavalla kuin ontologian nimissä, mutta joita ei oltu huomioitu esikäsittelyssä. Vertailussa on mahdollista syntyä vääriä osumia esimerkiksi lauserajojen ylittyessä. Tällaisia vääriä osumia pidettiin kuitenkin epätodennäköisinä, sillä joukko-osastojen nimiä ei yleisesti ottaen ole helppo sekoittaa muihin sanoihin.

Joukko-osastojen nimet ovat hyvin yksilöiviä, joten homonymia ei ole niiden osalta suuri ongelma. Talvi- ja jatkosodan joukko-osastot ovat kuitenkin erillisiä resursseja toimijaontologiassa, joten samannimisistä osastoista oli valittava valokuvan tai tapahtuman ajankohdan osasto. Koska kaikkien tapahtumien päivämäärät olivat tiedossa, oli ne helppo yhdistää joukko-osastojen aikakausiin. Kaikkien valokuvien päivämääriä sen sijaan ei ollut tiedossa. Alkuperäiset metadatatiedostot oli kuitenkin nimetty sodan mukaan, joten päivämäärättömille valokuville voitiin tämän perusteella määrittää aikakausi. Tämän luokittelun perusteella voitiin kandidaattijoukko-osastoista valita oikean aikakauden osasto.

Joissain tilanteissa yksi mainin voi tuottaa useamman kandidaatin myös siksi, koska ylemmän joukko-osaston nimi sisältyy pienemmän joukko-osaston nimeen: esimerkiksi merkkijonoon “II/JR 8” sisältyy paitsi joukko-osasto II/JR 8 (jalkaväkirykmentti 8, toinen pataljoona) myös sen yläosasto JR 8 (jalkaväkirykmentti 8). Tarkkuuden vuoksi tällaisissa tapauksissa valittiin tarkempi eli pidempään mainintaan kohdistuva kandidaatti. Toisaalta jos tarkkaa joukko-osastoa ei löytynyt ontologiasta, ei linkittymistä ylempään joukko-osastoon pyritty estämään, sillä tällainen linkki voidaan tulkita ainakin osittain oikeaksi.

Valokuvien kuvauksissa on käytetty joukko-osaston nimien lisäksi nelinumeroisia peitelukuja eli kenttäpostinumeroita. Nämä luvut olivat saatavilla toimijaontologiassa, mikä mahdollisti niiden käyttämisen linkityksessä. Haasteena peitelukujen käyttämisessä oli, että ne voivat sekoittua muihin lukuihin kuten painoihin tai vuosiin. Peiteluvut, jotka sekoittuvat sellaisiin vuosiin, joita tekstissä saatettiin mainita, suljettiin pois linkityksestä. Valokuvat on otettu toisen maailmansodan aikana, joten nämä vuodet usein esiintyivät tekstissä. Tämän lisäksi tekstissä mainitaan esimerkiksi rakennusten raken-

tamisvuosia. Poissuljettaviksi vuosiksi päätettiin valita vuodet 1800–1945. Muut luvut pyrittiin sulkemaan pois tarkastelemalla lukua ympäröivää tekstiä säännöllisten lausekkeiden avulla. Jos lukua seurasi määre kuten “kpl” tai “m” tai sitä edelsi sana “noin”, hylättiin lukuun täsmännyt joukko-osasto.

8.2 Henkilöt

Tapahtuma- ja valokuva-aineistojen kuvauksissa mainitaan usein henkilöitä, ja nämä pyrittiin yksilöimään sotatoimijaontologian avulla. Sotasampo tarjoaa toimijaontologian, joka sisältää noin 100 000 toiseen maailmansotaan liittyvää henkilöä, sisältäen muun muassa Kansallisarkiston tietokannan sodassa kaatuneista suomalaisista sotilastoimihenkilöistä [Les16]. Ontologia sisältää kattavaa tietoa sisältämistään henkilöistä, kuten syntymä- ja kuolinajat sekä -paikat, joukko-osastot, joissa henkilöt ovat palvelleet, ja sotilasarvot ylennyspäivineen. Joistain henkilöistä on saatavilla enemmän tietoa kuin toisista riippuen mistä lähteistä henkilön tiedot ovat peräisin.

Kuvaustekstien lisäksi valokuvametatiedoissa on erillisessä tietueessaan ilmoitettu valokuvan ottajan nimi. Näistä on erikseen muodostettu henkilöt toimijaontologiaan. Koska yhteys metatiedoissa mainituista nimistä ontologian henkilöihin oli tiedossa, voitiin valokuvaajat linkittää yksinkertaisesti kuvaajakentän merkkijonon perusteella. Sen sijaan kuvausteksteissä esiintyvien henkilöiden tunnistaminen ja linkittäminen oli monimutkaisempaa.

Kuvauksissa esiintyy henkilöiden nimiä eri muodoissa. Esimerkissä 1 on listattu tapoja viitata kenraaliluutnantti Karl Lennart Oeschiin SA-kuva-aineiston kuvateksteissä.

Esimerkki 1. Tapoja viitata Karl Lennart Oeschiin

1. kenraali Oesch
2. kenraaliluutnantti Oesch
3. kenraaliluutnantti K.L. Oesch
4. kenr.luutn. Oesch
5. kenraali Svensson, Mäkinen ja Oesch
6. kenraalit Walden, Oesch, Lundqvist ja Talvela

Jos Karl Lennart Oeschin voisi yksilöidä sukunimen perusteella, olisi esimerkin 1 lausekkeista helppo tunnistaa yksinkertaisesti vertaamalla tekstin sanoja toimijaontologian henkilöiden sukunimiin. Henkilöitä ei kuitenkaan voi yleisesti ottaen tunnistaa pelkän sukunimen perusteella, sillä

1. paikannimet sekoittuvat helposti sukunimiin (esimerkiksi Haapajärvi on kunta ja sukunimi) ja

2. monella henkilöllä on sama sukunimi.

Toisaalta koko nimen ja sotilasarvon vaatiminen laskisi saantia liikaa. Esimerkiksi “kenraaliluutnantti Oesch” ja “kenraaliluutnantti K.L. Oesch” ovat talvi- ja jatkosodan kontekstissa yhtä yksilöiviä käsitteitä [suo83].

Tunnistettavista henkilöistä suurin osa on sotatoimihenkilöitä, ja useimmissa tapauksissa käytettyjen aineistojen kontekstissa henkilöihin viitatessa mainitaan sotilasarvo, mikä auttaa tunnistamisessa. Täysin suoraviivaista tämäkään ei tunnistamisesta tee, sillä monet henkilöt ovat saaneet ylennyksiä sotien aikana. Upseerien ylennyshistoria auttaa upseerien tarkemmassa tunnistuksessa. Esimerkiksi tieto siitä, että Karl Lennart Oesch on ollut talvi- ja jatkosodan aikaan kenraaliluutnantti [Man], auttaa erottamaan hänet veljestään ja pojastaan, jotka myös mainitaan SA-kuvametatiedoissa.

Henkilöiden tunnistuksessa tavoitteena oli tunnistaa mahdollisimman paljon henkilöitä minimoiden kuitenkin virheelliset tunnistukset. Päädyttiin siis heuristiikkaan, joka sulkee pois liian epämääräiset viittaukset, tuottaa mahdollisimman paljon oikeita tunnistuksia mutta ei täysin takaa tunnistuksen oikeellisuutta. Koska useat seikat vaikuttavat henkilön yksilöintiin, kandidaatit pisteytettiin, ja parhaat pisteet saaneet kandidaatit linkitettiin. Ideaalitapauksessa kutakin mainintaa kohden löytyi kandidaatti, joka sai enemmän pisteitä kuin muut samasta maininnasta syntyneet kandidaatit. Tapauksessa, jossa useampi saman maininnan kandidaatti sai saman määrän pisteitä, valittiin kaikki tällaiset kandidaatit linkitettäviksi.

Ehdot kandidaattien valitsemiseksi ja hylkäämiseksi eivät ole ilmiselviä. Esimerkiksi voisi ajatella, että kaikki ennen tapahtuman ajankohtaa tai kuvanottamispäivää kuolleet henkilöt voisi sulkea pois. Hautajaisten takia kyseinen ehto ei kuitenkaan ole pätevä, ja valokuva-aineistossa on kuvia tunnettujen henkilöiden hautajaisista. Tästä syystä päädyttiin ratkaisuun, jossa yli kuukauden kuolleena olleiden henkilöiden tunnistusten pistemäärää vähennettiin. Kaikki henkilöiden yksilöinnissä huomioon otetut seikat on esitetty taulukossa 4. Syyt näiden valintaan ja tarkempi käsittely esitellään seuraavaksi.

Taulukossa 5 on esitetty, kuinka hyvin nimi yksilöi henkilön henkilöontologian sisällä. Henkilöontologian 99 483 henkilöstä vain 10 185 henkilöllä on ontologian sisällä uniikki sukunimi. Jos sukunimen lisäksi verrataan etunimiä, ja suljetaan pois sellaiset henkilöt, joilla on ainakin yksi sama etunimi, nousee uniikkien henkilöiden lukumäärä noin puoleen ontologian henkilöistä. Paras yksilöivyys saadaan vertaamalla koko nimeä sellaisenaan. Käytännössä kaikkia etunimiä käytetään kuitenkin harvoin henkilöön viitatessa, minkä takia koko nimen vaatiminen laskisi saantia suuresti linkityksen yhteydessä.

Sukunimen riittämättömyydestä henkilön yksilöimisessä kertoo se, että sukunimi “Hägglund” esiintyy henkilöontologiassa yhdeksän kertaa. Pelkän sukunimen perusteella linkittämällä valokuva-aineistosta löytyykin lähes kaksi miljoonaa linkkiä, kun aineiston sanat perusmuotoistetaan. Tämä

Tieto	**Käsittely**
Sukunimi	Vaaditaan
Etunimet	Etunimi vai etunimen etukirjain? Ensimmäinen etunimi?
Kuolinaika	Onko kandidaatti ollut elossa?
Sotilasarvot	Kuinka korkea-arvoinen kandidaatti on? Onko henkilöllä ollut mainittu sotilasarvo kyseisenä ajankohtana? Löytyykö kontekstista ristiriitainen sotilasarvo?
Maininnan pituus	Sisältyykö maininta jonkin toisen kandidaatin mainintaan?
Henkilön tunnettuus	Mistä lähteestä henkilön tiedot ovat?
Mannerheim-ristin ritarit	Onko kandidaatti Mannerheim-ristin ritari, ja onko kontekstissa maininta ritariudesta?
Joukko-osastot	Liittyykö linkitettävään resurssiin sama joukko-osasto kuin jossa kandidaatti on palvellut?

Taulukko 4: Henkilöiden yksilöinnissä huomioon otetut tiedot.

Yksilöinti	**Kpl**	**%**
Sukunimi	10 185	10.2
Sukunimi ja mikä tahansa etunimi	50 553	50.8
Koko nimi	92 098	92.6

Taulukko 5: Nimen yksilöivyys henkilöontologian sisällä.

tarkoittaisi, että kussakin valokuvan kuvauksessa olisi keskimäärin mainittu yli sata henkilöä. Sukunimen lisäksi vaadittiin siis vähintään yksi etunimi tai sotilasarvo. Kaikki sotilasarvot eivät kuitenkaan yksilöi henkilöitä yhtä hyvin: korkea-arvoisia henkilöitä on vähemmän kuin miehistöön kuuluneita, joten korkeat arvot, kuten kenraali, yksilöivät paremmin kuin esimerkiksi sotamies. Esimerkiksi sotamies Hägglundeja on ontologiassa neljä kappaletta mutta kenraaliluutnantti Hägglundeja vain yksi. Henkilöontologian henkilöistä onkin miehistöön kuuluneita 71%.

Alempien sotilasarvojen osalta henkilöontologia sisältää lähinnä sodissa menehtyneet henkilöt, mikä vaikeuttaa alempiarvoisten henkilöiden tunnistusta. Vaikka tekstissä mainittaisiinkin joku henkilö hyvin yksilöivässä muodossa, saattaa tämä jäädä tunnistamatta. Toisaalta henkilö saattaa vaikuttaa hyvältä kandidaatilta olematta kuitenkaan tekstissä mainittu henkilö, koska mainittua henkilöä ei löydy ontologiasta. Joissain tapauksissa saattaa henkilötietojen puutteesta ja osittaisesta osumasta johtuen muodostua myös ilmiselvästi väärä yksilöinti. Esimerkiksi "sotamies Heino Jauhiainen" yhdistyy sotamiehiin, joiden sukunimi on "Heino", sillä toinen kahden sanan mittainen n-grammi merkkijonosta on "sotamies Heino". Jos sotamies Heino Jauhiaista ei löydy henkilöontologiasta, "sotamies Heino" on pisin löydetty tunniste, joten kaikki sotamiehet, joiden nimi on Heino, ovat sopivia. Tällaisten virheellisten yksilöintien poissulkeminen ei ole yksinkertaista. Yllä mainitussa esimerkissä voitaisiin tutkia seuraavaa sanaa ("Jauhiainen"), todeta sen olevan erisnimi ison alkukirjaimen takia ja näin sulkea pois kyseinen yksilöinti. Tämä ei kuitenkaan toimisi yleisessä tapauksessa, sillä samalla suljettaisiin mahdollisesti pois valideja tapauksia, kuten "sotamies Heino Helsingissä". Edellä mainituista syistä miehistöön kuuluvia henkilöitä pidettiin tunnistuksessa vähemmän todennäköisinä osumina kandidaatteja arvioitaessa.

On myös tapauksia, joissa henkilön sotilasarvo sekä etu- ja sukunimi mainitaan tekstissä, mutta toimijaontologiasta löytyy samanniminen henkilö, jolla kyseistä sotilasarvoa ei kuitenkaan ole ollut ainakaan kyseisenä ajankohtana. Tällaisia tapauksia silmällä pitäen tunnistukseen johtaneen n-grammin ympäristö otettiin huomioon: tunnistuksen kandidaattia pidettiin vähemmän todennäköisenä yksilöintinä, jos n-grammia edelsi sotilasarvo, jota tunnistetulla henkilöllä ei toimijaontologian mukaan ole ollut. Sotasurmatietokannan osalta henkilöiden sotilasarvoista on tiedossa ainoastaan kuolinhetken sotilasarvo. Tästä syystä aineistoissa saattaa esiintyä henkilö, joka löytyy sotasurmakannasta, mutta jonka sotilasarvona mainitaan jokin muu kuin kuolinhetken sotilasarvo. Epäjohdonmukainen sotilasarvo maininnan yhteydessä ei kuitenkaan estä henkilön linkittymistä, jos muita yksilöimiseen johtavia todisteita on riittävästi. Suoraan n-grammia edeltävän sanan tutkiminen ei toimi joka tilanteessa. Otetaan esimerkiksi lauseke "luutnantti M. M. Kytölä". Lausekkeen n-grammilla "M. Kytölä" voidaan löytää kandidaatteja, mutta tässä tilanteessa n-grammia ei edellä sotilasarvo vaan

nimen etukirjain. Tämän takia sikäli kuin n-grammia edeltävässä sanassa oli iso alkukirjain, tätä edeltävä sana tarkastettiin myös.

Sotilasarvojen kirjoitusasut aiheuttivat myös haasteita, sillä tekstissä käytetään usein lyhenteitä kuten “kenr.luutn.” viitatessa arvoon “kenraaliluutnantti”. Tämän vuoksi kehitettiin sääntöjä, joiden avulla lyhenteet laajennettiin täyspitkiksi toimijaontologiaan vertaamista varten. Esimerkiksi “kenr.luutn. Oesch” voidaan muuntaa tunnistettavaan muotoon “kenraaliluutnantti Oesch” korvaamalla seuraavasti:

Esimerkki 2. “kenr.luutn.” ⟶“kenraali#luutn.” ⟶“kenraaliluutnantti”.

Sana “kenraali” voi viitata paitsi kenraalin arvoon myös mihin tahansa kenraalikuntaan kuuluvaan sotilasarvoon. Tästä syystä esimerkiksi Klaus Lennart Oeschia ei voi tunnistaa viittauksesta “kenraali Oesch” vain sotilasarvoja vertaamalla, sillä hänen korkein arvonsa oli kenraaliluutnantti [suo83]. Toimijaontologia sisältää sotilasarvohierarkian, jossa sotilasarvot on luokiteltu ryhmiin kuten *kenraalikunta*. Tästä syystä sana “kenraali” korvattiin sanalla “kenraalikunta”, ja henkilön sotilasarvoryhmää hyödynnettiin yksilöinnissä tarkan sotilasarvon lisäksi.

Kaikki aineistoissa esiintyvät henkilöviittaukset eivät vastaa tunnistuksessa vaadittua muotoa, kuten esimerkiksi “kenraalit Svensson, Mäkinen ja Oesch”. Kyseisellä tavalla esitetyt listat pyrittiin tunnistamaan automaattisesti ja muuntamaan tunnistukseen soveltuvaan muotoon, jossa listan jokaista nimeä edeltää sotilasarvo. Esimerkiksi muuntamalla yllä mainittu lauseke muotoon “kenraali Walden, kenraali Oesch, kenraali Lundqvist ja kenraali Talvela” voidaan listatut henkilöt tunnistaa sotilasarvon ja sukunimen perusteella. Muunnos toteutettiin käyttämällä säännöllisiä lausekkeita.

Edellä mainittujen kaltaisten listojen avaamisesta esitetyllä tavalla voi muodostua vääriäkin sotilasarvo-nimi-pareja. Esimerkiksi lausekkeessa “kenraali Oesch, Ryti ja Marski” sana “Ryti” viittaa todennäköisesti Suomen presidentti Risto Rytiin, eikä kenraaliin nimeltä Ryti. Tällaisia tapauksia ei kuitenkaan käytetyissä aineistoissa juuri esiinny.

Jotkin tapaukset voidaan käsitellä erityistapauksina. Esimerkiksi Carl Gustaf Emil Mannerheimiin viitataan usein pelkällä sukunimellä ja tässä tapauksessa sukunimi riittää yksilöimään henkilön. Lisäksi Mannerheimiin viitataan myös lempinimellä “Marski”. Muihinkin henkilöihin — esimerkiksi Väinö Tanner — viitataan pelkällä sukunimellä erityisesti tapahtumien kuvauksissa. Erikoistapaukset määritettiin osittain erikseen tapahtumien ja valokuvien osalta, sillä jotkin henkilöt, jotka ovat tapahtumien kuvauksissa yksilöitävissä esimerkiksi pelkän sukunimen perusteella — kuten Neuvostoliiton ulkoasiain kansankomissaari Vjatšeslav Molotov — eivät olekaan valokuvien kuvauksissa. Valokuvien kuvauksissa sanalla “Molotov” viitataan myös Molotov-nimiseen kylään.

Koska ARPA-järjestelmä palauttaa osumien lisäksi myös osumat tuottaneet n-grammit, voidaan näitä käyttää täsmentämiseksi entisestään: jos

yhdelle n-grammille on useampi osuma ja se on myös osa pidempää osuman tuottavaa n-grammia, todennäköisimmin osumista pisimpään n-grammiin kohdistuva osuma on oikea. Esimerkiksi tekstistä “kenraali K. Oesch” muodostuu kaksi vähintään kahden sanan mittaista n-grammia: “K. Oesch” ja “kenraali K. Oesch”. Näihin n-grammeihin kohdistuvista osumista on perusteltua valita henkilö, joka täsmää molempiin n-grammeihin sellaisten henkilöiden sijaan, jotka täsmäävät vain lyhyempään n-grammiin “K. Oesch”.

Kandidaattien tunnettuus vaikutti myös pisteytykseen: jos kandidaatin tietojen lähteenä oli Wikipedia, sai se pienen lisäyksen pisteytykseen. Samaten Mannerheim-ristin ritarit saivat saman pistelisän. Nämä henkilöt ovat muista lähteistä koottuja henkilöitä tunnetumpia, ja maininnoissa, jotka eivät muuten ole erityisen yksiselitteisiä, voidaan olettaa puhuttavan tunnetusta henkilöstä, jos muuta tapaa yksilöidä henkilöä ei löydy. Mainintaa ympäröivästä tekstistä myös viittausta Mannerheim-ristiin, jonka löytyessä kandidaatit, jotka olivat Mannerheim-ristin ritareita, saivat korkeammat pisteet.

Myös edeltäviä linkityksiä voidaan käyttää apuna yksilöinnissä. Jos esimerkiksi valokuvan kuvauksessa mainitaan jokin joukko-osasto ja henkilö, on kyseinen henkilö mahdollisesti palvellut mainitussa joukko-osastossa. Yksilöinnissä voidaan siis painottaa sellaisia henkilöitä, jotka ovat palvelleet jossain valokuvaan linkittyneessä joukko-osastossa.

8.3 Paikat

Paikannimien yksilöinnissä on henkilöiden tapaan ongelmana monitulkintaisuus: paikannimet sekoittuvat henkilöiden nimiin — esimerkiksi paikka nimeltä Kekkonen — ja monet paikat ovat samannimisiä — esimerkiksi Mustalampi-nimisiä paikkoja löytyy Suomesta satoja. Historiallisten paikkojen osalta monimerkityksisyyden lisäksi haasteena on paikkojen muuttuminen ajan kuluessa: toisen maailmansodan aikainen Helsinki ei ole alueellisesti sama kuin nykyinen. Toisaalta osaa silloisista kunnista ei ole enää lainkaan olemassa. Suomen rajat ovat myös muuttuneet viime sotien seurauksena, ja moni sotatapahtuma ja valokuva sijoittuukin alueille, jotka Suomi joutui luovuttamaan Neuvostoliitolle talvi- ja jatkosodan rauhanehtojen takia.

Talvi- ja jatkosodan aikaisten tapahtumien ja valokuvien paikkojen yksilöimiseen ja linkittämiseen tarvitaankin siis tietokanta historiallisista paikoista. Tätä varten käytettävissä oli ontologia historiallisista luovutetun Karjalan paikoista ja Suomen kunnista sotien aikana [HIT16]. Mahdollisimman suuren kattavuuden aikaansaamiseksi linkityksessä käytettiin apuna myös nykyisten paikkojen ontologiaa, Maanmittauslaitoksen Paikannimirekisteriä[20]. Tapahtumien kuvauksissa esiintyviä paikannimiä linkitettiin siis kolmeen eri paikkaontologiaan: historialliset kunnat, Karjalan paikat ja

[20] `http://www.ldf.fi/dataset/pnr/index.html`

Paikannimirekisteri.

Toisin kuin henkilöiden linkityksessä kullekin paikkaviittaukselle hyväksyttiin korkeintaan yksi linkitys siinäkin tapauksessa, että valinta jouduttiin tekemään mielivaltaisesti. Tämä tehtiin siitä syystä, että useamman samannimisen paikan näyttäminen samalle tapahtumalle aiheuttaisi sekaannusta tapahtumien visualisoinnissa. Lisäksi Karjalan paikat -ontologia sisältää useita samaa paikkaa kuvaavia resursseja. Tämä johtuu siitä, että ontologian paikat on poimittu karttalehdiltä [Ikk16], ja samoja alueita voi esiintyä useammalla karttalehdellä. Erityisesti suuret vesialueet levittäytyvät monelle karttalehdelle, ja ontologiasta löytyy paljon samannimisiä vesimuodostumia, joista moni kuitenkin viittaa itse asiassa samaan paikkaan. Tästä syystä kirjoittamishetkellä esimerkiksi Laatokalle on ontologiassa 22 URI-tunnistetta.

Taulukossa 6 on esitetty historiallisten kuntien ja Karjalan paikannimien määrät sekä paikannimien yksilöivyydet. Uniikkien paikannimien osuudet on laskettu paikan tyypin sisällä ja koko ontologian osalta.

Tyyppi	**Kpl**	**Yksilöivyys tyypin sisällä**	**Yksilöivyys ontologiassa**
Kunta	625	99%	76%
Kirkonkylä, kaupunki	50	100%	54%
Kylä	1544	88%	59%
Maastokohde	10 864	71%	66%
Rakennettu kohde	14 363	45%	40%
Symboli	29	7%	0%
Vesimuodostuma	5553	66%	63%

Taulukko 6: Paikkojen ja uniikkien paikannimien lukumäärät tyypeittäin historiallisten kuntien ja Karjalan paikkojen ontologioissa.

Koska samannimisiä paikkoja löytyy paitsi eri paikkaontologioista myös ontologioiden sisällä, usean osuman tapauksissa paikkoja painotettiin niiden tyypin mukaan. Historiallisten paikkojen ontologiassa kunta, “kirkonkylä, kaupunki” ja kylä ovat oman tyyppinsä sisällä melko yksilöiviä. Jotta koko Suomi voitiin kattaa tarkemmin kuin vain kuntien osalta, otettiin huomioon nyky-Suomen Paikannimirekisterin kaupungit ja kylät — tarkemmin sanottuna paikat, joiden tyyppi oli “kunta, kaupunki” tai “kylä, kaupunginosa tai kulmakunta”. Vastaavanlaista tietyn tyyppisiä paikkoja suosivaa heuristiikkaa ovat käyttäneet esimerkiksi Glover ja kumppanit [GTB+10], jotka toteuttivat paikkayksilöijän paikkojen yksilöintiin Ison-Britannian ja Irlannin historiallisista aineistoista.

Usean osuman tapauksessa osumia painotettiin seuraavalla tavalla paikan tyypin perusteella:

1. sotien aikainen kunta (historialliset kunnat)
2. kirkonkylä, kaupunki (Karjalan paikat)
3. kylä (Karjalan paikat)
4. vesimuodostuma (Karjalan paikat)
5. maastokohde (Karjalan paikat)
6. nykyinen kunta, kaupunki (Paikannimirekisteri)
7. nykyinen kylä, kaupunginosa tai kulmakunta (Paikannimirekisteri)

Suurimmassa osassa valokuvien metadataa paikkatieto oli ilmoitettu omana kenttänään, jos paikka oli tiedossa, minkä takia näiden valokuvien linkittämisessä paikkoihin ei paikannimien sekoittuminen henkilöiden nimiin ollut ongelma. Tapahtumien paikat oli yksilöitävä kuvaustekstistä, jolloin paikannimien ja henkilöiden nimien erottaminen oli haasteena. Virheellisten tunnistusten välttämiseksi joukko paikannimiä suljettiin pois linkityksessä. Poissuljettujen lista koostui paikannimistä, jotka löytyvät linkitettävistä ontologioista mutta johtavat käytännössä aina väärään tunnistukseen. Tällaisia paikannimiä ovat esimerkiksi Sillanpää ja Kekkonen, jotka viittaavat todennäköisemmin henkilöihin kuin paikkoihin.

Linkityksessä käytetty ARPA-palvelu mahdollistaa myös n-grammien valikoimisen niiden sanojen tunnistetun merkityksen perusteella. Sanoja voidaan suodattaa sen mukaan, minkä tulkinnan jäsentäjä valitsee niille. Esimerkiksi voidaan hyväksyä ainoastaan substantiivit tai sulkea pois suku- ja etunimet. Paikkojen tapauksessa voitaisiin joko valita sanoista ainoastaan paikannimiksi tulkitut sanat tai henkilöiden nimiin sekoittumisen ehkäisemiseksi sulkea pois etu- ja sukunimiksi tulkitut sanat. Käytännössä palvelun käyttämä jäsentäjä ei kuitenkaan osaa tässä käyttötapauksessa erottaa erityyppisiä erisnimiä toisistaan vaan antaa paikan- ja henkilönnimitulkinnoille saman painoarvon, silloin kun erottelusta olisi hyötyä. Tästä johtuen suodattimien käytöstä ei ole tämän työn kannalta mitään hyötyä. Itse asiassa etu- ja sukunimien poissuodattaminen ei vaikuta tuloksiin lainkaan.

Suodatus voidaan tehdä myös vahvemmin siten, että sana suljetaan pois, jos jokin mahdollinen tulkinta — ei siis välttämättä jäsentäjän lopullinen valinta — antaa sanalle jonkin paikannimestä poikkeavan merkityksen. Tällaisen suodattimen käyttö vähentää kuitenkin saantia, sillä esimerkiksi yksi jäsentäjän sanalle “Viipuri” antama tulkinta on sukunimi, vaikka se käytetyissä aineistoissa viittaa aina paikkaan.

8.4 Linkitysten toteutus

Aineistojen linkityksessä käytettiin SeCon ARPA-palvelua. ARPA [Mä14] on automaattinen annotointijärjestelmä, joka käyttää kieliteknologisia työkaluja syötteen normalisointiin ja SPARQL-palvelupistettä tekstissä esiintyvien käsitteiden hakuun. ARPA käsittelee — yleisimmässä tapauksessa perusmuotoistaa — syötetekstin, pilkkoo sen n-grammeiksi, hakee näille vastineita SPARQL-rajapinnasta ja palauttaa tunnistettujen entiteettien URI:t. ARPA voidaan konfiguroida siten, että se muodostaa n-grammeja, joissa vain viimeinen sana on perusmuodossa. Tämä ratkaisee naiivin perusmuotoistamisen ongelman liittyen sanaliittoihin. Esimerkiksi taivutettu muoto "Karjalan armeijan" saadaan näin oikeaan perusmuotoon "Karjalan armeija". ARPA:n käyttämä SPARQL-palvelupiste ja kysely ovat täysin konfiguroitavissa käyttötapaukseen sopiviksi. Näiden lisäksi muun muassa n-grammien maksimikoko ja sanojen taivutusmuodot ovat määriteltävissä.

ARPA käyttää syötteen normalisoinnissa SeCo Lexical Analysis Services (LAS)[21] -järjestelmää, joka kokoaa yhteen kieliteknologisia työkaluja ja tarjoaa verkkopalveluja niiden käyttämiseksi. LAS tarjoaa viisi palvelua: kielen tunnistus, tekstin perusmuotoistaminen, morfologinen analyysi, sanojen taivutus ja tavuttaminen [Mä14]. Palvelut tukevat useita kieliä, joista suomi on tämän työn kannalta oleellinen. Suomen kielen analyysi perustuu Helsingin yliopiston Helsinki Finite State Transducer (HFST) -kehykseen [LAH+11] ja Omorfi-tietokantaan [Pir15].

Linkitystä varten toteutettiin Python-ohjelmointikielellä kirjasto *ARPA Linker*[22], joka käy läpi syötteenä annetun RDF-tiedoston ja ARPA-palvelua käyttäen tuottaa tämän palauttamat linkit. Kirjasto ahdollistaa ARPA-palveluun lähetettävän tekstin esikäsittelyn ja vastausten tarkemman yksilöinnin. *ARPA Linker* käyttää RDF-muotoisen tiedon käsittelyyn RDFlib-kirjastoa[23].

ARPA Linker-kirjastolla toteutettu linkitysprosessi voidaan jakaa Hacheyn ja kumppaneiden [HRN+13] kappaleessa 4 esitellyn jaottelun mukaisesti *erottimeksi*, *etsijäksi* ja *yksilöijäksi*. Prosessi jakoineen on esitetty kuvassa 16.

Erotinvaiheessa teksti käy läpi esikäsittelyn, jossa entiteettiviittaukset pyritään muokkaamaan tunnistettavaan muotoon. Tämä voi tarkoittaa esimerkiksi jonkin tietyn entiteetin kirjoitusasun muuttamista samaan muotoon kuin linkitettävässä ontologiassa tai kappaleessa 8.2 mainittujen henkilölistausten käsittelyä. Esikäsittelyn jälkeen teksti pilkotaan ARPA-järjestelmää käyttäen n-grammeiksi, joita käytetään potentiaalisina nimettyinä entiteetteinä. Henkilöiden tapauksessa n-grammeista karsitaan sellaiset, jotka eivät kirjoitusasunsa puolesta viittaa henkilöihin. Koska kyse on erisnimistä, karsittiin esimerkiksi n-grammit, joissa ei ole isolla alkukirjaimella alkavia sanoja.

[21] `http://demo.seco.tkk.fi/las/`

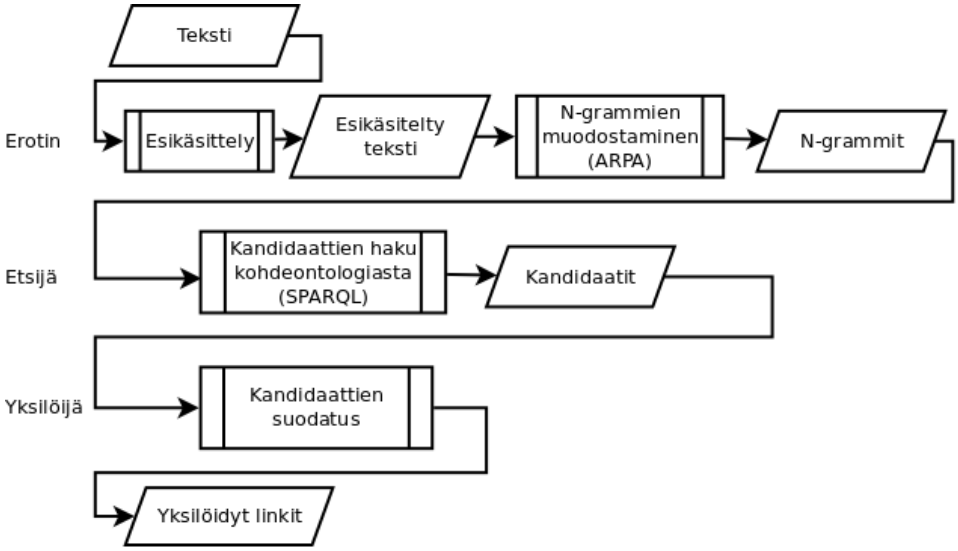
[22] `https://github.com/SemanticComputing/python-arpa-linker`

[23] `https://github.com/RDFLib/rdflib`

Karsiminen nopeuttaa varsinaista linkitysprosessia ja vähentää linkityksessä käytetyn laitteiston kuormitusta. Henkilöiden ja joukko-osastojen linkityksessä tekstistä muodostettiin n-grammeja, jossa $n \in [1, 5]$. Tästä syystä tekstikatkelmia syntyy huomattavasti enemmän kuin paikkojen linkityksen tapauksessa, jossa $n \in [1, 3]$.

Etsintävaiheessa kohdeontologiasta haetaan kandidaatit SPARQL-kyselyiden avulla vertaamalla resurssien merkkijonoesityksiä edellisessä vaiheessa muodostettuihin n-grammeihin. N-grammeja verrataan linkityskohteesta riippuen henkilöiden nimiin ja sotilasarvoihin, joukko-osastojen nimiin ja peitelukuihin tai paikkojen nimiin.

Kun kandidaatit on haettu, *yksilöijä* validoi ne pyrkimyksenä poistaa väärät kandidaatit. Esimerkiksi henkilöiden tapauksessa tässä vaiheessa kandidaatit käydään läpi ja ne pisteytetään sen mukaan, kuinka hyvin ne vastaavat aliluvussa 8.2 esitettyjä kriteerejä. Kandidaatit, jotka saavat määriteltyä kynnystä pienemmät pisteet, karsitaan.

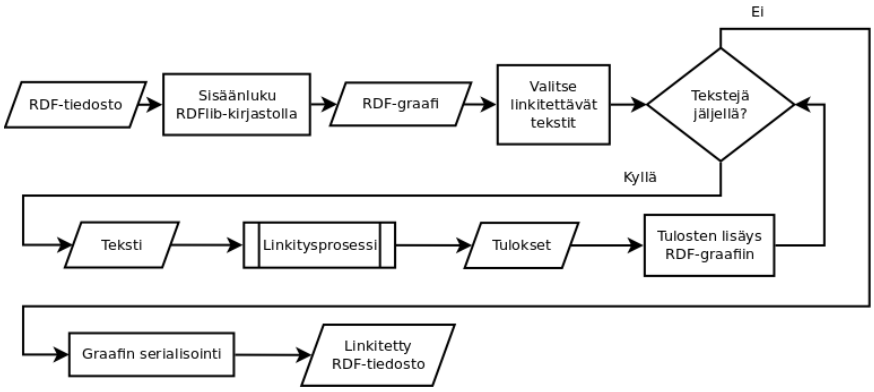


Kuva 16: Linkitysprosessi.

Kullekin linkitettävälle entiteettityypille (paikat, sotatoimihenkilöt ja joukko-osastot) luotiin oma ARPA-konfiguraatio ja linkitysohjelma käyttäen yhteisenä tekijänä yllä mainittua *ARPA Linker*-kirjastoa. Konfiguraatiot on julkaistu avoimena lähdekoodina GitHub-palvelussa[24]. Linkitettävä RDF-tiedosto luetaan kirjastolla graafiksi ja linkityksessä käytettävät tekstit poimitaan graafista annetun predikaatin mukaan. Tapahtumista henkilöitä ja paikkoja haetaan *dct:description*-ominaisuuden arvosta. Valokuvista henkilöitä haetaan samalla tavalla, mutta paikkoja ominaisuuden *wps:place_string*

[24] `https://github.com/SemanticComputing/warsa-linkers`

arvosta, johon alkuperäisessä metadatassa ollut tieto paikasta on tallennettu merkkijonona.



Kuva 17: Alkuperäinen linkitysten toteutus.

Alustavasti koko prosessi toteutettiin yhtenä suorituksena kuvan 17 mukaisesti. Jokainen linkitettävä teksti kävi yksitellen läpi linkitysprosessin (esikäsittely, kandidaattien haku ja yksilöiminen). Käsittelyn monimutkaistuessa linkityksen viemä aika kuitenkin kasvoi. Kun SA-kuva-datan linkitys henkilöihin kesti kaksi päivää ja kahdeksan tuntia, päätettiin prosessia muuttaa siten, että vaiheet voitiin ajaa erikseen.

Prosessi koostui lopulta kahdesta erillisestä kokonaisuudesta, jotka voitiin suorittaa omina ajoinaan:

1. n-grammien muodostus ARPA-järjestelmän avulla sekä
2. kandidaattien haku ja yksilöiminen.

Näistä ensimmäiseen kuului tekstin mahdollinen esikäsittely ennen n-grammien muodostamista. Jälkimmäinen kokonaisuus koostui kahdesta toimenpiteestä: kandidaattien noutaminen SPARQL-kyselyiden avulla kohdeontologiasta ja näiden karsinta.

Linkitysten yhteydessä data sai myös lopullisen luvussa 6 esitetyn muotonsa. Tapahtumadataan täydennettiin mallin mukaiset ominaisuudet, jotka yhdistävät tapahtumat ja linkitetyt entiteetit. Valokuvadata muunnettiin tässä vaiheessa CIDOC CRM:n mukaiseen muotoon säännöllisten lausekkeiden ja SPARQL-päivityskyselyiden avulla, ja tapahtumien tapaan dataan lisättiin linkityksiä kuvaavat ominaisuudet.

Linktysprosessi suoritettiin useaan otteeseen ja tulosten oikeellisuutta tarkasteltiin silmämääräisesti. Virheitä huomattaessa prosessia paranneltiin. Sotasampo-portaalissa on myös palautelomake, jonka kautta käyttäjät voivat antaa palautetta palvelusta. Käyttäjiltä saadun palautteen perusteella korjattiin myös linkityksiä erityisesti paikkojen suhteen. Asiantuntevien käyttäjien apu on erittäin hyödyllistä ei-itsestäänselvien virheiden huomaamisessa.

8.5 Tulokset ja arviointi

Linkitysten laadun arvioimiseksi valittiin aineistoista satunnaiset otannat nimettyjen entiteettien linkityksen tarkkuuden ja saannin mittaamiseksi. Valokuvista valittiin satunnaisotannaksi 100 kappaletta ja tapahtumista 50 kappaletta. Koska kävi ilmi, että joukko-osastojen tapauksessa mainintojen määrä jäi tällä otannalla hyvin pieneksi, otettiin niitä varten toiset samankokoiset otannat, joita käytettiin ainoastaan joukko-osastolinkitysten arvioinnissa. Käytettävissä ei ollut sovellusalan asiantuntijaa, joten arviointi oli tehtävä suurimmalta osin itse. Tämän ja mahdollisten puutteellisten tietojen takia linkitysten oikeellisuutta ei voitu välttämättä täysin varmistaa. Jos linkin oikeellisuudessa oli epäselvyyttä, jätettiin se pois arvioinnista. Esimerkiksi jos oli epäselvää, oliko linkittynyt henkilö kuvatekstissä nimetty henkilö, ei linkitys vaikuttanut arviointiin. Jos epäselvässä tapauksessa sama maininta oli tuottanut useamman kuin yhden linkin, voitiin kuitenkin olla varmoja, että kaikki paitsi yksi linkki olivat vääriä. Tällaisessa tapauksessa ylimääräiset linkit laskettiin vääriksi linkeiksi, ja yksi linkki jätettiin pois laskuista. Valokuvien arvioinnissa kaksi henkilöä ja kaksi paikkaa jätettiin arvioinnin ulkopuolelle. Tapahtumien arvioinnissa kaikki linkit saatiin arvioitua.

Vertailun vuoksi otannat linkitettiin myös yksinkertaisesti ilman esikäsittelyä ja yksilöintiä vertaamalla suoraan tekstissä esiintyviä sanoja linkitettävien entiteettien tekstimuotoisiin nimiin kohdeontologioissa yksilöimättä entiteettejä tämän tarkemmin. Nimet ovat ontologioissa ominaisuuksien *skos:prefLabel* (ensisijainen nimi) ja *skos:altLabel* (toissijainen nimi) arvoja. Henkilöresurssin ensisijainen nimi on toimijaontologiassa henkilön koko nimi. Henkilöiden osalta verrokkilinkitys tehtiin myös vertaamalla tekstiä edellä mainittujen ominaisuuksien sijaan henkilön sukunimeen. ARPA-asetukset, kuten n-grammien pituudet, pidettiin samoina kuin varsinaisessa linkityksessä. Esimerkiksi henkilöiden sukunimiä käyttävä verrokkilinkitys toteutettiin seuraavasti: 1) perusmuotoista teksti ja pilko se 1-, 2-, 3-, 4- ja 5-grammeiksi ARPA-järjestelmän avulla, 2) vertaa n-grammeja henkilöontologian henkilöiden sukunimiin ja 3) linkitä kaikki henkilöt, joiden sukunimi vastaa n-grammia.

Siinä missä henkilö- ja paikkamainintojen poimiminen tekstistä on ihmiselle melko suoraviivaista, joukko-osastojen tapauksessa oli tapauksia, joissa oli epäselvää pitäisikö jokin maininta luetella joukko-osastomaininnaksi. Tällaisia mainintoja ovat esimerkiksi “Nyhamnin linnake” ja “suojajoukot”. Koska tämä tulkinnanvaraisuus olisi voinut vaikuttaa arviointiin, pyydettiin toista tutkimusryhmän jäsentä merkitsemään joukko-osastomainnnat otannan teksteistä.

Tulokset on esitetty valokuvien osalta taulukossa 7 ja tapahtumien osalta taulukossa 8. Naiivit linkitykset on eroteltu taulukoissa alaindeksillä n: esimerkiksi “Paikat$_n$”. Lisäksi “Henkilöt$_{ns}$” kuvaa sukunimiä käyttävän linkityksen tuloksia. Taulukoissa käytetyt lyhenteet on esitetty seuraavassa:

Kohde Ontologia, jonka resursseihin linkitys kohdistui.

N Mainintojen lukumäärä tekstissä. Tämä luku ei sisällä arvioinnin ulkopuolelle jätettyjä mainintoja.

TP Oikeiden positiivisten (true positive) linkkien lukumäärä. Ts. sellaiset linkit, jotka yksilöivät nimetyn entiteetin oikein.

FP Väärien positiivisten (false positive) linkkien lukumäärä. Ts. sellaiset linkit, jotka yksilöivät väärän entiteetin maininnan perusteella.

FN Väärien negatiivisten (false negative) linkkien lukumäärä. Ts. sellaiset nimetyt entiteetit, jotka löytyvät kohdeontologioista, mutta joita ei ole linkitetty.

TN Oikeiden negatiivisten (true negative) linkkien lukumäärä. Ts. sellaiset nimetyt entiteetit, joita ei löydy kohdeontologioista, ja joita ei ole linkitetty.

P Tarkkuus. $P = \frac{TP}{TP+FP}$

$\mathbf{R}_{ont.}$ Saanti ontologiasta. $R_{ont.} = \frac{TP}{TP+FN}$

$\mathbf{R}_{kaikki}$ Saanti suhteessa kaikkiin mainintoihin riippumatta siitä, löytyykö kyseistä resurssia ontologiasta. $R_{kaikki} = \frac{TP}{N}$

$\mathbf{F}_1$ F_1-tulos, joka on tarkkuuden ja saannin harmoninen keskiarvo: $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$. Tässä $R = R_{ont.}$.

Kohde	**N**	**TP**	**FP**	**FN**	**TN**	**P**	$\mathbf{R}_{ont.}$	$\mathbf{R}_{kaikki}$	$\mathbf{F}_1$
Henkilöt	44	25	8	3	16	0.76	0.89	0.57	0.81
Henkilöt$_n$	44	1	11	27	16	0.08	0.04	0.02	0.05
Henkilöt$_{ns}$	44	22	2709	9	16	0.01	0.71	0.47	0.02
Paikat	91	54	15	16	21	0.77	0.76	0.59	0.77
Paikat$_n$	91	61	1947	10	21	0.03	0.84	0.66	0.06
Joukko-osastot	28	21	2	7	0	0.91	0.75	0.75	0.82
Joukko-osastot$_n$	28	7	10	21	0	0.41	0.25	0.25	0.31

Taulukko 7: Tarkkuus ja saanti valokuvien linkityksessä.

Kuten tuloksista nähdään, naiivi linkitystapa ei tuota hyviä tuloksia: joko tarkkuus tai saanti on kelvoton. Tulos voidaan todeta myös linkittämällä koko aineisto tällä tavoin. Naiivi linkitys ilman yksilöintiä tuottaa

Kohde	**N**	**TP**	**FP**	**FN**	**TN**	**P**	$\mathbf{R}_{ont.}$	$\mathbf{R}_{kaikki}$	$\mathbf{F}_1$
Henkilöt	34	26	0	7	1	1.00	0.79	0.76	0.88
Henkilöt$_n$	34	6	1	27	1	0.86	0.18	0.18	0.03
Henkilöt$_{ns}$	34	27	412	6	1	0.06	0.81	0.79	0.11
Paikat	72	28	4	7	37	0.88	0.80	0.39	0.84
Paikat$_n$	72	30	1048	5	37	0.03	0.86	0.42	0.05
Joukko-osastot	27	24	5	3	0	0.83	0.89	0.89	0.86
Joukko-osastot$_n$	27	17	9	10	0	0.65	0.63	0.63	0.64

Taulukko 8: Tarkkuus ja saanti tapahtumien linkityksessä.

esimerkiksi henkilöiden sukunimien tapauksessa erittäin suuren määrän linkityksiä suhteessa aineiston kokoon: valokuvien henkilölinkityksiä löytyy tällä tavoin lähes kaksi miljoonaa kappaletta. Tämä tarkoittaa keskimäärin yli 45 henkilöä yhtä valokuvaa kohden niissä valokuvissa, joiden kuvauksista muodostettiin vähintään yksi henkilölinkki. Näin montaa henkilöä ei kuitenkaan käytännössä tekstissä voi olla mainittuna jo senkään takia, että kuvausten yhteenlaskettu sanamäärä on alle miljoona. Esikäsittelyn ja yksilöinnin jälkeen henkilöitä löytyykin huomattavasti kohtuullisempi määrä: 23 293 linkkiä 16 638:sta valokuvasta.

Henkilöiden linkityksen toteutukseen käytettiin eniten aikaa, ja toteutus oli monimutkaisin. Yksilöinnin rikkaus näkyykin korkeana F_1-arvona (88%). Henkilöiden yleinen saanti (R_{kaikki}) valokuvien osalta on matala: 57%. Tämä johtuu siitä, että valokuvissa esiintyy monenlaisia henkilöitä aina venäläisistä sotavangeista lapsiin. Tapahtumien kuvauksissa sen sijaan mainitaan lähinnä tunnetumpia henkilöitä, joten tapahtumien osalta yleinen saanti oli korkeampi (76%). Sukunimiä käyttävä verrokkilinkitys saavutti tapahtumien tapauksessa hieman parempaan saantiin löytäen yhden henkilön enemmän. Yksilöivän linkityksen saantia voitaisiinkin mahdollisesti parantaa siten, että pelkän sukunimen sisältäviä mainintoja ei suljettaisi pois kandidaattihakuvaiheessa, vaan myös näitä pyrittäisiin tunnistamaan yksilöintivaiheessa.

Paikkojen linkitys osoittautui haastavaksi. Paikannimien monitulkintaisuuden vuoksi linkityksiä vääriin paikkoihin syntyy helposti, sillä osumien automaattinen varmennus on vaikeaa. Historiallisten paikkojen yksilöiminen vaatii tietoa, joka ei suoraan ilmene paikan mainitsevasta tekstistä. Henkilöiden tapauksessa tarkempaan yksilöintiin päästiin ottamalla huomioon esimerkiksi henkilön kaikki nimet ja mahdollinen sotilasarvo. Paikkojen kohdalla tämä ei ole mahdollista, sillä usein paikka pitäisi yksilöidä vain sen nimen perusteella. Vaikeuksista huolimatta paikkalinkityksessä päästiin

kuitenkin tyydyttäviin tuloksiin, vaikka yksilöintimenetelmä oli melko yksinkertainen: 77% valokuvissa ja 88% tapahtumissa. Valokuvien tapauksessa väärien positiivisten tulosten määrää vähensi se, että lähdetekstinä käytettiin paikkatietuetta, jolloin koko linkitettävän tekstin tiedettiin sisältävän vain paikkatietoa. Tällöin sekaannuksen mahdollisuutta esimerkiksi henkilöiden nimien kanssa ei juuri ollut. Toisaalta tässä suhteessa tärkeässä asemassa oli myös lista poissuljettavista paikannimistä.

Tulokset ovat lähellä Gloverin ja kumppaneiden [GTB+10] paikkayksilöijän tuloksia, joissa tarkkuudeksi arvioitiin 82%–92% riippuen aineistosta. Tulokset tosin eivät ole suoraan verrannollisia, sillä heidän aineistonsa olivat toki erilaisia. Heidän yksilöintimenetelmänsä oli myös monipuolisempi, ja esimerkiksi samassa kontekstissa mainittujen paikkojen välimatkan käyttäminen yksilöinnissä voisi olla hyödyllistä myös tämän työn aineistojen kohdalla.

Yleinen saanti (R_{kaikki}) jäi kuitenkin paikkojen osalta alhaiseksi: 59% valokuvien ja 44% tapahtumien tapauksessa. Tapahtumien kuvauksissa lähes puolet nimetyistä paikoista oli sellaisia, joita ei löytynyt kohdeontologioista. Huono saanti johtuu siitä, että historiallisten paikkojen ontologiasta puuttuu monia sotien kannalta hyvin keskeisiä paikkoja, kuten Karjalankannas, ja toisaalta erityisesti tapahtumissa mainitaan usein muita maita ja Suomen ulkopuolisia alueita, joita ei käytetyistä paikkaontologioista löydy.

Joukko-osastojen tunnistus onnistui tarkkuuden osalta hyvin: 91% valokuvissa ja 83% tapahtumissa. Tämä johtuu siitä, että osastoihin viitataan aineistoissa melko yhtenäisesti. Lisäksi joukko-osastojen nimet yksilöivät ne hyvin, eikä nimien monitulkintaisuus ollut suuri ongelma. Näistä syistä johtuen hyviin tuloksiin päästiin melko yksinkertaisella käsittelyllä, joskin mainintojen määrä oli myös pienin verrattuna muihin linkitettyihin entiteetteihin. Joukko-osastojen yksilöinnin suoraviivaisuudesta verrattuna muihin entiteettityyppeihin kertoo sekin, että naiivi linkitystapa pärjäsi osastojen tapauksessa selvästi parhaiten: F_1-arvo oli 31% valokuvissa ja 64% tapahtumissa siinä missä seuraavaksi paras naiivi F_1-tulos oli henkilöiden 11%.

Täysin ongelmatonta ei joukko-osastojen linkitys kuitenkaan ollut. Valokuvissa seitsemän joukko-osastoa jäi tunnistamatta. Tämä johtui siitä, etteivät viittaukset vastanneet ontologian nimiä: esimerkiksi erään valokuvan kuvaustekstissä “Veripalveluosaston R 31:n johtaja tohtori A. Heikel suorittaa veren laskun” viitataan osastoon Veripalveluryhmä 31 (Verip.R 31). Lisäksi tietokannan tekstihakuindeksi vaikutti linkitykseen: vaikka osastojen nimiä verrattiin ilman välimerkkejä, rajoitettiin entiteettien määrää kyselyssä ensin tekemällä tekstihaku n-grammilla. Indeksistä ei löydy esimerkiksi merkkijonolla “Kan.R.E” osastoa, jonka lyhenne on “Kan.RE”.

Kaiken kaikkiaan tarkkuudessa ja ontologioiden sisäisessä saannissa päästiin kaikkien aineistojen osalta hyviin lukemiin. Erityisesti tulokset olivat huomattavasti parempia kuin ilman yksilöintiä tehdyssä verrokkilinkityksessä.

Linkillä aineistosta toiseen on myös semanttinen merkitys. Esimerkiksi tapahtumien kuvauksista tunnistetut henkilöt kuvattiin yhdistämällä valokuva henkilöontologiaan CIDOC CRM:n ominaisuuden *P11 had participant* avulla. Tämä ominaisuus implikoi, että linkitetty henkilö on ollut osallisena kyseisessä tapahtumassa ollen samassa paikassa samaan aikaan [CDG+15]. Vaikka joku henkilö mainitaan tapahtuman kuvauksessa, ei henkilö kuitenkaan ole välttämättä ollut samassa paikassa tai muutenkaan varsinaisesti osallisena tapahtumassa. Valokuvien tapauksessa ei myöskään ole välttämätöntä, että kuvatekstissä mainittu henkilö olisi kuvassa tai läsnä, kun kuva on otettu. On kuitenkin hyvin vaikea päätellä automaattisesti, millä tavalla esimerkiksi kuvatekstissä mainittu henkilö liittyy valokuvaan. Linkitys toteutettiinkin nimenomaan nimettyjen entiteettien linkityksen näkökulmasta. Myös tulokset arvioitiin samasta näkökulmasta, eikä siinä otettu kantaa siihen, oliko henkilö varsinaisesti osallinen tapahtumaan tai esiintyikö hän valokuvassa.

Useimmiten mainittujen henkilöiden suhde tapahtumiin on, että he ovat osallisia kyseisissä tapahtumissa. Samaten useimmiten kuvatekstissä mainittu henkilö viittaa valokuvattuun henkilöön, mainittu paikka kuvaa tapahtuman paikkaa ja niin edelleen. Vaikka tästä voi seurata virheellisiä päätelmiä, semanttisesti rikkaampi linkitys nähtiin tärkeämpänä kuin tietojen täydellinen virheettömyys.

9 Datajulkaisu

Aineistot julkaistiin linkitettynä datana Linked Data Finland (LDF.fi) -alustalla osana Sotasampo-datapalvelua[25]. Tapahtuma- ja valokuva-aineistot julkaistiin omissa RDF-graafeissaan `http://ldf.fi/warsa/events` ja `http://ldf.fi/warsa/photographs`. Näiden lisäksi ajankohdat löytyvät graafista `http://ldf.fi/warsa/events/times`. Sotasammon aineistoja yhdistävä malli on määritelty graafissa `http://ldf.fi/schema/warsa`, josta löytyvät myös tämän työn aineistojen luokkien ja ominaisuuksien määritykset.

LDF.fi [HTAM14] on linkitetyn datan julkaisualusta, joka pyrkii parantamaan datajulkaisujen laatua tarjoamalla palveluita, jotka helpottavat kaikkien luvussa 2 seitsemän datajulkaisun laatua mittaavan tähden saavuttamista. Alusta tarjoaa aineistoille SPARQL-palvelupisteen, kotisivun ja erilaisia työkaluja aineiston hyödyntämiseen, kuten linkitetyn datan selaimen. Aineiston sivulta löytyy kyseisen julkaisun dokumentaatio, esimerkkiresurssejaja kyselyitä. Sotasammon datajulkaisu saavuttaakin datajulkaisun laadun seitsemästä tähdestä kuusi:

1. data on julkaistu avoimessa strukturoidussa muodosssa avoimella lisenssillä[26] (tähdet 1-3),

[25] `http://www.ldf.fi/dataset/warsa`

[26] CC BY 4.0 (`http://creativecommons.org/licenses/by/4.0/`)

2. resurssit on yksilöity HTTP-URI-tunnuksilla,

3. aineistot on linkitetty muihin aineistoihin,

4. julkaisun malli on eksplisiittisesti määritelty ja dokumentoitu.

Ainoastaan seitsemäs tähti jää vielä saavuttamatta, sillä aineistojen ja mallin vastaavuutta ei ole formaalisti varmistettu.

HTTP-pyynnön lähettäminen Sotasammossa yksilöiviin HTTP-URI-tunnisteisiin palauttaa resurssiin liittyviä tietoja, ja resurssit on linkitetty paitsi Sotasammon sisältämiin resursseihin myös ulkopuolisiin tietokantoihin. Esimerkiksi tapahtuman "Sotamarsalkka Mannerheim tarkasti Lapualla ulkomaisista vapaaehtoisista muodostetun Osasto Sisun." URI on `http://ldf.fi/warsa/events/event_536`, ja valokuvan, jossa "Sotamarsalkka Mannerheim saapuu seurueineen Mainilaan, josta hän Uuden Alakylän kohdalla erään venäläisen korsun katolta katseli Pietarin paloja", URI on `http://ldf.fi/warsa/photographs/sakuva_57717`. Sisältöneuvottelun vuoksi ihmiskäyttäjille tarjotaan HTML-sivu, joka sisältää resurssin tiedot, ja koneille tiedot tarjotaan sarjallistettuna halutussa muodossa.

Aineistot ovat käytettävissä SPARQL-rajapinnan[27] kautta, mikä mahdollistaa niiden monipuolisen kyselyn ja käytön sovelluskehityksessä.

10 Sovellukset

Tulosten visualisointia varten aineistoille toteutettiin verkkosovellukset osana Sotasampo-portaalia[28]. Koska aineistot haluttiin yhdistää saumattomasti, sovellukset tapahtumille ja valokuville toteutettiin osana yhteistä kokonaisuutta. Omien sovellustensa lisäksi tapahtumat ja kuvat ovat osa muita Sotasammon sovelluksia: esimerkiksi paikkanäkymässä tapahtumia ja valokuvia voi selata kartan avulla ja henkilönäkymässä esitetään kuhunkin henkilöön liittyvät tapahtumat ja valokuvat. Sovellus hakee datan SPARQL-kyselyjä käyttäen Linked Data Finland -palvelusta[29], jossa aineistot on julkaistu linkitettynä datana.

Tässä työssä toteutetulle kokonaisuudelle eroteltiin seuraavat käyttötapaukset:

1. Käyttäjä haluaa löytää tietoa Suomen vaiheista toisen maailmansodan aikana.

2. Käyttäjä on kiinnostunut sodanajan valokuvista ja haluaa hakea niistä häntä kiinnostavia kuvia.

[27] `http://ldf.fi/warsa/sparql`
[28] `http://www.sotasampo.fi`
[29] `http://www.ldf.fi`

3. Käyttäjä haluaa tutkia yksittäisen henkilön – esimerkiksi sukulaisensa – tarinaa sotien aikana.

4. Tutkija haluaa muodostaa uutta tietoa yhdistämällä tietoa eri lähteistä.

Näistä käyttötapausta 3 palvelee Sotasammon henkilösovellus[30], joka aineistojen välisten linkkien avulla tarjoaa käyttäjälle mahdollisuuden hakea henkilöitä ja nähdä näiden vaiheet sotien aikana visualisoituna aikajanalla ja kartalla. Toisaalta tutkija voi koostaa henkilön tiedot ja vaiheet vapaasti avoimen datajulkaisun avulla. Molemmat käytöt mahdollistaa tässä työssä toteutettu linkitys muihin aineistoihin ja avoin datajulkaisu. Datan julkaiseminen avoimesti SPARQL-palvelupisteen avulla palvelee erityisesti myös viimeistä käyttätapausta: tutkija voi vapaasti muodostaa kyselyitä dataan ja analysoida sitä. Täten toteutettavan sovelluksen toteutettaviksi jäävät käyttötapaukset 1 ja 2.

Sovelluksen toiminnallisuus toteutettiin JavaScriptillä — merkittävimmät apuna käytetyt kirjastot olivat AngularJS-sovelluskehys[31], Google Maps -ohjelmointirajapinta[32], Simile Timeline -aikajanakirjasto[33] ja Timemap.js[34], joka yhdistää kartan ja aikajanan.

AngularJS on JavaScript-ohjelmistokehys selainpohjaisten web-sovellusten toteuttamiseen [Goo16]. Koska vaatimuksena oli rakentaa sovellus SPARQL-palvelupisteen varaan, soveltuu AngularJS hyvin sovelluksen kehykseksi, sillä se ei esimerkiksi tee mitään oletuksia palvelimelta saatavan datan muodosta.

Vaikka vapaan lähdekoodin JavaScript-aikajanakirjastoja on tarjolla useita, kaikki eivät ole tarkoitettu sovelluksiin, jotka käsittelevät suurta määrää tapahtumia. Tässä työssä visualisoitavien tapahtumien määrä on melko suuri, joten tapahtumasovelluksen toteutuksen oli tuettava suuren tapahtumamäärän näyttämistä. Simile Timeline -kirjasto soveltuu suurellekin määrälle tapahtumia. Lisäksi Timemap.js-kirjasto yhdistää Simile Timeline -aikajanan karttaan, jolloin tapahtumat voidaan visualisoida paitsi ajallisesti myös spatiaalisesti. Ongelma näiden kirjastojen käytössä on, että kumpaakaan ei enää ylläpidetä. Tästä huolimatta kirjastot todettiin parhaiksi tämän työn käyttötarkoitukseen.

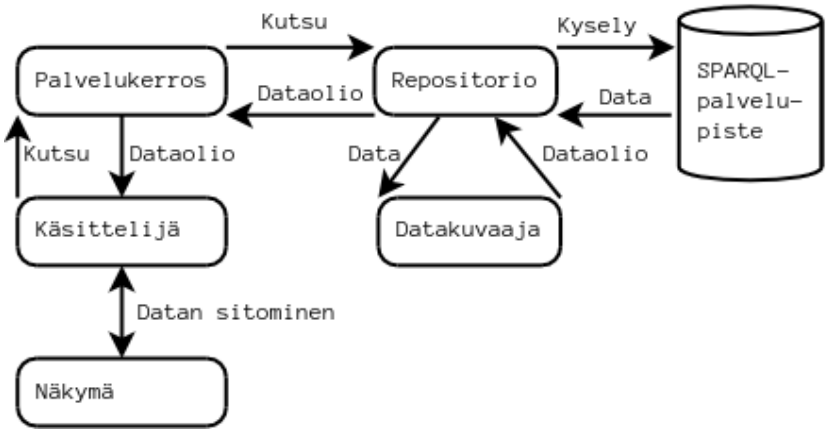
Sovelluksen kehityksen yhteydessä toteutettiin yleiskäyttöinen AngularJS-kirjasto SPARQL-kyselyiden tekemiseen ja tulosten käsittelyyn[35]. Tämän lisäksi toteutettiin SPARQL-fasettihakukirjasto *SPARQL Faceter* [KHH16a] valokuvasovelluksen käyttöön.

Sovelluksessa eri komponenttien vastuualueet on pyritty pitämään erillisinä, mikä helpottaa ylläpitoa [HL95]. Arkkitehtuurissa tietokantaliitäntä on

[30] `http://www.sotasampo.fi/fi/persons`
[31] `https://angularjs.org/`
[32] `https://developers.google.com/maps/`
[33] `http://www.simile-widgets.org/timeline/`
[34] `https://code.google.com/archive/p/timemap/`
[35] `https://github.com/SemanticComputing/angular-paging-sparql-service`

erotettu omaksi kerroksekseen. Tämä kerros toteuttaa SPARQL-kyselyt ja muuttaa tulokset datakuvaajan avulla olioiksi, joita muut sovelluksen osat käyttävät. Vaikka SPARQL-palvelupiste palauttaakin tulokset JavaScript-yhteensopivassa JSON-muodossa, on tulosten kuvaaminen yksinkertaisimmiksi olioiksi kuitenkin hyödyllistä. Datakuvaaja muuntaa SPARQL-tulokset muotoon, jota on helpompi käsitellä, ja jossa samaa resurssia kuvaavat tulosrivit on yhdistetty yhdeksi olioksi. Eri resurssityypeille, kuten esimerkiksi tapahtumat ja kuvat, on omat tietokantaliitännät eli repositoriot. Eri tyyppisille resursseille voidaan myös määrittää erilaisia datakuvaajia.

Palvelukerros toteuttaa liiketoimintalogiikan (business logic) ja pyytää tietokantaliitäntäkerrokselta dataa kuvaavat oliot. Käsittelijä (controller) määrittää toiminnot, jotka ovat käyttäjän saatavilla näkymässä. Käsittelijä saa palvelukerrokselta näkymässä näytettävän datan. Näkymässä data sidotaan HTML-elementteihin ja näytetään käyttäjälle verkkosivulla. Käyttäjän valitessa jonkin toiminnon sivulla kutsutaan siis sivun käsittelijässä määriteltyä funktiota, joka mahdollisesti pyytää jonkin palvelun ja siten repositorion kautta SPARQL-palvelupisteestä tietoja. Nämä tiedot kuvataan sopiviksi olioiksi ja esitetään käyttäjälle. Tämä sovelluksen arkkitehtuuri ja datan kulku on esitetty kuvassa 18.



Kuva 18: Sovelluksen arkkitehtuuri.

Seuraavissa kappaleissa kuvataan tapahtuma- ja valokuvasovellukset tarkemmin.

10.1 Tapahtumat

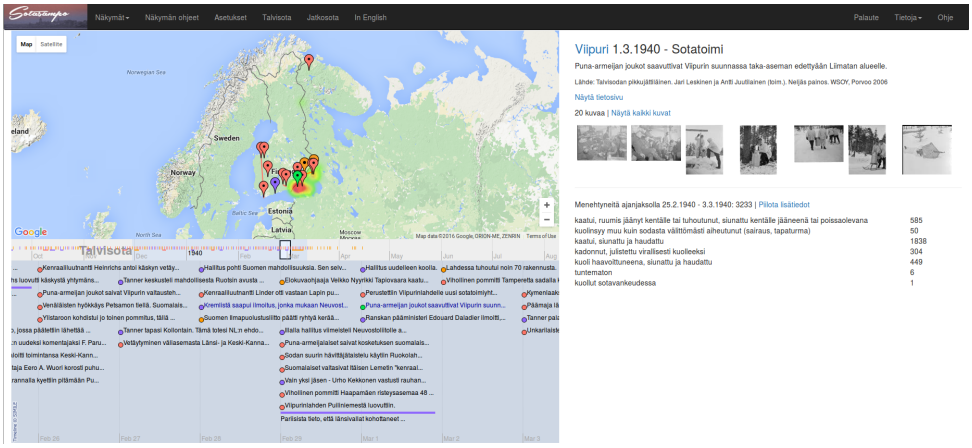
Tapahtumiin liittyy keskeisesti sekä aika- että paikkaulottuvuus ja niiden avulla voi hahmottaa sodan kulun ajan mukaan. Tästä syystä aikajana ja kartta ovat intuitiivinen tapa visualisoida niitä. Sotasammon tapahtumanäkymä yhdistää tapahtumien ajalliset ja paikalliset ulottuvuudet esittäen

ne käyttäjälle aikajanalla ja kartalla. Sovellus luo näin käyttäjälle tiedon hahmottamista helpottavan kontekstin.

Käyttäjä voi valita tapahtuman, jolloin siitä näytetään tietoa kuten kuvaus, paikkatiedot ja ajankohta. Tämän lisäksi näytetään tapahtumaan liittyvät valokuvat ja tapahtumaan linkitettyjen henkilöiden ja joukko-osastojen tiedot. Tapahtumiin liittyvät valokuvat haetaan tapahtuman ajankohdan ja paikan perusteella – valokuvien hakuun liittyvät asetukset ovat myös käyttäjän muokattavissa. Oletusarvoisesti valokuva näytetään tapahtuman yhteydessä, jos se on otettu samassa kunnassa korkeintaan kolme päivää tapahtuman jälkeen.

Kartalla esitetään tapahtumien lisäksi tietoa sodassa menehtyneistä Suomen sotilaista: kartalla näytetään lämpökartta menehtyneistä ja kartan oikealla puolella on tilastotietoa menehtyneistä siltä ajalta, joka on aikajanalla kullakin hetkellä valittuna.

Tapahtumasovelluksen käyttöliittymä on nähtävissä kuvassa 19. Käyttäjä on valinnut tapahtuman aikajanalta, jolloin kartan oikealle puolelle ilmestyy tietoa kyseisestä tapahtumasta, kuvia tapahtuman paikasta sekä hyperlinkkejä muihin linkitettyihin Sotasammon resursseihin. Kuvan tapahtuman tapauksessa tapahtuma on yhdistetty paikkaan, jolloin se näkyy aikajanan lisäksi kartalla, ja paikkalinkin valitsemalla käyttäjä pääsee Sotasammon paikkanäkymään[36], josta hän voi tutkia muita samaan paikkaan liittyviä resursseja. Tapahtumaan liittyy myös valokuvia, joita käyttäjä voi selata. Käyttäjä voi valokuvia selatessaan siirtyä valokuvan omalle sivulle, jolloin käyttäjä voi selata myös kuvaan liittyviä muita resursseja ja muita samankaltaisia kuvia.



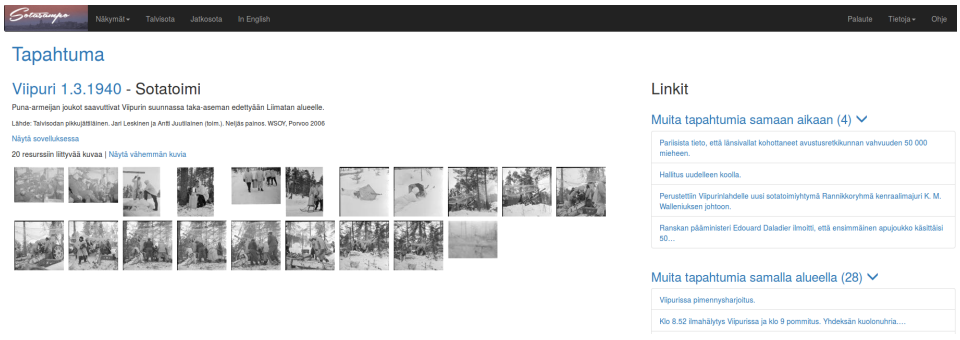
Kuva 19: Kuvankaappaus Sotasammon tapahtumasovelluksesta.

Tapahtumien visualisoinnin haasteena on samanaikaisten tapahtumien

[36] http://www.sotasampo.fi/fi/places

suuri määrä ja käyttäjälle mielenkiintoisten tapahtumien nostaminen esiin. Kaikkien tapahtumien esittäminen aikajanalla samanaikaisesti aiheuttaa aikajanan paisumisen hyvin suureksi ja vaikeuttaa mielekkäiden tapahtumien löytämistä. Tästä syystä aikajanalla näytetään poliittiset tapahtumat ja sotilaalliset tapahtumat taisteluita lukuun ottamatta. Taistelut on visualisoitu Sotasammon joukko-osastonäkymässä[37] joukko-osastoittain.

Aikajananäkymän lisäksi kullekin tapahtumalle tuotetaan myös tietosivu, jossa käyttäjälle tarjotaan tapahtuman tietojen ja siihen liittyvien valokuvien lisäksi suosituslinkkejä muihin tapahtumiin. Sivulla on suosituksia samaan aikaan ja toisaalta samalla alueella tapahtuneisiin tapahtumiin. Kuvassa 20 on kuvankaappaus erään tapahtuman tietosivusta.



Kuva 20: Kuvankaappaus tapahtuman tietosivusta.

10.2 Valokuvat

Kuten tapahtumiin myös valokuviin liittyy vahvasti paikka ja aika. Valokuvat voitaisiin siis myös visualisoida tapahtumien tapaan aikajanalla ja kartalla. Käyttötapauksena nähtiin kuitenkin ensisijaisesti valokuvien tutkiminen hakukäyttöliittymän avulla.

Valokuvia varten toteutettiin hakunäkymä, jossa valokuvia voi selata ja hakea fasettien avulla. Fasettihaun tarkoituksena on mahdollistaa asteittain etenevä haku ja tulosten rajaus. Fasetit ovat kategorioita, joiden arvoilla tulosjoukkoa voi rajata. Käyttäjän valitessa arvon fasetista, muiden fasettien valittavissa olevat arvot rajataan sellaisiin, joiden valinta tuottaa tuloksia. Tällä tavoin estetään turhauttavien tyhjien tulosjoukkojen ilmeneminen [Tun09].

Fasettihakua varten toteutettiin *SPARQL Faceter* [KHH16a] -kirjasto[38]. Sen pyrkimyksenä on tarjota helppokäyttöinen kirjasto SPARQL-perustaisen fasettihaun luontiin. Kirjasto on täysin selainpohjainen ja vaatii ainoastaan

[37] `http://www.sotasampo.fi/fi/units`
[38] `https://github.com/SemanticComputing/angular-semantic-faceted-search`

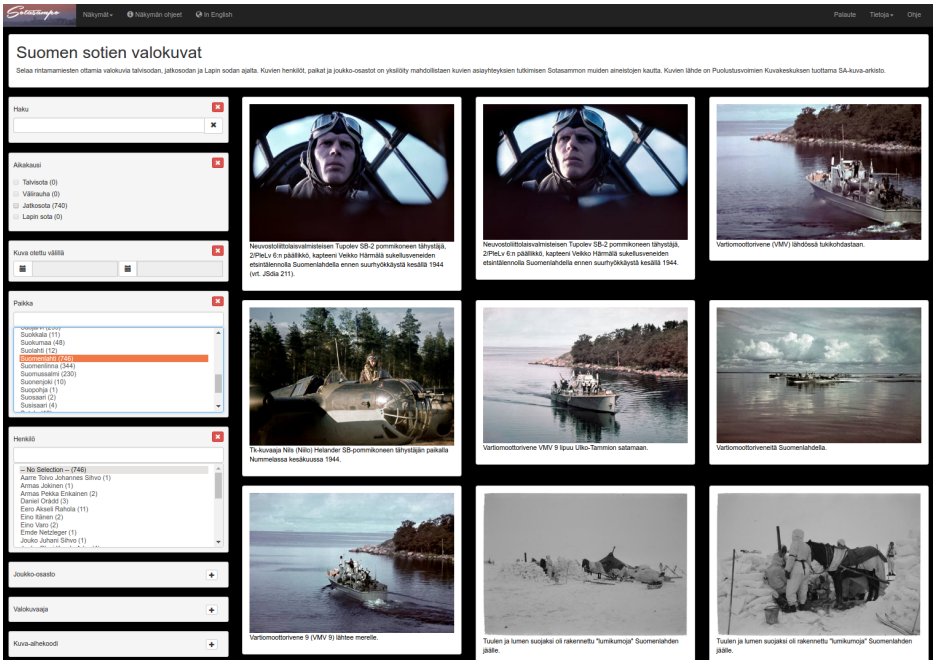
SPARQL-palvelupisteen, johon kyselyt kohdistetaan. Toteutuksessa on pidetty tärkeänä, että kirjaston voi ottaa käyttöön helposti, joten pakolliset konfigurointimäärittelyt on pidetty mahdollisimman vähäisinä. Jotta kirjasto olisi hyödyllinen mahdollisimman monessa kontekstissa ja erilaisilla datajoukoilla, tarjotaan myös mahdollisuus edistyneempään asetusten määrittelyyn.

SPARQL Faceter koostuu joukosta AngularJS-komponentteja, joita yhdistelemällä fasettihaku toteutetaan. Kirjasto määrittelee viisi fasettityyppiä: arvot listaava, valintaruutuja käyttävä, aikavälin määrittävä, hierarkkinen ja tekstihakuun perustuva fasetti. Arvot listaava fasetti – perusfasetti – listaa kaikki valittavat fasetin arvot. Näin toimii myös hierarkkinen fasetti mutta esittää arvot kahdessa tasossa määritellyn hierarkian mukaisesti. Valintaruutufasetissa kukin valintaruutu ja tämän valinnasta seuraava rajoite määritellään erikseen. Aikavälifasetti antaa käyttäjän rajoittaa tuloksia päivämäärien avulla. Tekstihakufasetti hakee käyttäjän antamaa tekstiä annetun ominaisuuden arvoista ja rajaa näin tulosjoukkoa. Näiden lisäksi kehittäjä voi määritellä omia sovelluskohtaisia fasetteja. Valokuvasovelluksen lisäksi kirjasto on käytössä Sotasammon Menehtyneet-sovelluksessa[39], joka tarjoaa käyttäjille mahdollisuuden hakea sodissa menehtyneitä henkilöitä.

Valokuvasovellus määrittää tekstihakukentän käyttäen tekstihakufasettia ja seitsemän fasettia: aikakausi, kuvanottopäivä, paikka, henkilö, joukko-osasto, valokuvaaja ja kuva-aihekoodi. Näistä aikakausi käyttää valintaruutufasettia, kuvanottopäivä aikavälifasettia, ja loput perusfasettia. Kuvassa 21 on kuvankaappaus valokuvasovelluksesta. Käyttäjä on valinnut ruudun vasemmassa laidassa olevasta paikkafasetista Suomenlahden, jolloin ruudun oikealla puolella näytetään kyseisessä paikassa otetut valokuvat. Lisäksi paikkafasetin alapuolella oleva henkilöfasetti näyttää henkilöt, joihin liittyviä valokuvia tulosjoukossa on ja näiden valokuvien lukumäärät. Käyttäjä voi henkilön valitsemalla rajoittaa tulosjoukon niihin valokuviin, jotka on otettu Suomenlahdella ja joissa kyseinen henkilö esiintyy.

Sovellusta voidaan arvioida vertaamalla sitä SA-kuvapalveluun [Puo], joka tarjoaa hakukäyttöliittymän SA-kuva-arkiston aineisoon. SA-kuvapalvelussa valokuvia voi hakea vapaalla tekstihaulla ja päivämäärän sekä sodan mukaan. Sotasampoon toteutettu valokuvasovellus toteuttaa samat hakuvaihtoehdot mutta näiden lisäksi myös fasetit kuvissa mainituille henkilöille, joukko-osastoille, kuvan ottajalle, ja kuvanottamispaikalle. Fasettien avulla hakua voi rajata tarkemmin esimerkiksi jonkun tietyn valokuvaajan mukaan, mikä ei onnistu SA-kuvapalvelun sovelluksessa. Useampien hakuehtojen lisäksi sovellus tarjoaa käyttäjälle lisätietoa kuvista: henkilölinkkien takia käyttäjä voi lukea kuka kuvan henkilö on ilman, että tietoa tarvitsisi etsiä jostain toisesta lähteestä. Kiinnostavan henkilön kaikki valokuvat voi hakea henkilöfasetin avulla sopivan hakusanan keksimisen sijaan. Paikkalinkkien avulla käyttäjä voi tutkia kartalla, missä kuva on otettu, ja paikan kautta löytää

[39] `http://www.sotasampo.fi/fi/casualties`



Kuva 21: Kuvankaappaus Sotasammon valokuvasovelluksesta.

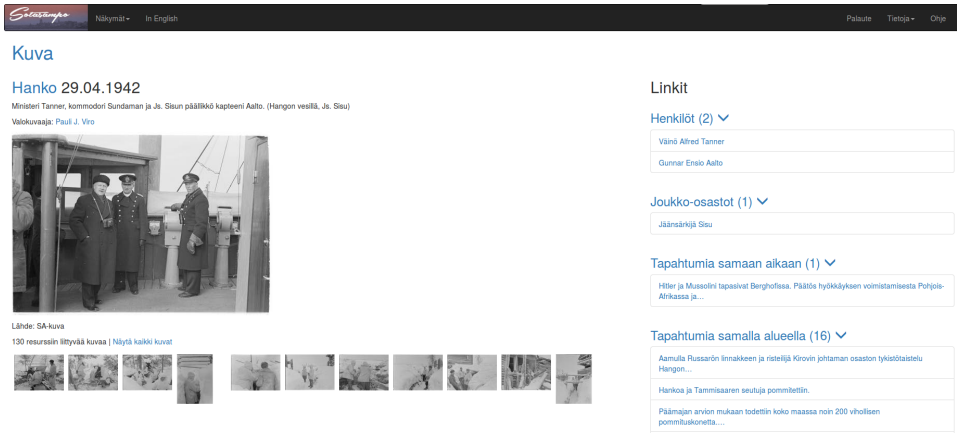
esimerkiksi tapahtumia samalla alueella.

SA-kuvapalvelussa valokuvien tekstit on tallennettu siten, että tekstin sanat löytyvät myös perusmuodossa [Puo]. Tämän takia hakusanalla “sotamies” löytyy kaikki kuvat, joissa mainitaan sana “sotamies” missä tahansa sijamuodossa. Koska tässä työssä toteutetun linkitetyn datan julkaisun tekstejä ei ole indeksoitu tällä tavoin, toteutetussa sovelluksessa hakusanalla “sotamies” löytyy vain sellaiset kuvat, joiden tekstissä sana on perusmuodossa. Kaikkien sijamuotojen löytämiseksi käyttäjä voi kuitenkin käyttää tähtimerkkiä (*): “sotamie*”.

Kuten tapahtumille myös valokuville muodostetaan omat tietosivunsa. Sivulla on suosituslinkkejä tapahtumiin samaan tapaan kuin tapahtumien sivulla valokuvan ajan ja paikan perusteella. Lisäksi käyttäjälle suositellaan muita valokuvia, jotka liittyvät kyseiseen valokuvaan. Erään valokuvan tietosivu on nähtävissä kuvassa 22.

11 Tulosten arviointi

Tässä luvussa käydään läpi tutkimuskysymykset ja niihin saadut vastaukset.



Kuva 22: Kuvankaappaus valokuvan tietosivusta.

11.1 Sotahistoriallisten aineistojen mallintaminen

Tapahtumia ja valokuvia toisesta maailmansodasta mallinnettiin linkitettynä datana. Mallintamisessa pyrittiin vastaamaan tutkimuskysymykseen:

Miten sotahistoriallisia aineistoja kannattaa mallintaa linkitettynä datana?

CIDOC CRM tarjoaa semanttisesti rikkaan mallin historiallisten aineistojen kuvaamiseen. Tapahtumaperustainen malli mahdollistaa erilaisten aineistojen harmonisoinnin, jolloin ne kirjavuudestaan huolimatta ovat yhteensopivia. Mallin avulla historiallisten aineistojen ajalliset muutokset voidaan kuvata tapahtumien kautta, mikä Sotasammossa on tärkeässä roolissa esimerkiksi henkilöiden sotilasarvojen osalta. Tapahtumat myös sitovat eri entiteetit ajanhetkiin ja paikkoihin, jolloin entiteetit linkittyvät toisiinsa luontevasti näiden kautta. Aineistojen yhteyksistä voidaan päätellä uutta tietoa käyttäen linkkejä muihin resursseihin ja mallissa määritettyjen ominaisuuksien semantiikkaa. Esimerkiksi henkilön osallisuudesta tapahtumaan voidaan päätellä henkilön sijainti tietyllä ajanhetkellä. Tällaisen datan analysoinnin lisäksi malli soveltuu myös sovelluskehityksen käyttöön. Tätä päätelmää tukevat esimerkiksi tässä työssä toteutetut sovellukset, jotka käyttävät mallin mukaisesti julkaistua dataa.

11.2 Aineistojen yhdistäminen muihin aineistoihin

Eri aineistojen yhdistämiseen liittyi kaksi tutkimuskysymystä, jotka käydään läpi seuraavassa.

Miten käytössä olevat aineistot kannattaa yhdistää muihin aineistoihin?

Aineistojen – erityisesti valokuva-aineiston – koosta johtuen aineistojen linkittäminen käsin ei ollut käytännössä mahdollista. Linkitystä varten toteutettiin linkityskirjasto, ja tämän avulla linkitysohjelmat kutakin kohdeontologiaaa varten. Automaattinen nimettyjen entiteettien linkitys vaati oman yksilöintiheuristiikan kehittämisen kullekin entiteettityypille. Koska ontologiat ovat saatavilla SPARQL-palvelupisteen kautta, voitiin tietoa kysellä monipuolisesti ja käyttää hyväksi resursseihin liittyviä tietoja. Erikoistuneiden heuristiikkojen ja kohdeontologioiden semanttisen rikkauden vuoksi linkitys onnistui hyvin kaikkien linkitettyjen entiteettien kohdalla: kaikissa linkityksissä päästiin satunnaisotannan perusteella yli 75% tuloksiin tarkkuudessa ja saannissa. Tulokset olivat huomattavasti paremmat verrattuna yksinkertaiseen linkitykseen, jossa tarkempaa yksilöintiä ei tehty.

Mitä lisäarvoa aineistojen yhdistämisestä muihin aineistoihin saavutetaan?

Yhdistämällä aineisto toisiin aineistoihin tuodaan aineistojen tietosisältö yhteen. Näin voidaan aineiston sisällöstä tarjota lisätietoa muiden aineistojen avulla, ja yhteyksistä voidaan päätellä uutta tietoa. Esimerkiksi, jos henkilö on yhdistetty tapahtumaan ja tämä paikkaan, voidaan päätellä, missä henkilö on ollut kyseisen tapahtuman aikaan. Näin linkitys mahdollistaa aineistojen tehokkaamman tutkimisen. Ilman eksplisiittisiä linkkejä tietoa jouduttaisiin hakemaan eri lähteistä, ja nimien synonymia ja toisaalta homonymia vaikeuttaisivat tiedon löytämistä. Koska Sotasammon aineistot on yhdistetty linkitettynä data, voidaan yhdellä SPARQL-kyselyllä selvittää esimerkiksi henkilön dokumentoidut liikkeet sotien aikana. Sotasammon henkilösovelluksessa[40] hyödynnetään muodostettuja linkkejä muodostamalla kullekin henkilölle tämän liikkeitä kuvaava aikajana. Tämä tarjoaa käyttäjille mahdollisuuden tutkia esimerkiksi sukulaisensa vaiheita sodassa.

Työn lähtökohtana oli erillisissä siiloissa, kuten tietokirjoissa, taulukoissa ja tietokannoissa, sijaitsevaa tietoa, jolloin lisätietoa näistä tietokohteista oli haettava eri lähteistä. Sotasammossa aineistot on yhdistetty paitsi toisiinsa myös muihin olennaisiin aineistoihin. Näin lisätietoa ei tarvitse etsiä muualta vaan se löytyy kunkin resurssin yhteydestä. Entiteettien vahva yksilöiminen mahdollistaa linkitettyjen aineistojen muodostaman kokonaisuuden selaamisen linkkien kautta.

[40] `http://www.sotasampo.fi/fi/persons`

11.3 Aineistojen visualisointi

Aineistoille toteutettiin sovellukset osana Sotasampoa. Sovellukset visualisoivat mallinnettuja aineistoja ja tarjoavat loppukäyttäjille tavan tutkia aineistoja osana suurempaa kokonaisuutta. Sovellusten toteutusten kautta vastataan viimeiseen tutkimuskysymykseen:

Miten mallinnetut aineistot kannattaa visualisoida?

Tapahtumia varten toteutetussa sovelluksessa käyttäjälle tarjotaan sodan kulun hahmottamista helpottava konteksti aikajanan ja kartan muodossa. Sovellus tarjoaa myös lisää tietoa tapahtumiin liittyvien kuvien avulla. Näiden lisäksi kartalla näkyvä sotilaiden kuolleisuus tarjoaa lisätietoa siitä, missä taisteluita on käyty kullakin ajanhetkellä.

Ongelmana sovelluksessa on käytetyn aikajanakirjaston ikä, mikä näkyy puutteellisena mobiililaitetukena. Aikajanan vierittäminen esimerkiksi älypuhelimella on hieman hankalaa. Lisäksi käytettävyyttä voisi parantaa lisäämällä mahdollisuuden suodattaa tapahtumia tapahtuman tyypin perusteella.

Valokuvia visualisoivan sovelluksen käyttötapaukseksi määritettiin valokuvien tutkiminen hakukäyttöliittymän avulla. Toteutettu sovellus tarjoaa käyttäjälle mahdollisuuden selata valokuvia, hakea niitä vapaan tekstihaun avulla ja rajata tuloksia fasettien avulla. Koska valokuvat liittyvät vahvasti muihin Sotasammon aineistoihin, näytetään kunkin resurssin tietosivulla siihen liittyvät valokuvat. Olemassa olevaan SA-kuvapalvelun valokuvasovellukseen verrattuna toteutettu sovellus tarjoaa rikkaamman tavan hakea valokuvia erilaisten valokuviin liittyvien tietojen perusteella.

Aineistojen linkitys toisiinsa mahdollistaa vaivattoman vaeltelun aineistojen välillä: esimerkiksi tapahtumasta voidaan siirtyä tarkastelemaan siihen liittyvää henkilöä; henkilön tiedoista voidaan siirtyä valokuviin, näiden tiedoista paikkoihin ja niin edelleen. Toteutetut sovellukset tukevat myös itse datajulkaisun hyödyllisyyttä ja suunnitellun mallin toimivuutta sovelluskehityksessä.

12 Yhteenveto

Tässä työssä toteutettiin Suomen talvi- ja jatkosodan tapahtuma- ja valokuvaontologiat osana Sotasampo-projektia. Aineistot mallinnettiin CIDOC CRM -standardin mukaisesti, jolloin ne ovat yhdenmukaisia paitsi toistensa myös muiden Sotasammon aineistojen kanssa.

Aineistoja rikastettiin yhdistämällä ne automaattisesti toisiin aineistoihin: paikkoihin, henkilöihin ja joukko-osastoihin. Koska eri aineistossa esiintyviin entiteetteihin viitattiin aineistoissa luonnollisen tekstin lomassa, oli apuna käytettävä kieliteknologisia työkaluja. Lisäksi automaattisessa linkityksessä

ongelmaksi nousi erityisesti tekstissä esiintyvien mainintojen monimerkityksisyys: esimerkiksi henkilöiden nimet ovat monitulkintaisia eri henkilöiden kesken, ja toisaalta ne sekoittuvat myös paikannimiin.

Linkitystä varten toteutettiin *ARPA Linker*-kirjasto, jonka avulla RDF-muotoisia aineistoja voidaan automaattisesti linkittää toisiin. Entiteettien yksilöimiseksi kehitettiin heuristiikkoja, joiden avulla monitulkintaisten entiteettivittausten tuottamista kandidaateista voitiin valita todennäköisimmin tarkoitetut entiteetit. Erilaiset entiteetit vaativat omat heuristiikkansa, jotta linkkien laatu saatiin riittävän hyväksi. Näin tuotettujen linkkien laatu arvioitiin mittaamalla tarkkuutta ja saantia käyttäen satunnaisia otantoja aineistoista. Toteutetut yksilöinnit paransivat linkkien laatua huomattavasti verrattuna verrokkilinkitykseen, jossa yksilöintiä ei tehty.

Julkaistujen aineistojen ja niiden välisten linkkien kautta voidaan löytää uutta tietoa esimerkiksi yksittäisten henkilöiden liikkeistä sotien aikana. Tämän työn tuloksena julkaistu tieto sodanajan tapahtumista ja näihin liittyvistä lisätiedoista oli lähtökohtaisesti ripoteltuna eri lähdeaineistoihin. Lisätietoa esimerkiksi tapahtumassa mukana olleista henkilöistä tai tapahtuman paikasta ei välttämättä ollut saatavissa samasta lähteestä.

Datajulkaisu tarjoaa aineistot koneluettavassa muodossa ja mahdollistaa uusien sovellusten rakentamisen dataan perustuen. Koska aineistot on julkaistu SPARQL-palvelupisteen avulla, voi niihin kohdistaa vapaasti esimerkiksi historiantutkijaa kiinnostavia kyselyitä, ja dataa voidaan näin analysoida.

Julkaistujen aineistojen hyödyntämiseksi toteutettiin kaksi sovellusta. Tapahtumasovellus esittää loppukäyttäjälle sodan kulun tarjoten lisätietoa tuotettujen linkkien avulla. Valokuvasovellus tarjoaa mahdollisuuden tutkia rintamamiesten ottamia kuvia semanttisen haun avulla. Toteutetut sovellukset osoittavat, että työssä suunniteltua datamallia ja -julkaisua voidaan hyödyntää sovelluskehityksessä. Sovellukset julkaistiin osana Sotasampo-portaalia, jossa ne yhdistyvät paitsi toisiinsa myös muihin sovelluksiin aineistojen välisten linkkien kautta. Portaalin sovellukset tarjoavat loppukäyttäjille keskitetyn lähteen Suomen vaiheista toisen maailmansodan aikaan.

Tämän työn tuloksia on esitelty myös Aalto-yliopiston Semanttisen laskennan tutkimusryhmässä kirjoitetuissa kansainvälisissä tutkimusartikkeleissa [HTM+17, HTM+15, HHL+16, KHH16a, KHH+16b].

Lähteet

[BHBL09] Bizer, Christian, Heath, Tom ja Berners-Lee, Tim: *Linked data-the story so far*. Semantic Services, Interoperability and Web Applications: Emerging Concepts, sivut 205–227, 2009.

[Biz09] Bizer, C.: *The Emerging Web of Linked Data*. IEEE Intelligent Systems, 24(5):87–92, Sept 2009, ISSN 1541-1672.

[BL06] Berners-Lee, Tim: *Linked data – design issues*. 2006. `http://www.w3.org/DesignIssues/LinkedData.html`, vierailtu 20.6.2016 .

[BP06] Bunescu, Razvan C ja Pasca, Marius: *Using Encyclopedic Knowledge for Named Entity Disambiguation.* Teoksessa *Proceedings of EACL 2006, the 11th Conference of the European Chapter of the Association for Computational Linguistics*, sivut 9–16, 2006.

[CDG+15] Crofts, Nick, Doerr, Martin, Gill, Tony, Stead, Stephen ja Stiff, Matthew: *Definition of the CIDOC Conceptual Reference Model*. 2015. `http://www.cidoc-crm.org/docs/cidoc_crm_version_6.2.1.pdf`.

[CFTW13] Clark, Kendall, Feigenbaum, Lee, Torres, Elias ja Williams, Gregory: *SPARQL 1.1 Protocol*. W3C Recommendation, W3C, maaliskuu 2013. `http://www.w3.org/TR/2013/REC-sparql11-protocol-20130321/`, vierailtu 17.11.2016 .

[CID11] CIDOC CRM Special Interest Group: *How To Implement CRM Time in RDF*, 2011. `http://www.cidoc-crm.org/docs/How_to%20implement%20CRM_Time_in%20RDF.pdf`.

[Cuc07] Cucerzan, Silviu: *Large-Scale Named Entity Disambiguation Based on Wikipedia Data*. Teoksessa *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, sivut 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Doe03] Doerr, Martin: *The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata*. AI magazine, 24(3):75, 2003.

[Goo16] Google: *What Is Angular?*, 2016. `https://docs.angularjs.org/guide/introduction`, vierailtu 4.11.2016 .

[GOS09] Guarino, Nicola, Oberle, Daniel ja Staab, Steffen: *What Is an Ontology?* Teoksessa Staab, Steffen ja Studer, Rudi (toimittajat): *Handbook on Ontologies*, sivut 1–17. Springer Berlin Heidelberg, 2009, ISBN 978-3-540-92673-3.

[GTB+10] Grover, Claire, Tobin, Richard, Byrne, Kate, Woollard, Matthew, Reid, James, Dunn, Stuart ja Ball, Julian: *Use of the Edinburgh geoparser for georeferencing digitized historical collections.* Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368(1925):3875–3889, 2010, ISSN 1364-503X. `http://rsta.royalsocietypublishing.org/content/368/1925/3875`.

[HB11] Heath, Tom ja Bizer, Christian: *Linked data: Evolving the web into a global data space.* Synthesis lectures on the semantic web: theory and technology, 1(1):1–136, 2011.

[HHL+16] Hyvönen, Eero, Heino, Erkki, Leskinen, Petri, Ikkala, Esko, Koho, Mikko, Tamper, Minna, Tuominen, Jouni ja Mäkelä, Eetu: *WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History.* Teoksessa Harald Sack, Eva Blomqvist, Mathieu d'Aquin Chiara Ghidini Simone Paolo Ponzetto Christoph Lange (toimittaja): *The Semantic Web – Latest Advances and New Domains (ESWC 2016).* Springer-Verlag, May 2016.

[HIT16] Hyvönen, Eero, Ikkala, Esko ja Tuominen, Jouni: *Linked Data Brokering Service for Historical Places and Maps.* Teoksessa *Proceedings of the 1st Workshop on Humanities in the Semantic Web - WHiSe*, May 2016.

[HL95] Hürsch, Walter L ja Lopes, Cristina Videira: *Separation of Concerns.* 1995.

[HL15] Hienert, Daniel ja Luciano, Francesco: *Extraction of Historical Events from Wikipedia.* Teoksessa Simperl, Elena, Norton, Barry, Mladenic, Dunja, Della Valle, Emanuele, Fundulaki, Irini, Passant, Alexandre ja Troncy, Raphaël (toimittajat): *The Semantic Web: ESWC 2012 Satellite Events*, nide 7540 sarjassa *Lecture Notes in Computer Science*, sivut 16–28. Springer Berlin Heidelberg, 2015, ISBN 978-3-662-46640-7. `http://dx.doi.org/10.1007/978-3-662-46641-4_2`.

[HMPR04] Hevner, Alan R, March, Salvatore T, Park, Jinsoo ja Ram, Sudha: *Design science in information systems research.* MIS quarterly, 28(1):75–105, 2004.

[HRN+13] Hachey, Ben, Radford, Will, Nothman, Joel, Honnibal, Matthew ja Curran, James R.: *Evaluating entity linking with wikipedia.* Artificial Intelligence, 194:130–150, 2013, ISSN 0004-3702. http://www.sciencedirect.com/science/article/pii/S0004370212000446.

[HS11] Han, Xianpei ja Sun, Le: *A Generative Entity-mention Model for Linking Entities with Knowledge Base.* Teoksessa *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, sivut 945–954, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics, ISBN 978-1-932432-87-9. http://dl.acm.org/citation.cfm?id=2002472.2002592.

[HSZ11] Han, Xianpei, Sun, Le ja Zhao, Jun: *Collective Entity Linking in Web Text: A Graph-based Method.* Teoksessa *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, sivut 765–774, New York, NY, USA, 2011. ACM, ISBN 978-1-4503-0757-4. http://doi.acm.org/10.1145/2009916.2010019.

[HTAM14] Hyvönen, Eero, Tuominen, Jouni, Alonen, Miika ja Mäkelä, Eetu: *Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets.* Teoksessa *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers*, sivut 226–230, May 2014.

[HTM+15] Hyvönen, Eero, Tuominen, Jouni, Mäkelä, Eetu, Dutruit, Jérémie, Apajalahti, Kasper, Heino, Erkki, Leskinen, Petri ja Ikkala, Esko: *Second World War on the Semantic Web: The WarSampo Project and Semantic Portal.* Teoksessa *Proceedings of the ISWC 2015 Posters & Demonstrations Track.* CEUR-WS Proceedings, October 2015. http://www.ceur-ws.org/Vol-1486, Vol 1486.

[HTM+17] Heino, Erkki, Tamper, Minna, Mäkelä, Eetu, Leskinen, Petri, Ikkala, Esko, Tuominen, Jouni, Koho, Mikko ja Hyvönen, Eero: *Named Entity Linking in a Complex Domain: Case Second World War History.* Teoksessa *Proceedings, Language, Technology and Knowledge 2017. June 19-20, Galway, Ireland.* Springer-Verlag, February 2017. http://ldk2017.org/, Accepted.

[Ikk16] Ikkala, Esko: *Suomalainen historiallisten paikkojen ja karttojen ontologiapalvelu.* Pro Gradu -työ, Aalto-yliopisto, sähkötekniikan korkeakoulu, August 2016.

[KHH16a] Koho, Mikko, Heino, Erkki ja Hyvönen, Eero: *SPARQL Facet-er – Client-side Faceted Search Based on SPARQL.* Teoksessa

Proceedings of the ESWC Developers Workshop 2016. CEUR Workshop Proceedings, May 2016.

[KHH+16b] Koho, Mikko, Hyvönen, Eero, Heino, Erkki, Tuominen, Jouni, Leskinen, Petri ja Mäkelä, Eetu: *Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data*. Teoksessa *Proceedings of the 1st Workshop on Humanities in the Semantic Web - WHiSe*, May 2016.

[LAH+11] Lindén, Krister, Axelson, Erik, Hardwick, Sam, Silfverberg, Miikka ja Pirinen, Tommi: *HFST–Framework for Compiling and Applying Morphologies*. Teoksessa *Proceedings of Second International Workshop on Systems and Frameworks for Computational Morphology*, sivut 67–85, 2011. `http://www.helsinki.fi/%7Etapirine/publications/Pirinen-sfcm-2011.pdf`.

[Les16] Leskinen, Petri: *Sotilashenkilöiden ja joukko-osastojen mallintaminen ja käyttö toimijaontologiana*. Pro Gradu -työ, Aalto-yliopisto, perustieteiden korkeakoulu, Dec 2016.

[LJ05] Leskinen, Jari ja Juutilainen, Antti (toimittajat): *Jatkosodan pikkujättiläinen*. WSOY, Finland, 2005.

[LJ06] Leskinen, Jari ja Juutilainen, Antti (toimittajat): *Talvisodan pikkujättiläinen*. WSOY, Finland, neljäs painos, 2006.

[LME12] Lin, Thomas, Mausam ja Etzioni, Oren: *Entity Linking at Web Scale*. Teoksessa *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, sivut 84–88, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. `http://dl.acm.org/citation.cfm?id=2391200.2391216`.

[LWH+13] Li, Yang, Wang, Chi, Han, Fangqiu, Han, Jiawei, Roth, Dan ja Yan, Xifeng: *Mining Evidences for Named Entity Disambiguation*. Teoksessa *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, sivut 1070–1078, New York, NY, USA, 2013. ACM, ISBN 978-1-4503-2174-7. `http://doi.acm.org/10.1145/2487575.2487681`.

[Man] Mannerheim-ristin ritarien säätiö: *Mannerheim-ristin ritarit – Karl Lennart Oesch*. `http://www.mannerheim-ristinritarit.fi/ritarit?xmid=111`, vierailtu 3.10.2016 .

[MTLH16] Mäkelä, Eetu, Törnroos, Juha, Lindquist, Thea ja Hyvönen, Eero: *WW1LOD – An application of CIDOC-CRM to World War 1 Linked Data*. International Journal on Digital Libraries, 2016.

[Mä14] Mäkelä, Eetu: *Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text*. Teoksessa *Proceedings of the ESWC 2014 demonstration track, Springer-Verlag*, May 2014.

[NDO05] Nagypál, Gábor, Deswarte, Richard ja Oosthoek, Jan: *Applying the Semantic Web: The VICODI Experience in Creating Visual Contextualization for History*. Literary and Linguistic Computing, 20(3):327–349, 2005. `http://llc.oxfordjournals.org/content/20/3/327.abstract`.

[NS07] Nadeau, David ja Sekine, Satoshi: *A survey of named entity recognition and classification*. Lingvisticæ Investigationes, 30(1):3–26, 2007.

[Pir15] Pirinen, Tommi A.: *Omorfi—Free and open source morphological lexical database for Finnish*. Teoksessa *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, numero 109, sivut 313–315. Linköping University Electronic Press, 2015.

[Puo] Puolustusvoimien kuvakeskus: *SA-kuva-arkisto*. `http://sa-kuva.fi/`, vierailtu 29.11.2016 .

[RA07] Raimond, Yves ja Abdallah, Samer: *The Event Ontology*, lokakuu 2007. `http://motools.sf.net/event/event.122.html`, vierailtu 5.7.2016 .

[Sha10] Shaw, Ryan: *LODE: An ontology for Linking Open Descriptions of Events*, 2010. `http://linkedevents.org/ontology/2010-10-07/`, vierailtu 5.7.2016 .

[SHBL06] Shadbolt, Nigel, Hall, Wendy ja Berners-Lee, Tim: *The semantic web revisited*. Intelligent Systems, IEEE, 21(3):96–101, 2006.

[suo83] *Suomi Sodassa – Talvi- ja jatkosodan tärkeät päivät*. Valitut Palat, toinen painos, 1983, ISBN 9519078940.

[The11] The British Museum: *The British Museum is the first UK arts organisation to publish its collection semantically*, 2011. `http://www.britishmuseum.org/about_us/news_and_press/press_releases/2011/semantic_web_endpoint.aspx`, vierailtu 13.7.2016 .

[Tun09] Tunkelang, Daniel: *Faceted search*, nide 1 sarjassa *Synthesis lectures on information concepts, retrieval, and services*. Morgan & Claypool Publishers, 2009.

[VK14] Vrandečić, Denny ja Krötzsch, Markus: *Wikidata: A Free Collaborative Knowledgebase*. Commun. ACM, 57(10):78–85, syyskuu 2014, ISSN 0001-0782. http://doi.acm.org/10.1145/2629489.

[vMS+09] van Hage, Willem, Malaisé, Véronique, Segers, Roxane, Hollink, Laura ja Schreiber, Guus: *The Simple Event Model Ontology*, 2009. http://semanticweb.cs.vu.nl/2009/11/sem/, vierailtu 5.7.2016 .

[vMS+11] van Hage, Willem Robert, Malaisé, Véronique, Segers, Roxane, Hollink, Laura ja Schreiber, Guus: *Design and use of the Simple Event Model (SEM)*. Web Semantics: Science, Services and Agents on the World Wide Web, 9(2):128 – 136, 2011, ISSN 1570-8268. http://www.sciencedirect.com/science/article/pii/S1570826811000199, Provenance in the Semantic Web.

[Wil11] Williams, Hugh: *Re: DBpedia: limit of triples*, 2011. http://lists.w3.org/Archives/Public/public-lod/2011Aug/0028.html, vierailtu 17.11.2016 .

[WLC14] Wood, David, Lanthaler, Markus ja Cyganiak, Richard: *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation, W3C, helmikuu 2014. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, vierailtu 26.10.2015 .