Harri Kiiskinen
FT, erikoistutkija, kulttuurihistoria, Turun yliopisto

Semanttinen metadata MoMaF-hankkeen aineistonhallinnassa

Tässä katsauksessa tarkastelen *Movie Making Finland: Finnish fiction films as big data, 1907–2017* (MoMaF) -hankkeen yhteydessä tekemääni metadatatyötä. Lähtökohtaisesti metadata (tai metatieto) on tietoa datasta. Yhden metadata voi kuitenkin yhtä hyvin olla toisen dataa, kuten vaikkapa elokuvien näyttelijä- ja esitystiedot, jotka itse elokuviin nähden ovat metadataa mutta elokuvia tutkivalle voivat olla keskeinen osa tutkimusaineistoa. Tässä ”metadata” tarkoittaa tietoa, joka kuvaa tutkijalle elokuvia kuvailevan aineiston ominaisuuksia. Näin metatietoa voisi olla vaikkapa tieto siitä, mistä elokuvien metadatatiedot Kansallisen audiovisuaalisen instituutin Elonet-tietokannasta löytyvät. Tässä tapauksessa se, mitä arkistokäytössä kutsutaan elokuvien metadataksi, onkin osa tutkimusaineistoa ja siten muuttunut dataksi. Metadata ei siis ole tiedon absoluuttinen vaan kontekstisidonnainen määre, jonka merkitys riippuu tiedon käyttäjästä ja käyttötarkoituksesta.

Toisaalta voidaan yhtä hyvin pitäytyä jaottelussa, jossa varsinainen data ja metadata erotetaan toisistaan. Näin on tehty myös MoMaF-hankkeessa, jossa varsinaista dataa ovat elokuvat ja erityisesti digitaaliset elokuvatiedostot. Näiden digitaalisten tiedostojen sisältöä kuvaava aineisto on määritelty metadataksi riippumatta sen lähteestä ja muodosta. MoMaF-hankkeessa metadatan käsittely tarkoittaa kaikkea digitaalisia elokuvatiedostoja kuvailevan aineiston käsittelyä.

Aloitan käsittelemällä sitä, minkälaisia metadata-aineistoja olen hankkeessa käsitellyt sekä sitä, miten ne suhtautuvat itse dataan. Sen jälkeen esittelen järjestelmiä ja teknisiä ratkaisuja, joita hankkeessa on aineiston käsittelyssä hyödynnetty. Lopuksi pohdin vielä tehtyjen valintojen ja ratkaisujen toimivuutta ja soveltuvuutta.

Metadata-aineistot

Hankkeen ulkoiset aineistot

Hankkeen ulkopuolisista lähteistä kerätyt aineistot ovat lähtökohtaisesti vain luettavissa olevaa aineistoa. Ulkopuolisista lähteistä kerättyä aineistoa olen käsitellyt lähinnä raaka-aineena omien, pidemmälle jalostettujen aineistojen synnyttämisessä. Luettelen seuraavaksi lyhyesti aineistot, joiden käsittelyn tarkempaan kuvaukseen palaan tekstissä tuonnempana.

Ulkopuolisista lähteistä MoMaF-hankkeeseen on kerätty aineistoa Elonetistä, Internet Movie Databasesta (IMDb) ja Wikidata-tietokannasta.

Elonet

Suomalaista elokuvaa koskevan tiedon pääasiallisena lähteenä toimi Kansallisen audiovisuaalisen arkiston tuottama kansallisfilmografia, joka löytyy digitaalisessa

muodossa Elonet-tietokannasta. Kyseinen verkkotietokanta tarjoaa kunkin elokuvan kuvailutiedot, jotka ladattiin hankkeen käyttöön yksinkertaisen lukijaohjelman avulla.

IMDb

Kansainvälisestä IMDb-elokuvatietokannasta on hankkeessa ladattu lähinnä täydentävää elokuviin liittyvää metadataa, kuten näyttelijöiden valokuvia, joita on käytetty apuna kasvojentunnistuksessa.

Wikidata

Wikidata-tietokanta on tarjonnut täydentävää tietoa esimerkiksi näyttelijöistä. Suuresta määrästä Elonet-tietokantaan tallennetuista näyttelijöistä löytyy myös Wikidata-tietue, jonka tietoa on käytetty hyväksi muun muassa näyttelijöiden sukupuolien määrittämisessä.

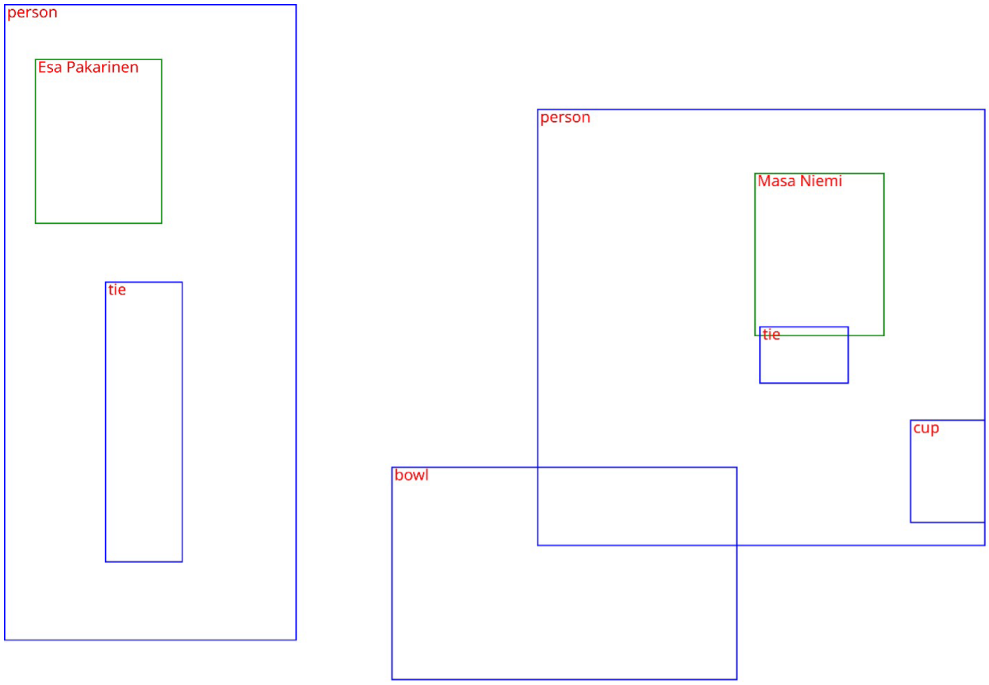
Itse tuotetut aineistot

MoMaF-hankkeessa on tuotettu kahdenlaista metadataa: ulkoisten tietolähteiden perusteella elokuvia kuvaavaa aineistoa sekä projektissa kehitettyjen analyysimenetelmien tuottamaa tulosaineistoa. Lähtökohtana olivat Elonet-tietokannasta ladatut tiedostot, jotka muunnettiin ohjelmallisesti semanttiseksi verkoksi. Näin Elonetin kansallisfilmografia-aineiston perusteella rakennettiin sekä elokuvista että näyttelijöistä semanttinen tietokanta.[1] Tietoa täydennettiin tarpeen mukaan hakemalla lisää aineistoa sekä IMDb-tietokannasta että Wikidatasta. Kaikki aineistot tallennettiin samaan kantaan.

Elokuvia ja näyttelijöitä koskevan metatietoaineiston lisäksi projektissa on tuotettu erilaisia tunnistusaineistoja, jotka on myös koodattu semanttisesti ja tallennettu samoihin tietojärjestelmiin. Näiden pohjaksi luotiin tietomalli ajallisten annotaatioiden tekemiseksi sekä kasvojen- ja objektintunnistuksessa hyödynnetty tietomalli, jolla kunkin objektin ruudulla rajaava kehys saadaan tallennettua siten, että tunnistukseen liittyvän annotaation sijaintia ruudulla on mahdollista seurata elokuvan ruudusta toiseen. Hankkeessa *annotaatio* tarkoittaa siis elokuvaan liittyvää tietoa, joka kuvaa jotain määritettyä osaa elokuvan audiovisuaalisesta kokonaisuudesta ja kestosta. Annotaatio voi osoittaa esimerkiksi sen, että ”Tauno Palon kasvot esiintyvät ruudulla ajankohdasta 1.12.34,5 ajankohtaan 1.12.52,2 alueella, joka ulottuu vaakasuunnassa 10 % leveydestä 62 % leveyteen ja pystysuunnassa 12 % korkeudesta 78 % korkeuteen”, tai sen, että ”ääniraidalla soi trumpetti 3,4 sekunnin ajan alkaen ajankohdasta 45.3,4”.

Kuvassa 1 on esimerkki annotaatioiden visuaalisesta kuvauksesta elokuvasta *Pekka ja Pätkä pahassa pulassa* (1955). Selvyyden vuoksi kuvasta on jätetty pois elokuvan kuvarajausta osoittava ruutu. Sininen kehys osoittaa objektintunnistuksen tuloksia ja vihreä kasvojentunnistusta.

1 ”Semanttinen tieto”, ”semanttinen verkko”, ”semanttinen tietokanta” ovat kaikki toisiinsa liittyviä, lähes samaa tarkoittavia käsitteitä, joiden taustalla on ajatus tiedon järjestämisestä siten, että tiedon muodostavat pienet yksiköt liittyvät toisiinsa. Näitä liitoksia kuvataan ominaisuuksilla, jotka antavat liitoksille merkityksiä. Esimerkiksi henkilöön liitetty päivämäärä voi tarkoittaa syntymäpäivää mutta yhtä hyvin myös vaikka työsuhteen alkamispäivää tai päivää, jona hän osti itselleen auton. Toisaalta voimme määritellä esimerkiksi yleisen tietotyypin ”tapahtuma”, johon liittyy aina tapahtuman päivämäärä. Tapahtumalla voi olla eri alaluokkia tapahtuman tyypin mukaan, vaikkapa juuri syntymä, auton osto tai työsuhteen alkaminen. Eri tapoja jäsentää tietoa semanttisesti on lukuisia, mutta keskeistä on, että tietoon itseensä liitetään sitä kuvailevia ominaisuuksia siten, että tietojoukko, tai semanttinen verkko, tavallaan selittää itse itsensä.



Kuva 1. Esimerkki annotaatioiden visuaalisesta kuvauksesta elokuvassa *Pekka ja Pätkä pahassa pulassa* (1955).

Tekniset ratkaisut

Tietomuodot

Metatietojen tallennusmenetelmiksi on monia vaihtoehtoja, joilla on keskenään hyvin erilaisia ominaisuuksia. MoMaF-hankkeessa valinnan keskeinen kriteeri oli se, että eri tietolähteistä tulevaa dataa pitää olla mahdollista integroida aiempaan dataan mahdollisimman hyvin, ja että hankkeen alkuvaiheessa ei vielä ole tarkkaa käsitystä siitä, missä muodossa ja minkälaista hankkeen aikana tuotettu data on luonteeltaan. Lisäksi oli varauduttava siihen, että dataa tuotetaan suuria määriä ja että järjestelmän tulisi kyetä integroimaan todennäköisesti vähintään satoja miljoonia tietoyksiköitä. *Tietoyksikkö* tarkoittaa tässä yhteydessä pienintä yksittäistä tietoa, joka toisaalta saattaa olla elokuvan koko sisältöseloste, toisaalta yksittäiseen ruutuun liittyvän visuaalisen annotaation yhden kehyksen vasemman alakulman vaakakoordinaatti prosentuaalisena osuutena ruudun vaakamitasta. Yksittäisen tietoyksikön koko saattaa siis vaihdella yhdestä desimaaliluvusta usean kilotavun kokoiseen tekstiin.

Pelkästään alkuperäisen elokuva-aineiston tarkastelu osoitti, että aineiston tallentaminen perinteisen relaatiotietokannan muodossa olisi vaatinut pitkän työn tietokannan rakenteen suunnittelussa. Tähän on lähtökohtaisesti kaksi syytä:

1. Relaatiotietokannat perustuvat lähtökohtaisesti taulukkomuotoiseen tiedon tallentamiseen. Tällöin esimerkiksi yhtä elokuvaa varten on taulukossa yksi rivi ja jokainen palsta kuvaa erillistä elokuvaan liittyvää tietoa. Yksinkertaisten tietojen, kuten vaikkapa julkaisuvuoden tai sisältökuvauksen, hallinta on yksinkertaista, mutta se vaatii joka tapauksessa taulukon palstojen määrittämistä etukäteen. Tämä edellyttää hyvää ennakkokäsitystä tallennettavan tiedon luonteesta ja sisällöstä sekä selvää ajatusta siitä, mitä tietoa hankkeen aikana tullaan tarvitsemaan.

2. Relaatiotietokanta – osin nimensä vastaisesti – ei ole kovinkaan joustava ja sujuva tietoyksiköiden välisiä suhteita käsiteltäessä. Toisaalta elokuva-aineisto on lähes läpikotaisin pelkästään eri entiteettien (elokuvien, ihmisten, yhtiöiden) välisiä suhteita eri muodoissaan.

Lisäksi oli tietenkin otettava huomioon hankkeen keskeiset tavoitteet: tutkia suomalaisen elokuvan historiaa koneoppimisjärjestelmien avulla ja synnyttää siten uutta tietoa suomalaisesta elokuvasta. Tähän päämäärään pyrkiminen tulisi hankkeen kuluessa näkymään kahdella metadatan käsittelyyn vaikuttavalla tavalla:

1. Hankkeen kuluessa tultaisiin tuottamaan suuri määrä elokuvia kuvaavaa aineistoa, jonka muodosta ja laadusta ei hankkeen alussa vielä ole selvää käsitystä. Hankkeelle tulisi myös sen aikana suunnitella ja kehittää tallennusmuodot, jotka liittäisivät uuden tiedon jo olemassa olevaan aineistoon sujuvasti ja mielellään saman järjestelmän sisällä.
2. Hankkeen päättyessä nämä aineistot tulisi julkaista sellaisessa muodossa, joka mahdollistaisi niiden jatkokäytön. Koska voidaan olettaa, ettei hankkeen lopussa kuitenkaan olisi käytettävissä suuria resursseja aineiston järjestämiseen, olisi tärkeätä miettiä koko metadatan hallinta siten, että siitä olisi mahdollisimman pienellä vaivalla saatavissa ote jossain avoimessa, yleisesti luettavassa muodossa, jotta tämän otteen voisi tallentaa soveltuvaan pitkäaikaissäilytykseen.

Selvitystyö johti nopeasti siihen tulokseen, että RDF-muotoinen semanttinen data voisi toimia parhaiten tämän tyyppisessä hankkeessa. RDF-data[2], tai semanttinen data, rakentuu yksinkertaisista subjekti–predikaatti–objekti-lauseista tai kolmikoista, joilla, kuten luonnollisissakin kielissä, subjektiin liitetään ominaisuus tavalla, jota kuvaa predikaatti. Tällaisia väitelauseita tai toteamuksia yhdistämällä voidaan kuvata monimutkaisiakin asiakokonaisuuksia yksinkertaisten väitelauseiden yhdistelmänä.[3]

Jos vertaamme tätä tiedon esittämisen tapaa esimerkiksi taulukkoon, voimme kuvata asiaa seuraavasti: subjekti vastaa taulukon riviä, jokainen predikaatti vastaa taulukon yhtä palstaa ja jokaisen solun sisältö vastaa yhtä objektia.

Näiden väitekolmikoiden lisäksi RDF määrittelee käsitteen ”graafi”, jolla kuvataan tällaisten väitekolmikoiden joukkoa. Jälleen vertailukohtana taulukkomuotoa käyttäen, ”graafi” voisi tarkoittaa yhtä kokonaista taulukkoa. Lähtökohtaisesti graafi kuvaa siis toisiinsa jollain tavalla liittyvien väitekolmikoiden joukkoa.

RDF-muotoinen data ei itsessään ole lähtökohtaisesti semanttista, mutta siitä saadaan tällaista hyvin yksinkertaisesti. Sekä subjektin, predikaatin että objektin asemassa RDF-datassa voi esiintyä IRI.[4] Semanttinen web taas pohjautuu juuri URI/IRI-muotoisten käsitteiden käytölle; tällaisista *Lähikuvan* lukijoille tutuin lienee *Dublin*

2 RDF on lyhenne sanoista *Resource Description Framework*. Sillä tarkoitetaan käsitteellistä tiedon esittämisen välineistöä erityisesti verkossa. RDF ei itsessään ole varsinaisesti kieli vaan abstrakti tietomalli, jonka pohjalle on kehitetty kieliä ja määritelmiä. Ks. https://www.w3.org/TR/rdf11-concepts/.

3 Johdatuksena RDF-muotoisen datan käyttöön ks. Hyvönen 2018.

4 IRI on kuin URI mutta IRI-nimissä suurin osa Unicode-merkistön merkeistä on käytettävissä. IRI voi siis koostua hyvin vaikkapa kiinan kielen kirjoitusjärjestelmän merkeistä, kun taas URI on rajattu alkuperäiseen ASCII-merkistöön. Tästä on selvää hyötyä jo suomenkielisen aineiston käyttämisessä, sillä tietokantaan tallennettavien tunnisteiden muodostamisessa voidaan tällöin käyttää apuna myös suomenkielistä tekstiä ilman monimutkaisia muunto-operaatioita.

Core -skeema.[5] Se määrittää joukon metatietokenttiä, joiden avulla voidaan kuvata teosten metatietoja. Semanttinen datajoukko muodostuu siten, että erityisesti väitekolmikoiden predikaatteina käytetään tällaisten valmiiden skeemojen määrittämiä käsitteistöjä. Toisaalta myös subjektien ja objektien asemassa voidaan käyttää URI/IRI-muotoisia käsitteitä ja tällöin viitata asioihin yleisesti tunnetuilla ”nimillä”. Tällaisesta esimerkkinä voivat toimia esimerkiksi Yleinen suomalainen ontologia (YSO), jossa esimerkiksi ”elokuva” on <http://www.yso.fi/onto/yso/p1235>, tai Wikidata, jossa ”Helsinki” on <https://www.wikidata.org/wiki/Q1757>.

Parhaimmillaan semanttinen data tarkoittaa siis sitä, että yhdessä hankkeessa syntynyttä aineistoa kuvataan käsitteillä, jotka ovat laajalti tunnettuja. Näin hankkeessa kuvatut aineistot liittyvät näiden käsitteistöjen avulla hankkeen ulkopuolisiin entiteetteihin viittaamalla niihin yleisesti tunnetuilla nimillä. Semanttisen datan käytön perusideologian mukaista siis olisi, että tietomallissa erityisesti predikaatteina käytettäisiin jotain jo valmiiksi luotua, olemassa olevaa sanastoa. Tällöin tuotettu tieto olisi lähtökohtaisesti yhteensopivaa kaikkien samaa sanastoa käyttävien tietomassojen kanssa – ainakin ajatuksen ja tekniikan tasolla. Tätä ratkaisua harkittiin vakavasti hankkeen alkuvaiheessa, mutta valmiiden sanastojen käyttö hylättiin periaatteessa kahdesta syystä:

1. Valmiita, hyödyntämiskelpoisia sanastoja, jotka olisivat tarjonneet ratkaisuja muihin kuin aivan triviaaleihin datan kuvailun ratkaisuihin, ei onnistuttu löytämään. Lähinnä puuttumaan jäivät valmiit ratkaisut, joilla olisi voitu kuvata elokuvien ja eri tekijöiden välisiä monimutkaisia suhdeverkostoja sekä elokuviin tehtävää kuvailevaa annotointi-aineistoa. Vain osittain soveltuvien sanastojen käyttö olisi ollut rajoittavaa ja luultavasti hankaloittanut työtä huomattavasti.[6]
2. Tutkimushanke eroaa kulttuuriperintöaineiston tallentamisesta jo lähtökohtaisesti. Tutkimushankkeen on voitava käsitellä aineistoa tavalla, joka on hankkeen tarkoitusperien kannalta järkevintä, eikä sitoutua olemassa oleviin ratkaisuihin vain sen takia, että sellaisia on olemassa. Mikäli jokin hyödyllinen tietomalli olisi ollut saatavilla, sen käyttö olisi helpottanut hankkeen toteutusta. Mutta koska näin ei ollut, varmin ratkaisu oli luoda dataa varten oma tietomalli ja dokumentoida se RDFS- ja OWL-skeemojen[7] avulla.

5 *Dublin Core* -metadataskeemasta ja sen sisältämistä elementeistä ks. esim. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.

6 Elonet-tietokannasta saatava data näyttäisi ainakin XML-tiedostojen sisältämien viitteiden perusteella noudattavan EN-15907-standardia, joka taas näyttäisi olevan jonkinlainen elokuvien metatietojen vaihtamiseen kehitetty metadatasanasto ja -rakenne. Se ei kuitenkaan sisällä sanastoja ja rakenteita kuin elokuvien perustavan metadatan kuvailuun, eikä sitoutuminen tämän käyttöön siksi tuntunut tarpeelliselta ratkaisulta. Standardista lisää mm. <http://www.filmstandards.org/fsc/index.php/EN_15907>.

7 RDFS ja OWL-skeemat ovat keinoja kuvata semanttisen datan käsitteistöjä tavalla, joka dokumentoi eri käsitteiden välisiä suhteita ja mahdollistaa kuvailun perusteella yksinkertaisten loogisten johtopäätösten tekemisen. Niiden pääasiallinen tarkoitus on toimia semanttisen datan rakenteen kuvaajina, ja siten ne toimivat samalla myös dokumentaationa tallennettavan datan muodosta. Ne eivät ole skeemoja samalla tavalla kuin relaatiotietokannan skeemat, jotka määrittävät miten tietoa tulee ja on mahdollista tallentaa. Aiheesta lisää mm. Hyvönen 2018 ja Allemang & Hendler 2011.

Tallennus ja käsittely

Soveltuvaan tietokantaan tallennettuna tällaisesta tiedosta voidaan nk. SPARQL-kieltä[8] käyttäen helposti hakea esimerkiksi kaikki elokuvan *Pojat* (1962) tekemiseen osallistuneet henkilöt ja vaikka heidän syntymävuotensa. SPARQL muistuttaa varsin paljon relaatiotietokantojen hakemiseen käytettyä SQL-kieltä, mutta koska tietoaineisto ei ole taulukkoina vaan eri tietoyksiköitä yhdistävinä linkkeinä, hakulausekkeiden muotoilu on jossain määrin erilaista.

RDF-muotoisen datan tallennusalustaksi valittiin Apache Jena -tietokanta, joka on erityisesti kehitetty RDF-datan käsittelyyn, sekä tämän päälle rakennettu Apache Jena Fuseki -SPARQL-palvelin, joka mahdollistaa RDF-datan tallentamisen ja hakemisen muista ohjelmista ja koneista käsin.[9]

Tietokanta asennettiin Turun yliopiston IT-palveluiden tarjoamalle palvelimelle, ja se muodosti hankkeen metadatan käsittelyn ja tallentamisen perustan. Aineiston hakua ja muokkausta varten on käytettävissä suojattu tietokantayhteys, joka mahdollistaa aineistojen hakemisen ja tallentamisen myös palvelimen ulkopuolelta.

Elonet-tietokannan aineisto

Elokuvia koskevan metatiedon lähtökohtana oli KAVIn suomalaisen elokuvan kansallisfilmografia. Sen aineisto ladattiin suoraan Elonetistä XML-tiedostoina[10] ja muunnettiin RDF-muotoon tarkoitusta varten tehdyillä XSLT-kielisillä muunnosmalleilla.[11] RDF-aineiston tietomalli on hanketta varten kehitetty, koska valmista, suoraan tarkoitukseen sopivaa tietomallia ei ollut saatavilla.

RDF-datan eli semanttisen datan keskeinen ominaisuus on mahdollisimman pysyvien tunnisteiden käyttö. Tämä tarkoittaa sitä, että aineistoa tallennettaessa asioille annetaan nimet, joita käytetään säännönmukaisesti niihin viitatessa. Semanttisen verkon ideana olisi lisäksi niin sanottu globaali yksilöitävyys siten, että esimerkiksi henkilöön viitattaisiin tunnisteella, joka olisi sama kuin mitä käytetään vaikkapa Wikidatassa tai joissain muissa laajalti käytetyissä, avoimissa palveluissa. Tällainen yksilöitävyys on tietenkin kannatettavaa, mutta tällaista valmista aineistoa käytettäessä on otettava huomioon muitakin lähtökohtia.

Ensimmäinen ongelma syntyy entiteettien tunnistamisesta. Jos meillä on valmis tietokanta, josta löytyy henkilö nimeltä ”Mari Kuusinen”[12] ja sitten löydämme Wikidatasta henkilön nimeltä ”Mari Kuusinen”, mistä voimme tietää, että nämä ovat yksi ja sama henkilö? Tämä ei ole vaikea ongelma ratkaistavaksi, mutta ratkaisu vaatii työtä ja aikaa, ja jos tarkoituksena on, että tuotava aineisto on mahdollisim-

8 SPARQL on RDF-datan käsittelyyn kehitetty kieli. Tarkemmin SPARQL-kielestä mm. Hyvönen 2018 sekä <https://www.w3.org/TR/sparql11-query/>.

9 Apache Jena -tietokannasta ja Fuseki-SPARQL-palvelimesta lisää mm. <https://jena.apache.org/>.

10 XML (*Extensible Markup Language*) on merkintäkieli, joka soveltuu tietoyksiköiden kuvailuun. Esimerkiksi HTML on XML-kieli. Näiden ero on kuitenkin siinä, että XML ei sinänsä määritä, minkä nimisiä elementtejä merkintäkielessä voidaan käyttää. HTML taas koostuu joukosta elementtien nimistä, joiden avulla kuvataan tekstidokumentti XML-muodossa. Samoin esimerkiksi TEI on XML-muotoinen määritelmä, jonka avulla voidaan kuvata monimutkaisiakin tekstidokumentteja.

11 XSLT (*XSL Transformations*) on kieli, jonka avulla XML-tiedostoja voidaan muuntaa toisiksi XML-tiedostoiksi. Se mahdollistaa selkeiden kuvausten tekemisen lähtötiedoista kohdetiedoiksi. RDF-muotoisen datan yksi tallennusmuoto on RDF/XML, joka on siis nimensä mukaisesti XML-tyyppiä, ja XSLT-muunnosten avulla tämän tuottaminen on suhteellisen yksinkertaista.

12 Kuvitteellinen tässä yhteydessä.

man nopeasti hyödynnettävissä, ei ole mahdollista ratkaista tuhansien ihmisten ja elokuvien kohdalla tällaisia linkitysongelmia ennen aineiston käyttöönottoa. Niinpä päätimme muodostaa Elonetin metadatan aineiston pohjalta tunnisteet kullekin elokuvalle, henkilölle, ryhmälle jne. ja täydentää tätä tietoa ulkoisilla tunnisteilla vasta jälkeenpäin.

Näiden projektin sisäisten tunnisteiden luomisessa käytettiin menetelmiä, joilla samoista lähtöarvoista päästään aina samaan tunnisteeseen, eli samalle elokuvalle tai henkilölle tuli aina sama tunniste muunnosprosessin eri ajokerroilla. Elonet-tietokanta onneksi sisälsi jo valmiiksi tietokannan sisäisiä tunnisteita esimerkiksi juuri elokuville. Esimerkiksi elokuva *Tanssi yli hautojen* vuodelta 1950 on Elonet-tietokannassa saanut tunnisteen ”113501”, minkä pohjalta sille tehtiin MoMaF-hankkeen oma tunniste: <http://momaf-data.utu.fi/elonet_elokuva_113501>.

Tämä on tieteellisen hankkeen tarpeisiin muunnettavan aineiston kohdalla hyödyllistä, sillä tällöin aineiston tuonti ja muunnos ovat toistettavissa ilman, että mahdollisesti aineistoon jo tehdyt linkitykset katoavat. Tämä mahdollisti muun muassa sen, että hankkeen aikana Elonet-tietokannasta haettiin kahteen otteeseen elokuvista uudet metatiedot, jotka sitten päivitettiin tietokantaan. Koska tietokantaan tuotettu oma aines viittasi elokuvaan aina samalla tunnisteella, oli mahdollista korvata vain Elonetista peräisin oleva metadata ilman, että muu metadata-aineisto tietokannassa olisi vaatinut päivitystä.

Näin muunnettu ja tuotu aineisto tallennettiin Apache Jena -tietokannassa yhteen nimettyyn graafiin eli nimettyyn RDF-kolmikoiden ryhmään (ks. yllä). Esimerkiksi graafi <http://momaf-data.utu.fi/kf-data> sisältää kaiken Elonetin kansallisfilmografia-tietokannasta tuodun aineiston. Näin nimettyjen graafien käyttö aineiston hallinnassa mahdollistaa aineistokokonaisuuksien dokumentaation, muokkaamisen ja korvaamisen. Nimettyjen graafien avulla myös aineiston haut on mahdollista kohdentaa vain tiettyihin osiin ainestoa esimerkiksi vertailtaessa eri prosessien tuottamia tuloksia.

Elonet-aineiston lisäksi tietokantaan haettiin lisätietoa näyttelijöistä Wikidata-palvelusta sekä näyttelijöiden valokuvia IMDB-tietokannasta. Näiden löytämiseen tarvittava informaatio löytyi Wikidatan avulla, sillä Wikidataan on tallennettu henkilöiden ja elokuvien yhteyteen niiden Elonet-tunniste, jota käytettiin hankkeen omien tunnisteiden luomiseen. Elonet-tunnisteen avulla oli helppo liittää toisiinsa Wikidatan tiedot elokuvista ja ihmisistä.

Analyysien tulokset

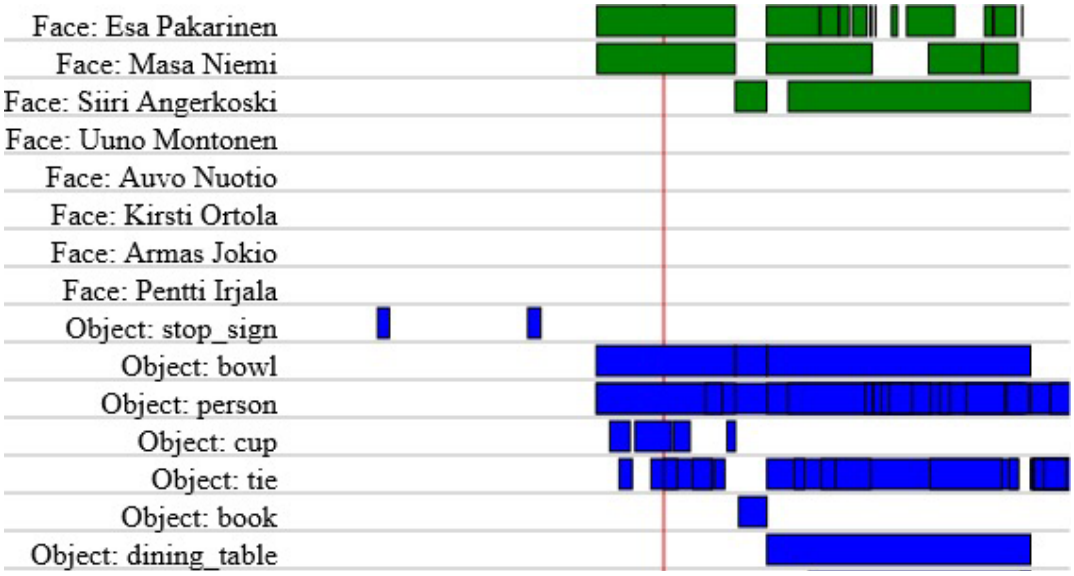
MoMaF-hankkeessa elokuvia on käsitelty erilaisin algoritmisin menetelmin muun muassa kasvojen tunnistuksen, esineiden tunnistuksen, äänten tunnistuksen ja puheen tunnistuksen tarpeisiin. Kukin menetelmä tuottaa omanlaistaan tulostietoa. Tulokset on pyritty integroimaan samaan tietokantaan, johon elokuvien metatieto on tallennettu.

Yllä mainittujen tunnistusmenetelmien tuloksena syntyy elokuvia koskevaa annotaatiodataa. Kukin väline tuottaa tätä hieman omassa muodossaan, mutta lähtökohta kaikessa annotoinnissa on sama: se viitaa tiettyyn elokuvatiedostoon, sillä on alku- ja loppukohdat ja se liittää elokuvatiedoston tiettyyn osaan jonkin tiedon, kuten henkilön, esineen, äänen tai tekstin. Kunkin annotaatiotyypin kohdalla on omat erityispiirteensä, mutta lähtökohtana voidaan silti käyttää samaa perusrakennetta, jolla eri menetelmien tuottama kuvaileva aineisto saadaan lähtökohtaisesti yhteensopivaksi. Näin on mahdollista hakea esimerkiksi tietty kohta elokuvasta ja saada kaikki kyseistä kohtaa koskeva annotaatioaineisto yhdellä haulla.

Kasvojen ja esineiden tunnistuksen tuottaman aineiston yksi piirre on se, että kasvot ja esineet eivät yleensä täytä koko kuvaa, vaan tunnistealgoritmit tuottavat yleensä laatikon, jonka sisälle tunnistettu kasvo tai esine kuvassa jää. Näin samassa kuvassa saattaa olla useitakin tunnistettuja kasvoja ja esineitä. Annotaatioiden kirjaamisessa onkin syytä olla tarkkana, jotta kaikki menee oikein. Lisäksi kasvojen ja esineiden tunnistusalgoritmit käsittelevät yleensä yksittäisiä kuvia, mutta yksittäisiä filmin ruutuja koskevien annotaatioiden tallentaminen ei välttämättä tuota kovinkaan hyödyllistä dataa. Objektien rajalaatikot ja saman objektin rajalaatikoiden muodostama ”rata” onkin syytä tallentaa yhteen annotaation yksittäisiä ruutuja koskevien rajalaatikoiden sarjana.[13] Niinpä kuhunkin objekti- ja kasvojentunnistukseen liittyy sarja kehyksiä, jotka kuvaavat kyseisen kohteen kuva-alaa ruudulla. Kuva 1 on esimerkki yhtä ruutua koskevasta visuaalisten annotaatioiden datasta. Kuvassa 2 taas nähdään esimerkki siitä, miten annotaatioiden rakentaminen sarjoiksi mahdollistaa niiden visuaalisen havainnollistamisen.

Äänten- ja puheentunnistuksessa vastaavia ongelmia ei ole. Puheentunnistuksessa voidaan kuitenkin keskustella siitä, kuinka pienten yksiköiden tasolla annotaatiot luodaan, toisin sanoen kirjataanko annotaatioiksi sanoja vai niiden muodostamia suurempia kokonaisuuksia. Tältä osalta hankkeessa ei vielä ole päästy täsmälliseen lopputulokseen.

Näiden lisäksi tietokantaan tallennettiin myös *Suomen kansallisfilmografian* kuvailevan tekstiaineksen sisältö. Tälle aineistolle rakennettiin lisäksi Apache Lucene -pohjainen teksti-indeksi, jonka avulla voitiin tehdä tehokkaasti vapaatekstihakuja elokuvien sisältöä vapaamuotoisesti kuvaavista kentistä. Tätä aineistoa käytettiin myös tekstintunnistusmenetelmien kouluttamisessa ja elokuvia kuvaavan aineiston analyysissä (Ginter et al. 2022).



Kuva 2. Visuaalinen esitys kasvon- ja objektintunnistusten tuottamien sarjojen havainnollistamisesta. Kohteena on elokuva *Pekka ja Pätkä pahassa pulassa* (1955). Aika kulkee vaakasuunnassa, ja värilliset palkit kuvaavat kunkin tunnisteen esiintymää elokuvassa.

13 ”Saman objektin” määritelmä tässä tapauksessa ei ole yksiselitteinen. Tässä yhteydessä on käytetty Jorma Laaksosen MoMaF-projektissa kehittelemää heuristiikkaa samojen objektien tunnistamiseksi peräkkäisissä ruuduissa. Tämän avulla on ollut mahdollista yhdistää peräkkäisissä ruuduissa esiintyvä objekti sarjaksi, joka on tunnistettavissa.

Jatkokäyttö

Metadata-aineiston käytön suunnittelun yksi lähtökohta jo hankkeen alkuvaiheessa oli sen jatkokäytettävyyden mahdollistaminen. Myös tämä seikka puolsi aineiston käsittelyä RDF-datana esimerkiksi perinteisen relaatiotietokannan sijaan. Yksittäisten analyysien yhteydessä aineistoa varmasti käsiteltiin siten kuin sen kannalta oli parasta, mutta tulosten integroimisessa toisiinsa hyödynnettiin RDF-muotoisen aineiston tarjoamia sisäisen viittaamisen mahdollisuuksia. Aineiston jatkokäytön kannalta on hyödyllistä, että RDF tukee erityisesti semanttisen verkon tarpeita.

Käytännössä tämä tarkoittaa sitä, että hankkeen kuluessa synnytetyt aineistot syötetään yhteiseen tietokantaan siten, että ne linkittyvät olemassa oleviin aineistoihin. Lopputuloksena syntynyt datajoukko on mahdollista ottaa ulos tietokannasta yhtensä suurena RDF-tyyppisenä aineistokokonaisuutena, joka on sellaisenaan vietävissä mihin tahansa muuhun RDF-tietokantaan.

Lopuksi

Monen toimijan tutkimushankkeessa syntyvän heterogeenisen aineiston yhteensovittamiseen ja tallentamiseen semanttinen tietokanta ja sen ympärille rakennettu metatiedon ja tulosaineiston käsittely toimii hyvin mutta vaatii panostusta koko projektin ajan. RDF-muotoisen aineiston tuottaminen ei välttämättä ole lähtökohtaisesti tutkijoille entuudestaan tuttua. Siksi datan tallennusrutiinien kehittäminen tuottaa jonkin verran lisätyötä. Tällaisessa suurten aineistomassojen käsittelyyn ja analyysiin pohjautuvassa hankkeessa moni käytetty työväline tuottaa taulukkomuotoista dataa, jonka muuntaminen RDF-muotoon ei yleensä onnistu automaattisesti ilman esivalmisteluita. Toisaalta ei myöskään ole vaikeaa kehittää kuvauksia taulukkomuotoisen datan palstojen muuntamiseksi RDF-muotoisen datan vaatimiksi lauseiksi.[14]

Joka tapauksessa keskitetty tulosten ja datan integrointi hankkeen aikana vaatii asialle omistettuja resursseja ja koordinointia. Mutta tämän tulisi joka tapauksessa olla lähtökohtana usean toimijan hankkeissa, joissa yhdistetään eri tutkimusperinteistä nousevia tutkimusryhmiä omine toimintatapoineen.

Teknisesti valittu ratkaisu toimii tarpeisiinsa nähden suhteellisen hyvin. Tiedon jäsentäminen ja uuden tiedon integroiminen olemassa olevaan malliin on sujuvaa ja joustavaa, vaikkakin mallin jäsentäminen ja hahmottaminen on työlästä. Samaan aikaan aineiston hakuominaisuudet ovat erinomaiset, ja hyvinkin monimutkaisten kysymysten esittäminen aineistolta on mahdollista. Kääntöpuolena on hakuominaisuuksien esittelyn vaikeus ilman asialle omistettua käyttöliittymää.

Vaikka hankkeessa on keskitytty valitsemaan menetelmiä ja työkaluja, jotka soveltuvat ennen kaikkea MoMaF-hankkeen käyttöön, erityisesti Aalto-yliopiston piirissä toteutettujen Sampo-hankkeiden pohjalta on syytä olettaa, ettei näiden menetelmien hyödyllisyys rajoitu tähän, vaan niillä on laajempiakin hyödyntämismahdollisuuksia esimerkiksi kulttuuriperintöaineistojen tarkastelussa.[15]

14 Valmiitakin muuntimia esimerkiksi CSV- ja TSV- (pilkuin tai tabulaattorimerkein kentiksi jaoteltu data) muotoisen taulukkoaineiston tuomiseksi RDF-muotoon on olemassa. Tällaisten hyödyntäminen voi olla järkevää, mikäli ohjelmointiosaamista ei löydy, mutta RDF-datan tuottaminen esimerkiksi Python-kielellä on aika yksinkertaista.

15 Sampo-hankkeista tunnetuimmat lienevät tällä hetkellä Sotasampo <https://www.sotasampo.fi/> ja Kirjasampo <https://www.kirjasampo.fi/>.

Lähteet

Allemang, Dean & Hendler, James (2011) *Semantic web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann Publishers.

Ginter, Filip; Kiiskinen, Harri; Kanerva, Jenna; Chang, Li-Hsin & Salmi, Hannu (2022) Deep Learning and Film History: Model Explanation Techniques in the Analysis of Temporality in Finnish Fiction Film Metadata. *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference 2022*, CEUR Workshop Proceedings vol. 3232, 50–62.

Grósz, Tamás; Kallioniemi, Noora; Kiiskinen, Harri; Laine, Kimmo; Moisio, Anssi; Römpötti, Tommi; Virkkunen, Anja; Salmi, Hannu; Kurimo, Mikko & Laaksonen, Jorma (2022) Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Toward a Multimodal Analysis of Audiovisual Data. Proceedings of *the 6th Digital Humanities in the Nordic and Baltic Countries Conference 2022*, CEUR Workshop Proceedings vol. 3232, 63–78.

Hyvönen, Eero (2018) *Semanttinen web. Linkitetyn avoimen datan käsikirja*. Helsinki: Gaudeamus.