Roope Ahola

Tietomallien linkitysmenetelmät ja rooli linkitysprosesseissa

Tietotekniikan
pro gradu -tutkielma
5. marraskuuta 2024

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Kokkolan yliopistokeskus Chydenius

Tekijä: Roope Ahola
Yhteystiedot: roope.ahola@live.com
Puhelinnumero: 040-837 0560
Ohjaaja: Risto T. Honkanen
Työn nimi: Tietomallien linkitysmenetelmät ja rooli linkitysprosesseissa
Title in English: Data model linking methods and the role of data models in linking process
Työ: Tietotekniikan pro gradu -tutkielma
Sivumäärä: 72+9
Tiivistelmä: Työn tutkimuskysymkset ovat: 1) Miten pienen perehtymisen jälkeen domain-osaajien on mahdollista tuottaa linkittämiseen vaadittava määrittelydokumentti? 2) Miten paljon teknistä osaamista määrittelydokumentin muodostaminen vaatii domain-osaajalta? 3) Mikä on linkityksessä tuotetun määrittelydokumentin jatkohyödynnettävyys toisissa linkityksissä? 4) Mitkä tekijät määrittelydokumenteissa vaikuttavat yhdistetyn datan laatuun?

Tutkimusmenetelmänä käytettiin tapaustutkimusta, josssa tapauksina olivat valitut linkitysmenetelmät ja niitä sovellettiin joukkoliikenteen kahden tietomallin välisten entiteettien linkityksissä. Tiedonkeruumenetelminä toimivat tehdyt linkitysharjoitukset sekä semi-strukturoidut haastattelut domain-asiantuntijoiden ja dataintegraattorien kanssa. Haastattelut litteroitiin ja analysoitiin luokittelemalla keskeiset havainnot.

Päätuloksina todettiin, että erityisesti domain-osaajan aikaisempi kokemus sekä tietotekninen taitotaso vaikuttavat siihen, miten helposti valittavalla linkitysmenetelmällä saadaan tuotettua määrittelydokumentti. Toisen tutkimuskysymyksen osalta vastaus ei ole yksiselitteinen. Linkitystaulukko on yleisesti helpommin lähestyttävä ei-teknisen henkilön toimesta, mikäli halutaan varmemmin tuotettua määrittelydokumentti dataintegraattorien käytettäväksi. RML-menetelmän käyttökokemusta voidaan parantaa yhteensopivilla käyttöliittymäeditoreilla. Dataintegraattorit korostivat, että paikallisen tason linkityksissä linkitysmenetelmällä ei ole suurta merkitystä, vaan tärkeintä on iteraatio sidosryhmien välillä. Kolmannen tutkimuskysymyksen osalta todettiin, että RML-menetelmällä tuotettu määrittelydokumentti on helpommin tarkistettavissa syntaktisten virheiden varalta ja se mahdollistaa paremman hyödynnettävyyden erilaisissa graafitietokantojen avulla tehtävissä linkityksissä. Viimeisen tutkimuskysymyksen löydettiin kaksi tekijää, jotka vaikuttavat yhdistetyn datan laatuun. Määrittelydokumentin syntaktinen laatu on

tärkeää, kun määrittelydokumenttia käytetään suoraan teknisessä dataintegraatiossa. Toisena tekijänä tunnistettiin määrittelydokumentin kyky tukea sidosryhmien iteraatiota.
Avainsanat: Tietomallit, Datan linkitysmenetelmät, liikenteen tietomallit
Abstract: The research questions are: 1) How can domain experts produce the required specification document for linking after a little familiarisation? 2) How much technical knowledge does it require for a domain expert to generate a configuration document? 3) What is the usefulness of the final output (the definition document) produced in the linking process for further use in other linking processes? 4) What factors in the specification documents affect the quality of the linked data?

The research method used was a case study, where selected linking methods were applied to the linking of entities between two data models of public transport. The data collection methods used were the linking exercises performed and semi-structured interviews with domain experts and data integrators. Recordings of the interviews were made and the data was analysed by transcribing and classifying the key findings.

The main conclusions were that the domain expert's previous experience and level of IT skills have a particular impact on the ease with which the chosen linking method can produce a specification document. As regards the second research question, the answer is not unequivocal. A linking table is generally easier to approach by a non-technical person if a more robust specification document is to be produced for use by data integrators. The user experience of the RML method can be enhanced by compatible editor software. The data integrators pointed out that for local level linking, the linking method is not very important, but the most important thing to build a common understanding is the iteration between the stakeholders. Regarding the third research question, it was found that the specification document produced using the RML method is easier to check for syntactic errors and also allows for better usability for linking to different graph databases. The final research question identified two factors that affect the quality of the combined data. The syntactic quality of the specification document is important when the specification document is used directly in technical data integration. The second factor identified was the ability of the specification document to support stakeholder iteration.
Keywords: Data models, data linking methods, mobility data

Esipuhe

Tässä työssä tarkastelen tietomallien linkittämiseen liittyviä menetelmiä ja prosesseja, joita olen havainnut oman kokemukseni kautta. Tietomallit ovat keskeisiä nykyaikaisessa tiedonhallinnassa, mutta niiden tehokas linkittäminen tuo mukanaan monia haasteita, kuten yhteensopivuusongelmia ja prosessien monimutkaisuutta. Näiden haasteiden ymmärtäminen on ollut tärkeä osa tutkimustani. Toivon, että löydökseni tarjoavat uusia näkökulmia ja ratkaisuja alalla.

Haluan myös kiittää ohjaajaani Risto Honkasta, joka on antanut arvokasta tukea ja ohjausta tämän työn aikana. Kiitokset myös kaikille haastattelututkimukseen osallistuneille, joiden näkemykset ja kokemukset ovat olleet korvaamattomia tutkimukseni tueksi. Erityiset kiitokset perheelleni, joka on ollut tukena ja kannustamassa minua koko tämän matkan ajan.

Toivon, että tämä työ inspiroi muita ja edistää keskustelua tietomallien linkittämisestä ja niiden prosesseista.

Sanasto

CEN	The European Committee for Standardization
CSV	Comma Separated Values
GTFS	General Transit Feed Specification
GTFS rt	General Transit Feed Specification realtime
IFOPT	Identification of Fixed Objects in Public Transport
JSON	JavaScript Object Notation
LOT	Linked Open Terms
MMTIS	Multimodal Travel Information Service
NeTEx	Network Timetable Exchange
ODIN	Open Mobility Data In the Nordics
ORSD	Ontology Requirement Specification Document
POI	Point Of Interest
R2RML	RDB to RDF Mapping Language Schema
RDF	Resource Description Framework
RML	RDF Mapping Language
SPARQL	RDF query language
Transmodel	Reference Data Model For Public Transport
UML	Unified Modeling Language
XML	Extensible Markup Language
XSD	XML Schema Definition

Sisällys

Liitteet

1 Johdanto

Erilaiset trendit sekä muutostarpeet ympäristöystävällisempien ja parempien joukkoliikennepalveluiden tarjoamiseksi ovat kiristäneet lainsäädännön vaatimuksia joukkoliikenteen tietomallien osalta [19, s. 14]. Palvelutasovaatimusten kasvaessa erilaiset avoimen datan strategiat sekä tietomallien yhteensopivuuteen panostaminen ovat lisääntyneet [22] [19, s.18]. Tämä mahdollistaa uusien palveluiden syntymisen sekä palvelulaadun parantamisen. Tavoitteet yhteensopivuuden parantamisesta ovat johtaneet erilaisten tietomallien linkitysmenetelmien muodostamiseen ja toteutustapoja onkin syntynyt vuosien aikana lukuisia. Esimerkiksi Dimou et al. [9] sekä Scrocca et al. [23] esittelevät hyvin erilaisista lähestymiskulmista toteutettuja linkitysmenetelmiä liiketoimintatavoitteiden saavuttamiseksi. Tässä työssä tuodaan esille tarpeiden, sitä tukevan lainsäädännön, teknisten ratkaisujen ja toimintaprosessien välisiä riippuvuuksia, kun tarkastellaan joukkoliikenteen suurempaa uudistusta kohti yhteensopivampaa järjestelmäkokonaisuutta.

Tutkimuskysymyksinä esitetään:

- Miten pienen perehtymisen jälkeen domain-osaajien on mahdollista tuottaa linkittämiseen vaadittava määrittelydokumentti?
- Miten paljon teknistä osaamista määrittelydokumentin muodostaminen vaatii domain-osaajalta?
- Mikä on linkityksessä tuotetun määrittelydokumentin jatkohyödynnettävyys toisissa linkityksissä?
- Mitkä tekijät määrittelydokumenteissa vaikuttavat yhdistetyn datan laatuun?

Tutkimuskysymykset vastaavat muun muassa VTT:n raportissa [19] esitettyihin tarpeisiin laadukkaan ja täsmällisen datan tuottamiseksi liiketoiminnan tarpeisiin sekä työssä esitettävien linkitysmenetelmien soveltuvuuteen erilaisissa tapauksissa.

Työn teoriaosuudessa käydään läpi joukkoliikenteen palvelutietoja kuvaavien yleisten tietomallistandardien ja -implementaatioiden ominaispiirteitä. Tietomallien osalta esitetään myös näkökulmia, joiden pohjalta tietomallit ovat kehittyneet. Lisäksi esitetään kirjallisuuden perusteella joukkoliikenteen tietomallien linkityksissä käytettyjä menetelmiä.

Tutkimusmenetelmänä käytetään tapaustutkimusta, jossa tapauksina ovat valitut linkitysmenetelmät ja niitä sovelletaan joukkoliikenteen kahden tietomallin välisten entiteettien linkityksissä. Tiedonkeruumenetelminä toimivat tehtävät linkitysharjoitukset sekä semi-strukturoidut haastattelut domain-asiantuntijoiden ja dataintegraattorien kanssa. Haastattelut litteroitiin ja analysoitiin luokittelemalla keskeiset havainnot.

Päätuloksina todetaan, että erityisesti domain-osaajan aikaisempi kokemus sekä tietotekninen taitotaso vaikuttavat siihen, miten helposti valittavalla linkitysmenetelmällä saadaan tuotettua määrittelydokumentti. Toisen tutkimuskysymyksen osalta vastaus ei ole yksiselitteinen. Linkitystaulukko on yleisesti helpommin lähestyttävä ei-teknisen henkilön toimesta, mikäli halutaan varmemmin tuotettua määrittelydokumentti dataintegraattorien käytettäväksi. RML-menetelmän käyttökokemusta voidaan parantaa yhteensopivilla käyttöliittymäeditoreilla. Dataintegraattorit korostavat, että paikallisen tason linkityksissä linkitysmenetelmällä ei ole suurta merkitystä, vaan tärkeintä yhteisen ymmärryksen muodostamiseksi on iteraatio domain-asiantuntijan ja dataintegraattorin välillä. Kolmannen tutkimuskysymyksen osalta todetaan, että RML-menetelmällä tuotettu määrittelydokumentti on helpommin tarkistettavissa syntaktisten virheiden varalta ja lisäksi se mahdollistaa paremman jatkohyödynnettävyyden erilaisissa graafitietokantojen avulla tehtävissä linkityksissä. Yleisesti haastateltavat kokevat, että linkitystaulukko tarjoaa paremmin mahdollisuuden lisätä implisiittisiä kommentteja linkityksien osalta, mikä tukee iteraatiota eri sidosryhmien välillä. Viimeisen tutkimuskysymykseen löydetään kaksi tekijää, jotka vaikuttavat yhdistetyn datan laatuun. Määrittelydokumentin syntaktinen laatu on tärkeää, kun määrittelydokumenttia käytetään suoraan teknisessä dataintegraatiossa. Toisena tekijänä tunnistetaan määrittelydokumentin kyky tukea sidosryhmien iteraatiota. Vertailtavista linkitysmenetelmistä on mahdotonta arvioida paremmuutta yksiselitteisesti, sillä usean asian kokonaisuus määrittelee, mikä on paras menetelmä käytettäväksi.

Toinen luku käsittelee joukkoliikenteen tietomalleihin liittyvää lainsäädäntöä ja tavoitteiden saavuttamiseksi vaadittavaa yhteistyötä eri sidosryhmien välillä tietomallien sekä datan laadun parantamiseksi. Kolmannessa luvussa esitellään työn kannalta keskeisiä joukkoliikenteen tietomalleja sekä niiden rooleja eri joukkoliikenteen toiminnoissa. Luvussa esitellään kaupallisista lähtökohdista muodostetun GTFS-tietomallin erityispiirteitä ja viranomaisvetoisista lähtökohdista kehittyä Transmodel-viitekehystä sekä sen implementaationa toimivaa NeTEx-tietomallia. Neljäs

luku käsittelee tietomallien linkitysmenetelmiä ja havainnollistaa niiden tarpeellisuuden eheän tiedon tarjoamiseksi. Erityisesti käydään läpi, miksi linkitysmenetelmiä tarvitaan ja mikä merkitys on tietomallistandardeilla linkityksissä ja luvussa esitellään erilaisia menetelmiä, jotka lähestyvät tiedon yhteensovittamiseen liittyviä haasteita erilaisista näkökulmista. Nämä esitetyt linkitysmenetelmät on kehitetty erilaisista näkökulmista ja niiden ominaisuuksia on havainnollistettu joukkoliikennekontekstin osalta. Vastaavasti on nostettu esille linkitysmenetelmien käytössä esiintyviä haasteita, joita tulisi huomioida onnistuneen lopputuloksen kannalta. Viidennessä luvussa esitetään tutkimuksen toteutusvaiheeseen liittyviä asioita ja esitetään tutkimuksen tarkoitus ja tutkimuskysymykset. Erityisesti havainnollistetaan joukkoliikenteen toimintakentän muutoksista johdettavia motivaatiotekijöitä, joiden pohjalta olen tunnistanut tutkimuskysymyksiä. Näiden lisäksi käydään tutkimusasetelmaan liittyvia asioita, kuten tutkimusmenetelmä, tiedonkeruumenetelmät ja datan analyysi. Tämä havainnollistaa teoriaa toteutettavan tutkimuksen sovellettavista menetelmistä ja eri vaiheet sidosryhmät, valitut linkitysmenetelmät ja kuvataan tutkimuksen eteneminen tutkijan näkökulmasta. Kuudes luku esittelee tutkimuksen tulokset, jossa esitetään keskeisimmät havainnot domain-asiantuntijoiden ja dataintegraattorien haastatteluista. Luvun lopussa on haastatteluiden tuloksien yhteenveto sekä niitä verrataan tutkimuskysymyksiin. Seitsemännessä luvussa analysoidaan teorian sekä empiirisessä tutkimusvaiheessa saatujen tuloksien pohjalta tietomallien linkitykseen liittyviä tekijöitä, linkitysmenetelmien eroavaisuuksia ja niiden ominaisuuksia toisiinsa. Erityisesti luvussa pohditaan määrittelyprosessien suhdetta linkitysmenetelmiin ja pyritään tunnistamaan lopputuloksen kannalta olennainen riippuvuus linkitysmenetelmän ja toiminta- sekä applikaatioprosessien välillä. Luvun lopussa peilataan koko toimintakentän riippuvuuksia toisiinsa sekä isojen osakokonaisuuksien vaikutusta tutkimuksen havaintoihin sekä linkitysmenetelmiin. Viimeisessä luvussa on esitetty työn yhteenveto ja johtopäätökset.

2 Joukkoliikenteen toimintakentän muutokset sekä tavoitteet

Joukkoliikenteen palvelutasot ja asiakkaiden vaatimustasot ovat kasvaneet ja näiden mahdollistamiseksi lainsäädännön vaatimukset tuottavat uusia vaatimuksia sekä vaikutuksia joukkoliikenteen palvelutuottajien toimintaan.

Julkisella sektorilla on tarpeen kustannussäästöpaineiden sekä ilmastonmuutoksen myötä tehostaa sekä monipuolistaa tarjottavia joukkoliikenteen palveluita. Julkisen sekä yksityissektorin osalta odotetaan avoimempia, kustannustehokkaita palveluita ja yhteensopivia järjestelmiä, jotka tukevat monipuolisia jakamistalouden palveluita [19, s.18-20]. Useampi lähde, kuten Scrocca et al. [23] sekä Ruckhaus et al. [22] mainitsevat kasvavan palvelulaadun, vaatimukset yhteensopivista järjestelmistä sekä liiketoimintatavoitteiden vaikuttavan palvelutuottajien toimintaan. Liiketoimintaa tukevat erilaiset analytiikkapalvelut ja niiden muodostamat kyselyt kasvattavat vaatimuksia tietomallien semanttisille linkityksille, mikä lisää edelleen big datan hallinnan monimutkaisuutta [1].

Aliluvussa 2.1 käsitellään joukkoliikenteen tietomalleihin liittyvää lainsäädäntöä sekä tavoitteiden saavuttamiseksi vaadittavaa yhteistyötä eri sidosryhmien välillä. Aliluvussa 2.2 vastaavasti esitellään eri tasoilla tehtävää yhteistyötä tietomallien sekä datan laadun parantamiseksi.

2.1 Lainsäädäntö sekä kansainvälinen yhteistyö

Vuoden 2015 hallitusohjelmassa kirjattiin tavoitteeksi liikenteen tietojen jakamista ja tietovarastojen tehokkaampaa käyttöä, jotta voidaan tukea liikenteen palvelukentän kehittymistä sekä mahdollistaa yhtenäisten matkaketjujen muodostuminen kansallisella tasolla. VTT:n tutkimusraportissa [19] nostetaan esille, että kansallisella tasolla kaikki joukkoliikenteen toimijat tahtovat tiedon kattavampaa koontia, joukkoliikenteen datan laadun parantamista ja muodostamista tulevaisuuden tarpeita palvellen.

Euroopan komission strategisena tavoitteena on edistää liikenteen tieto- ja viestintätekniikan roolia osana älykkäiden liikenteen palveluiden kehittämiseksi ITS-

direktiivin (2010) avulla. Vuonna 2017 tuli voimaan MMTIS-asetus[1], joka velvoittaa liikennepalveluiden järjestäjiä avaamaan palveluiden osalta asetuksessa määritetyt tiedot digitaalisessa yhteensopivassa muodossa. Lisäksi asetuksessa velvoitetaan tietojen jakamista kansallisten yhteyspisteiden (NAP) kautta kaikkien saataville [19, s.18-20]. Myös Scrocca et al. [23] ja Ruckhaus et al. [22] toteavat, että EU:n asetus 2017/1926 vaatii tiedon jakamista yhteensopivassa CEN standardoidussa Transmodel-pohjaisessa tietomallissa kansallisten yhteyspisteiden kautta.

VTT:n raportin [19, s.18-20] mukaan erilaisten kunnianhimoisten tutkimusohjelmien avulla pyritään edistämään yhtenäisiä matkaketjuja sekä mahdollistamaan multimodaalinen joukkoliikenne koko Euroopan tasolla ja tavoitteen saavuttamiseksi ratkaisuja kehitetään eurooppalaisten Transmodel-pohjaisten standardien päälle. EU:n tavoitteena on kehittää kattava sisärajat ylittävä markkina eri toimialoilla, kuten älykkään liikennejärjestelmän osalta, mitä tukevat kattava datastrategia sekä "Yhteinen eurooppalainen liikkuvuuden data-avaruus". Scrocca et al. [23] toteavat EU:n tavoitteen olevan kirkas ja määrätietoinen heterogeenisten tietolähteiden rajaamisen osalta sekä tiedon laadun ja yhteensopivuuden osalta. Bellini et al. toteavat [1] lainsäädännön asettamien velvoitteiden muodostavan kuitenkin haasteita datan omistajuuden sekä luotettavuuden osalta, miten varmistetaan omistajien oikeuksien säilyminen sekä datan luotettavuus hyödyntäjien näkökulmasta.

2.2 Yhteistyö tietomallien yhteensovittamisessa

EU:n ja kansallisten lainsäädäntöjen sekä strategioiden myötä yhteistyö eri tasoilla on tiivistynyt. Erityisesti pohjoismainen yhteistyö joukkoliikenteen matkatiedon sekä palveluiden kehittämiseksi on tiivistynyt ”Open Mobility Data in the Nordics”(ODIN) -yhteistyöprojektin toimesta, jonka tavoitteina ovat mm. Matkatiedon sisällön sekä laadun varmistaminen ODIN-projektin ohjauksessa toimivan standardityöryhmän avulla, pohjoismaisten ratkaisujen yhteiskehittäminen sekä yhteistyö EU-yhteensopivuuden varmistamiseksi. Erityisesti ODIN-yhteistyössä edistetään pohjoismaisen ”NeTEx Nordic Profile” kehitystyötä, jossa kehitetään pohjoismaista tietomalli-implementaatiota CEN standardin NeTEx-tietomallin osalta [19, s.18-20]. Lisäksi EU-tasolla on aktiivinen yhteistyöryhmä, NeTEx-CEN Working Group TC278/-

[1]Komission delegoitu asetus (EU) 2017/1926, annettu 31 päivänä toukokuuta 2017,Euroopan parlamentin ja neuvoston direktiivin 2010/40/EU täydentämisestä EU:n laajuisten multimodaalisten matkatietopalvelujen tarjoamisen osalta

WG3/SG9 [17], tietomallin eri implementaatioiden yhteensovittamisen sekä tietomallin kehitystarpeiden tunnistamiseksi.

Myös Scrocca et al. [23] sekä Ruckhaus et al. [22] esittävät SNAP-projektin, jossa espanjalaisten viranomaisten olemassa olevia tietomalleja yhteensovitetaan EU-lainsäädännön mukaisiin vaatimuksiin. Edellä mainitun lainsäädännön vaatimuksien mukaiset tietomallit ovat laajoja, mikä vaatii syvää asiantuntemusta kontekstista. Viime vuosina kansalliset sekä kansainväliset organisaatiot ovat pyrkineet luomaan keskitettyjä data hub-ratkaisuja, joissa on kyvykkyys käsitellä eri tietomalleihin perustuvaa dataa keskitetysti ja perustuen erilaisiin yhteensopivuutta painottaviin tietomalleihin [1]. Mylonas et al. [15] nostavat esille NAPCORE-organisaation, jonka tavoitteena on edistää kansallisten yhteyspisteiden yhdenmukaisia toimintoja, tarjoamaan koulutusta ja materiaaleja laadukkaan datan tuottamiseksi sekä jakamaan tietoisuutta olemassa olevista standardeista.

3 Tietomallit osana joukkoliikennettä

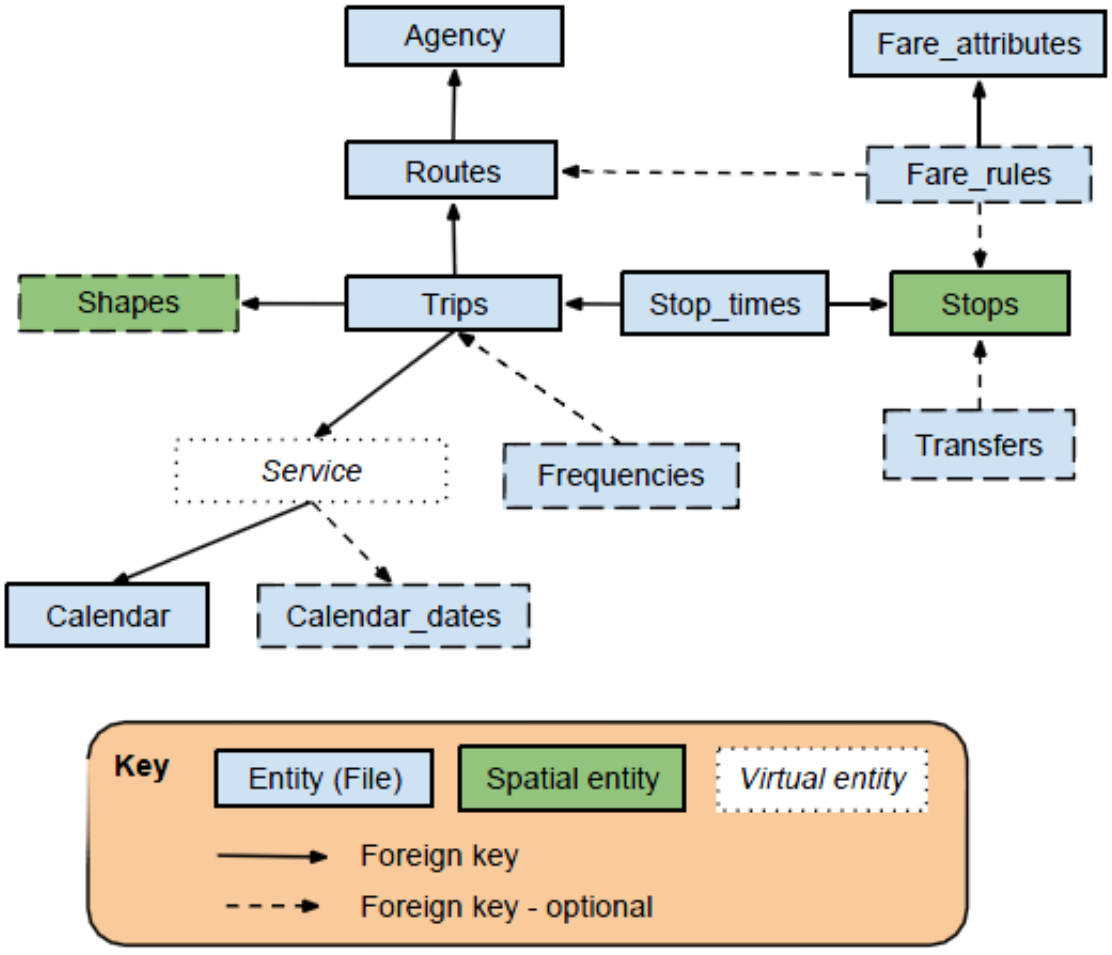
Joukkoliikenteen tietomalleilla kuvataan topologiaa, liikenteen infrastruktuuritietoja, matkatietoa palveluista ja hintatietoja [19, s. 21]. Joukkoliikenteen tietomalleja on sekä kaupallisista lähtökohdista että viranomaisnäkökulmasta kehitettyjä. Tietomalleilla esitetään koneluettavassa muodossa staattisia ja dynaamisia dataa joukkoliikenteen verkoston, palvelutarjonnan sekä hintojen esittämiseksi. Staattisella tiedolla tarkoitetaan tietoa, joka muuttuu harvemmin ja on järkevämpää hyödyntää eri palveluissa eräajona ladattavana pakettina tai tiedostoina. Vastaavasti dynaaminen tieto on reaaliaikaista ja täydentää usein staattista tietoa.

Tässä luvussa esitellään työn kannalta keskeisiä joukkoliikenteen tietomalleja sekä niiden rooleja eri joukkoliikenteen toiminnoissa. Aliluvussa 3.1 käydään läpi kaupallisista lähtökohdista muodostetun GTFS-tietomallin erityispiirteitä, kun aliluvut 3.2 ja 3.3 esittelevät viranomaisvetoisista lähtökohdista kehittyä Transmodel-viitekehystä sekä sen implementaationa toimivan NeTEx-tietomallin.

3.1 GTFS

Yleisin joukkoliikenteen tietomalli on Googlen kehittämä GTFS (General Transit Feed Specification), joka toimii erityisesti aikataulupalveluiden tarjoamiseen soveltuvana tietomallina. Yhdysvalloissa Oregonin osavaltiossa toimiva Portlandin seudullinen joukkoliikenneviranomainen TriMet sekä Google aloittivat heinäkuussa 2005 projektin, jossa tarkoituksena oli muodostaa yleisesti käytössä oleviin reittioppaisiin soveltuva tietomalli joukkoliikenteen palveluiden kuvaamiseksi [2]. Google kehitti staattisen GTFS-tietomallin v. 2006 aikataulutietojen esittämiseksi ja käyttää kyseistä tietomallia Google Transit -palvelussaan. GTFS-tietomallia ei ole standardoitu virallisen standardointitahon toimesta. GTFS-tietomalli on muodostunut laajimmaksi käytössä olevaksi tietomalliksi erityisesti joukkoliikenteen aikataulupalveluissa [19, s.21]. Myös Ruckhaus et al. toteavat [22] GTFS:n olevan yksinkertaisuutensa sekä suurien reitityspalveluiden tukemana standardina erittäin suosittu ja yleisesti käytetty standardi.

GTFS koostuu useampia CSV-tiedostoja sisältävästä ZIP-paketista, jossa yksittäi-



Kuva 3.1: GTFS -tietomalli [13]

nen tekstitiedosto kuvaa tietomallin taulua, kuten reitti- tai vuorotietoja [10]. GTFS:n tietomallin taulujen dataa yhdistetään toisiinsa relaatioavaimien avulla, jolloin yksittäisestä matkasta (trip) viitataan ajettavaan reittiin (route) ja reitistä on viite liikennöitsijään (agency), kuten kuva 3.1 havainnollistaa.

GTFS sisältää lukuisia eri laajennuksia, kuten GTFS-Flex [19, s.21], jonka avulla voidaan kuvata kutsujoukkoliikenteen palveluita monipuolisemmin kuin tavallisen reittipohjaisen joukkoliikenteen palveluita. GTFS-standardin puutteita sekä tarpeita on täydennetty laajennuksin erilaisiin käyttötarpeisiin, kuten kutsujoukkoliikenteen tietoja mallintava GTFS-Flex [19].

3.2 Transmodel

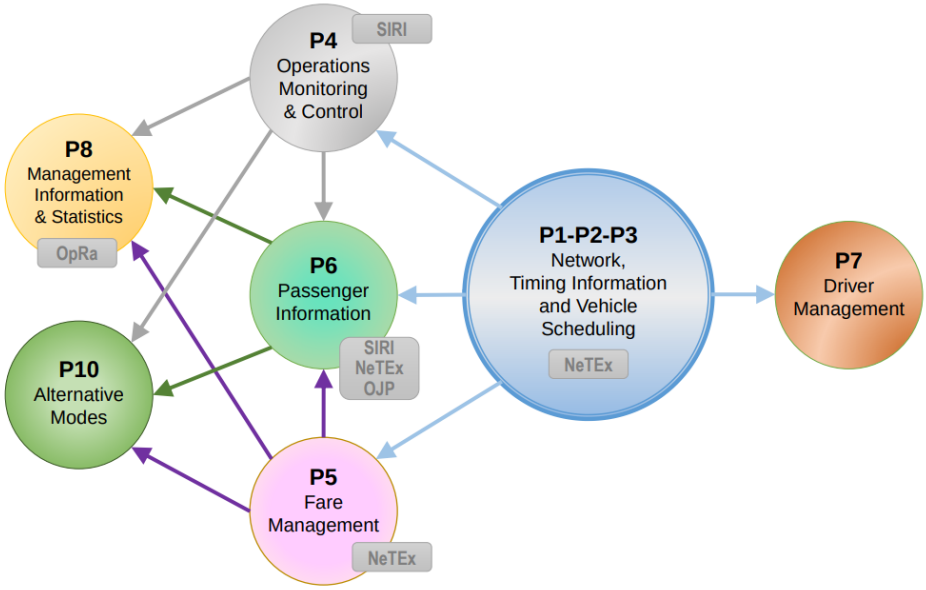
Transmodel on EU-direktiivin 2010/40/EU[1] pohjalta muodostettu älykkään liikenteen tietomallikehys, jonka tavoitteena on edistää liikenteen tietomallien yhteensopivuutta Euroopan laajuisesti [22]. Se lisää toiminnallisuuksia joukkoliikenteen informaation sekä palvelun tarjoamiseksi sekä mahdollistaa järjestelmien käyttämän datan yhteensopivuuden yhtenäisten määrityksien, rakenteiden sekä merkityksien avulla [25]. Transmodel on koko joukkoliikenteen kattava viitekehys, mikä mahdollistaa tietomallien kehittämisen myös tulevaisuuden tarpeita varten. Transmodel sisältää 8 eri osiota ja kuvaa laajasti eri asioiden riippuvuuksia, mutta ei määrittele tarkasti yksittäisten tietotyyppien sisältöä. Osiot jakaantuvat [25]:

- Osa 1: yleinen konsepti
- Osa 2: Joukkoliikenteen infrastruktuuri- ja verkkotiedot
- Osa 3: Aikataulutiedot sekä ajoneuvojen aikataulutus
- Osa 4: Operatiivinen monitorointi sekä hallinta
- Osa 5: Matkaoikeuksien sekä hinnoittelutietojen hallinta
- Osa 6: Matkustajainformaatio
- Osa 7: Kuljettajien hallinta

[1]EUROOPAN PARLAMENTIN JA NEUVOSTON DIREKTIIVI 2010/40/EU, annettu 7 päivänä heinäkuuta 2010, tieliikenteen älykkäiden liikennejärjestelmien käyttöönoton sekä tieliikenteen ja muiden liikennemuotojen rajapintojen puitteista

- Osa 8: Informaation ja tilastojen hallinta
- Osa 10: Vaihtoehtoiset kulkumuodot

Kuvassa 3.2 esitetyssä Transmodelin viitekehyksessä sinisellä oleva osa-alue käsittää edellä mainitut osa-alueet 1-3 ja on keskeinen osa joukkoliikenteen suunnitelman mallintavaa tietoa. Muut osiot laajentavat tietomallia eri näkökulmista, kuten osa 5, joka käsittää matkaoikeuksiin sekä hinnoitteluun liittyvät laajennukset.



Kuva 3.2: Transmodel -viitekehyksen sisältämät osiot[25]

Yleinen konsepti

Transmodelin ensimmäinen osa sisältää datakehykset, joita käytetään kaikissa eri osissa noudattamaan yhtenäistä datan rakennetta. Lisäksi ensimmäisessä osassa määritetään versiointi-, omistajuus- ja voimassaolotiedot datan elinkaaren hallitsemiseksi sekä käyttämiseksi ehdoissa määritetyllä tavalla. Tämän takia dataentiteeteillä on määritelty omistajuus-, versiotiedot sekä datan käyttämistä ohjaavat ehdot [26]. Liitteessä A on esitetty Transmodelin versiointimalli [24]. Ensimmäisessä osiossa esitetään uudelleen hyödynnettävät datatyypit, jotka eivät välttämättä ole suoraan joukkoliikenteen kontekstiin sidottuja, mutta niiden pohjalta joukkoliikenteen tietotyypit ovat periytyneet tai niitä käytetään Transmodelin toimesta.Näitä datatyyppe-

jä ovat paikkaan, liikennemuotoon, saavutettavuuteen ja kalentereihin liittyvät tiedot.Esimerkiksi POINT, LINK, LAYER, joiden avulla pisteet ja niiden väliset linkit voidaan yhdistää eri näkökulmia kuvaavien kerroksien välillä.

Joukkoliikenteen infrastruktuuri- ja verkkotiedot

Toinen osa sisältää verkko- ja infrastruktuuritietoja,jotka koostuvat tai periytyvät infrastruktuuripisteistä sekä linkeistä. Esimerkiksi reitti ROUTE muodostetaan peräkkäin olevista linkeistä infrastruktuurikerroksella ja näin periytyvän LINK SEQUENCE-tietotyypistä [26]. Osa 2 määrittelee myös muita pisteestä POINT ja LINK SEQUENCE -tyypistä periytyviä tietotyyppejä, joilla kuvataan joukkoliikenteen palvelua tai toimintaa pisteessä/linkillä, kuten matkustajalle tärkeä JOURNEY PATTERN, mikä koostuu ohitusajat sisältävistä pysäkeistä STOP POINT. Tämä malli on esitetty liitteessä B.

Aikataulutiedot sekä ajoneuvojen aikataulutus

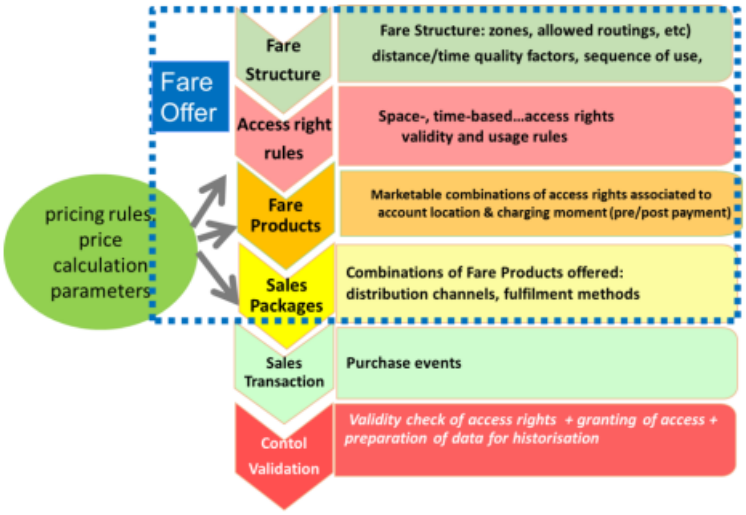
Transmodel osa 3 käsittää aikataulu- ja kalenteritiedot, kuten suunnitelman mukaiset lähtötiedot, reittien pysäkkivälien ajoitustiedot sekä suunnitellun liikenteen erilaiset säännöt mm. vaihtoyhteyksille [26]. Esimerkiksi osa 3 määrittää ajoneuvon yksittäisen matkan aikataulutiedot VEHICLE JOURNEY -mallin avulla, jossa tiettynä päivätyyppinä DAY TYPE suoritetaan matka tietyllä matkan reitillä JOURNEY PATTERN.JOURNEY PATTERN sisältää vastaavasti aikatietoja tietyissä pisteissä TIMING POINT [24]. Tämä malli on liitteessä C.

Operatiivinen monitorointi sekä hallinta

Osa 4 sisältää operatiiviseen toimintaan liittyvät suunnitellun liikenteen tiedot sekä reaaliaikaisesti esitettävät päiväkohtaiset muutokset kaluston käytöstä reiteillä, häiriötiedottamisen sekä odotusaikatietoja matkustajille [27]. Suunnitelman mukainen operatiivinen tieto sisältää ennakkoon suunniteltuja päiväkohtaisia tietoja ajoneuvojen matkoista (VEHICLE JOURNEY), jotka suoritetaan määriteltynä päivänä (OPERATING DAY). Tiettynä päivänä ajettava ajoneuvon vuoro on DATED VEHICLE JOURNEY, mikä on yhdistetty myös kuljettajien työvuoroihin sekä saatavilla olevaan kalustoon. Nämä kaikki tiedot ovat osana tuotantosuunnitelmaa (PRODUCTION PLAN), jonka elementit esitetty liitteessä D.

Matkaoikeuksien sekä hinnoittelutietojen hallinta

Osa 5 sisältää tietomallin tariffien, tuotteiden sekä myynnin rakenteista ja niiden parametreista, joita tarvitaan tuotteiden, taksojen sekä käyttöehtojen suunnittelussa myyntiä varten sekä operatiivisissa prosesseissa, kuten myyntitilanteissa, matkaoikeuden tarkastamisessa ja validoinnissa [28]. Kuvassa 3.3 oleva Fare Offer -osio käsittää suunnitteluvaiheessa muodostettavan datan, joka koostuu Fare Structure, Access right rules, Fare Products, Sales Packages -osista.



Kuva 3.3: Transmodel osa 5 määrittelemät osiot, tuotteiden myynnin suunnittelemiseksi sekä operatiivisten toimintojen mahdollistamiseksi [28].

Fare Structure -osassa määritetään tilaan tai aikaan liittyviä tietotyyppejä (`FARE STRUCTURE ELEMENT`), joiden avulla määritetään taksojen muodostumista. Tilaan perustuvat määrittelyt voivat sisältää esimerkiksi vyöhykkeiden tai pisteestä pisteeseen muodostuvan reitin osalta. Lisäksi voidaan määrittää myös tasataksa tai progressiivisia sääntöjä. Aikaan liittyvät tietotyypit määrittävät esimerkiksi kyseisen taksan voimassaolon eri vuorokauden aikoina.

Access right rules -osassa määritetään käyttöoikeussääntöjä (`CONTROLLABLE ELEMENT`), jotka määrittävät käyttöoikeuksia aikaisemmin mainittujen `FARE STRUCTURE ELEMENT` -tietotyyppien osalta. Yksittäinen tuote sisältää usein useampia `FARE STRUCTURE ELEMENT` -tietotyyppejä ja nämä yhdessä muodostavat `VALIDABLE ELEMENT` -tietotyypin, joka voidaan ajatella kokoelmana sallittavia vaihtoehtoja matkaoikeuden käyttämiseksi (esim. pienempi vyöhykevalinta lukijalaitteelta leimattaessa, vaikka

matkaoikeus olisi laajempi) [24]. Liitteessä E on esitetty tarkemmin näiden tietotyyppien suhdetta toisiinsa.

Fare Product -osa määrittelee FARE PRODUCT -tietotyypin, joka on immateriaalinen markkinoitavissa oleva tuote. Tähän tietotyyppiin vaikuttavat myös maksutapa ja mahdollisesti muut ehdot, jolloin FARE PRODUCT lopullinen hinta vaihtelee asiakkaasta riippuen.

Sales Packages -osa määrittää kullekin myyntikanavalle myyntiin tarjottavat tuotteet ja niiden myymiseen liittyvät ehdot. SALES OFFER PACKAGE on asiakkaalle ostettavaksi tarjottava tuote, josta muodostuu ostotilanteessa fyysinen TRAVEL DOCUMENT, joka sisältää ostetun matkaoikeuden tiedot CUSTOMER PURCHASE PACKAGE -elementissä. Kuvan 3.3 Sales Transaction -osion määrittelemänä voidaan mallintaa transaktio-tapahtumat ostotilanteesta.

Kuvan 3.3 ja Transmodelin viidennen osan Control- ja Validation-osiot määrittelevät operatiivisessä toiminnassa suoritettavien tarkastus- ja validointitapahtumien tietotyypit. Näitä tietotyyppejä voidaan hyödyntää erityisesti tapahtumalokien muodostamiseksi.

Matkustajainformaatio

Osa 6 sisältää matkustajainformaation esittämiseen liittyvät tietomallit, erityisesti kuluttajapalveluiden tueksi ja täydentää myös osan 4 reaaliaikainformaationa tarjottavaa tietoa. Osa 6 täydentää tietomallia asiakkaan matkan näkökulmasta ja esittelee TRIP -tietotyypin (matka) pohjalta tietomallin täydennykset matkustajan näkökulman esittämiseksi. Aikaisemmin esitetty ajoneuvon VEHICLE JOURNEY voi toteuttaa palvelunäkökulmasta SERVICE JOURNEY, joka vastaavasti on osa matkustajan matkan TRIP PATTERN:ia. Lisäksi osassa esitetään viitekehys asiakasinformaation esittämiseksi [29].

Kuljettajien hallinta

Osa 7 sisältää tietomallin kuljettajien työvuorojen suunnittelemiseksi, kuten auto- ja kuljettajakierrot, joiden pohjalta määritetään operatiivisen tuotannon suunnitelma PRODUCTION PLAN yhdessä suunniteltujen aikataulujen kanssa. Lisäksi osassa määritetään kuljettajien sekä ajoneuvojen operatiivisen toiminnan seuraamisessa käytettävät tietotyypit, kuten kirjautumistiedot lähdölle tai taukojen alkamis- ja päättymisajat [30].

Informaatio ja tilastojen hallinta

Osassa 8 täydennetään muiden osien tietomalleja joukkoliikenteen toteumatietojen tilastoimiseksi sekä joukkoliikenteen palvelutason kuvaamiseksi, kuten matkustajamäärätiedoilla toteutuneissa lähdöissä sekä tilastointia toteutuneessa palvelussa [30].

Vaihtoehtoiset kulkumuodot

Osa 10 sisältää täydennyksiä uusien multimodaalisten palveluiden sekä kulkumuotojen tukemiseksi olemassa olevissa osissa 1-8. Kuvassa 3.2 on esitetty tietomallistandardit, jotka implementoivat sekä pohjautuvat Transmodelin tiettyihin osioihin [25].

3.3 NeTEx

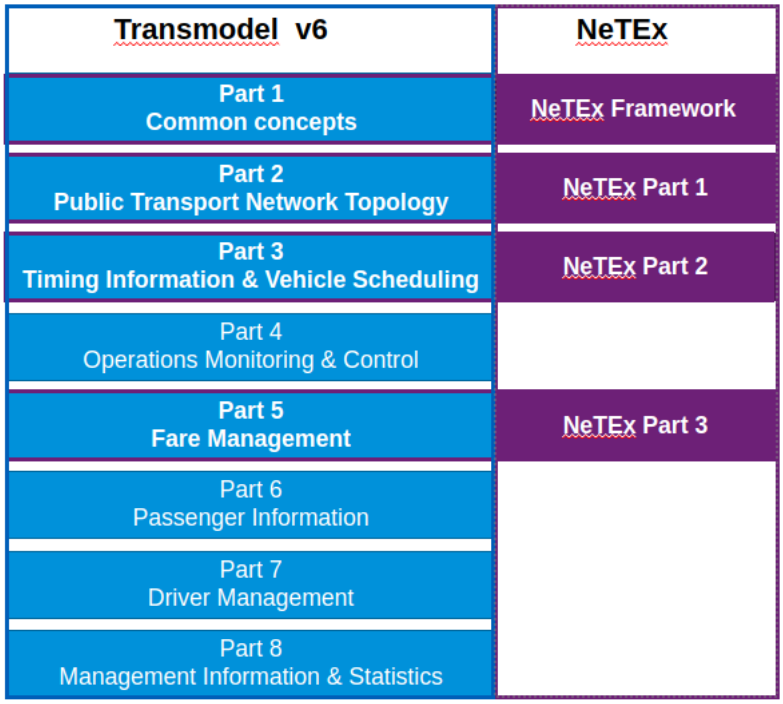
Viranomaisvetoinen NeTEx-standardi (Network Timetable Exchange) puolestaan on huomattavasti laajempi CEN-standardoitu tietomalli/implementaatio, joka perustuu Transmodel -viitekehykseen ja jakautuu kolmeen eri osaan [17, 19].

- NeTEx osa 1, liikenteen verkkotopologia, kuten reitit ja pysäkkien sijaintitiedot
- NeTEx osa 2, joukkoliikenteen aikataulutiedot, kuten vuorojen lähtöajat
- NeTEx osa 3, tuote- ja hintatiedot

NeTEx-standardi implementoi Transmodelin osat, kuten kuvissa 3.4 ja 3.5 esitetään. NeTEx sisältää tarkemmin määriteltynä tietotyyppien attribuutit sekä tietomallin skeeman, mikä mahdollistaa yhteensopivan tietomallin soveltamisen joukkoliikenteen tietojärjestelmissä [17].

NeTEx -standardi on huomattavasti GTFS:ää laajempi XML-pohjainen tietomalli, jonka tehtävänä on tarjota tietomalli koko EU:n laajuiseen joukkoliikennetiedon vaihtoon sisärajat ylittäen NeTEx:n avulla on mahdollista mallintaa joukkoliikenteen infrastruktuuri, palvelut sekä hinta- ja tuotetiedot monipuolisemmin kuin väljemmin määritelty GTFS-tietomalli [19, s.22, 127].

VTT toteaa, että NeTEx tietomallin keskeneräisyys sekä monimutkaisuus verrattuna yksinkertaisempaan GTFS:ään vaikeuttaa joukkoliikenteen tietomallien siir-

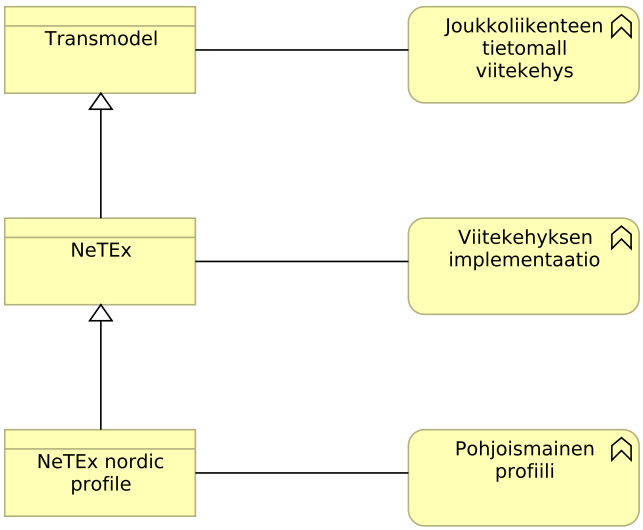


Kuva 3.4: NeTEx ja Transmodel -osien vastaavuudet toisiinsa.

tymistä NeTEx:iin [19, s. 62]. Kaupallisen GTFS-standardin vahva asema ja tuki eri palveluissa sekä yksinkertaisuus vaikeuttavat siirtymistä viranomaisvetoiseen NeTEx -tietomalliin. Kuitenkin EU-lainsäädännön tavoitteet sekä yhteensopivuus multimodaalisessa joukkoliikenteessä vaativat siirtymisen asteittain kohti NeTEx-mallia.

NeTEx osa 1, Verkkotopologia

NeTEx osa 1 sisältää mallin verkkotopologiasta sekä kaikissa NeTEx:n eri osissa käytettävien kehyksien määrittelyt [5]. Osan 1 määrittelyt jakautuvat Network Description, Fixed Object sekä Tactical Planning Components -malleihin. Network Description -malli koostuu mm. verkkoinfrastruktuurin, linjan, reitin kuvaavista tietotyypeistä. Fixed Object -malli sisältää vastaavasti eri tyyppisiä paikkoja kuvaavia tietotyyppejä, kuten alue, pysäkki, pysäköinti sekä erilaiset paikkojen varusteluun liittyvät tyypit. Tactical planning -mallissa on kuvataan tietotyypit erilaisista pisteiden ja niiden välisten linkkien muodostamista ketjuista, joita on esitetty erillisistä näkökulmista. Esimerkiksi matkaketju `JOURNEY PATTERN`.



Kuva 3.5: NeTEx-tietomalli implementoi Transmodel-viitekehyksen mukaisen mallin.

NeTEx osa 2, aikataulu- ja palvelutiedot

NeTEx osa 2 sisältää erilaisten matkojen sekä palveluiden kuvaavat mallit. Journey ja JourneyTimes -mallissa kuvataan matkan (`JOURNEY`) tietoja eri näkökulmista [6]. Esimerkiksi `VEHICLE JOURNEY` kuvaa ajoneuvon matkaa operatiivisessa liikenteessä. Vastaavasti `VEHICLE JOURNEY` toteuttaa liikennepalvelua kuvaavan `SERVICE JOURNEY`, joka mallintaa matkan toisesta näkökulmasta. Osa 2 sisältää myös matkaan eri elementtien tietomallit, kuten aikataulutietoja matkan pysäkkiketjulle (`JOURNEY`,`JOURNEY PATTERN`, `JOURNEY PATTERN RUN TIME`). Muita matkan (`JOURNEY`) liittyviä tietomalleja on vaihtoyhteyksien sekä matkan palveluihin liittyvien tietotyyppien kuvaamiseksi.

NeTEx osa 3, tuote- ja hintatiedot

NeTEx osa 3 [7] sisältää erilaisten joukkoliikenteen tuotteiden, hinnoittelun sekä myynnissä tarvittavan datan tietomallit. Esimerkiksi Fare Product -mallissa määritetään matkustusoikeuteen liittyviä tietotyyppejä, kuten myytävä immateriaalinen tuote (`FARE PRODUCT`), jonka aktivoituminen tai veloitushetki määritetään tietotyypillä `CHARGING MOMENT`. Veloitushetkellä tarkoitetaan esimerkiksi ennen tuotteen käyttöä tehtävää veloitusta tai käytön jälkeen tehtävää veloitusta. Tuotteen

(FARE PRODUCT) muodostama matkustusoikeus (ACCESS RIGHT IN PRODUCT) tarkistetaan kulkuneuvoon nousun yhteydessä esitetyn tunnisteen (VALIDABLE-ELEMENT avulla. NeTEx osa 3 Fare Zone -malli vastaavasti yhdistää taksatiedot infrastruktuurielementteihin, kuten aikataulutetun pysäkin (SCHEDULED STOP POINT) osalta FARE SCHEDULED STOP POINT yhdistää vyöhyketiedon kyseiseen pysäkkiin [16]. Tämä mahdollistaa asiakkaan tuotteen sisältämän matkustusoikeuden ehtojen tarkistamisen erilaisiin tekijöihin perustuen. Liitteen F kaavio havainnollistaa edellä mainittuja riippuvuussuhteita toisiinsa.

NeTEx:n periytymisestä ja suhteesta Transmodel-malliin

NeTEx-projektin tarkoituksena on ollut yhdistää aikaisemman Transmodel v5.1 sekä IFOPT -standardin (Identification of Fixed Objects in Public Transport) mukaisten tietomallien konseptit. Tämä mahdollistaa joukkoliikenteelle oleellisten infrastruktuuritietojen, kuten pysäkit ja POI-pisteet (Point Of Interest), saamisen osaksi joukkoliikenteen palvelutietojen malleja. NeTEx-standardin laajennettua käsittämään tuote- ja hintatietoja (osa 3), Transmodel-viitekehykseen sisällytettiin NeTEx:n muutokset konseptien harmonisoinnin vuoksi. Tuorein Transmodel-versio (v6) sisältää nämä muutokset [25]. Kuten kuvassa 3.2 on esitetty, NeTEx on rikastanut Transmodel-viitekehyksen tietomalleja tarvittavilta osin toimiakseen joukkoliikenteen datan rajapintastandardina [17].

4 Tietomallien linkittäminen toisiinsa

Tässä luvussa käsitellään tietomallien linkitysmenetelmiä ja havainnollistetaan niiden tarpeellisuus eheän tiedon tarjoamiseksi. Aliluvussa 4.1 käydään läpi, miksi linkitysmenetelmiä tarvitaan ja mikä merkitys on tietomallistandardeilla linkityksissä. Aliluku 4.2 esittelee erilaisia menetelmiä, jotka lähestyvät tiedon yhteensovittamiseen liittyviä haasteita erilaisista näkökulmista. Aliluvussa esitetyt linkitysmenetelmät on kehitetty erilaisista näkökulmista ja niiden ominaisuuksia on havainnollistettu joukkoliikennekontekstin osalta. Vastaavasti aliluvussa 4.3 on nostettu esille linkitysmenetelmien käytössä esiintyviä haasteita, joita tulisi huomioida onnistuneen lopputuloksen kannalta.

4.1 Linkitysmenetelmiä tarvitaan tietomallien ja datan yhteensovittamiseksi

Dimou et al. nostavat esille artikkelissaan [9, s.1], että tavoitteet datan avoimen jakamisen osalta aiheuttavat tilanteen, jossa useista lähteistä on saatavilla dataa eri implementaatioina sekä standardin mukaisina. Tämä aiheuttaa tilanteen, jossa eri lähteistä koostettu data sisältää duplikaatteja esimerkiksi tunnisteiden sekä tietosisältöjen osalta, kun lähteet ovat keränneet dataa eri näkökulmista sekä kontekstista. Mylonas et al. [15] toteavat artikkelissaan, että heterogeenisistä datalähteistä koostettu data sisältää paljon päällekkäisyyksiä ja voi muodostua haasteeksi datan yhdenmukaistamiselle. Bellini et al. korostavat artikkelissaan [1] avoimen joukkoliikennedatan suurimmaksi ongelmaksi sen, että data saadaan vaihtelevasti eri datalähteistä, jotka ovat hajallaan sekä tuottavat datan eri tietomalleihin pohjautuen ja erilaisissa serialisaatioissa. Myös Scrocca et al. [23] toteavat, että joukkoliikenteen tietomallit ovat olleet yleisesti ottaen heikkoja yhteensopivuudeltaan ja tämän vuoksi EU:n tavoitteiden (esimerkiksi asetus 2017/1926)[1] saavuttamiseksi on kehitetty Transmodel-standardi, joka toimii kattavana konseptuaalisena viitekehyksenä [23].

[1]Komission delegoitu asetus (EU) 2017/1926, annettu 31 päivänä toukokuuta 2017, Euroopan parlamentin ja neuvoston direktiivin 2010/40/EU täydentämisestä EU:n laajuisten multimodaalisten matkatietopalvelujen tarjoamisen osalta

Transmodel on erittäin laaja sekä kattava viitekehys, johon myös NeTEx-tietomalli pohjautuu.

Dimou et al. [9] toteavat Transmodel-viitekehyksen tarvetta puoltavasti, että edellä mainittujen ongelmien pienentämiseksi datan muodostamisessa tulisi hyödyntää olemassa olevia tunnisteita sekä noudattaa yhtäläisiä konsepteja tietomallien osalta, jotta vältetään datan duplikaatteja ja haasteita tiedon yhdistämisessä. Tämä korostaa tarvetta modulaarisen, yhteensopivuutta edistävän, helposti mukautettavan menetelmän kehittämiseksi, jotta tietovarantojen inkrementaalinen kehittäminen sekä koonti yhteensopivassa muodossa olisi mahdollista. Esimerkiksi Ruckhaus et al. [22] ovat nostaneet esille Ciudades Abiertas -projektin, jossa espanjalaiset kunnat sekä yksityiset toimijat muodostivat yleiset avoimen hallinnon periaatteet, jotka helpottavat olemassa olevien mallien uudelleenkäytetettävyyttä. Suomessa vastaavanlaista yhteistyötä toteutetaan matkatietotyöryhmän toimesta [19, s. 101-102]

Laaja kansainvälinen viitekehys sekä siitä johdetut tietomallit mahdollistavat yhteensopimattomien tietomallien linkittämisen Transmodel-pohjaiseen tietomalliin.Tämä tarjoaa välillisen linkityksen eri tietomallien osalta, mikä vähentää linkityksien määrän $O(n^2)$ luokasta luokkaan $O(n)$ [23]. Näin mahdollistetaan kattavien, useita liikkumismuotoja käsittävän palveluportfolion muodostuminen.

4.2 Linkitysmenetelmiä

Eri tietomallien muunnoksissa on toteutettu useita eri ratkaisuja, jotka ovat perustuneet ennalta määritettyyn serialisaatioon tai tietokantatyyppiin [23]. Vastaavan havainnon tietomallien linkitysmenetelmien kohdentumisesta eri serialisaatioihin tai tietokantamalleihin nostavat Dimou et al. artikkelissaan [9]. Lähteenä olevat artikkelit nostavat muutaman geneerisen menetelmän tietomallien ontologiseen linkittämiseen useista eri lähteistä.

Kohdan 4.2.1 menetelmä tarjoaa joukkoliikenteen domain-asiantuntijoille menetelmän muodostaa tietomallien linkittämisen taulukon avulla, mikä toimii aineistona koneelliselle linkittämiselle sekä datakonversioiden muodostamiselle [3]. Edellä mainitussa menetelmässä keskitytään menetelmän käyttöön sekä hyödyntämiseen käytännön läheisten käyttötapauksien avulla. Vastaavalla tavalla kohdassa 4.2.2 esitetään LOT-prosessimalli [22] tietomallien kehittämiseksi ja eri vaiheissa vaadittavat asiantuntijat, datan sekä työkalut datan käsittelemiseksi. LOT-menetelmä esitellään yleisenä menetelmänä, joka ei ota kantaa käsiteltävään domainiin. Kaksi muuta esi-

teltävää menetelmää, RML [9] ja Chimera [23], lähestyvät tietomallien linkittämistä teknisestä näkökulmasta ja niissä esitetään myös ohjelmallisia ratkaisuja tietomallien linkittämiseksi. Chimera kykenee esitellyistä menetelmistä muodostamaan myös kohdemallin dataa osana menetelmää, kun muut esiteltävät menetelmät keskittyvät ontologiamallien määrittelydokumenttien tai aineiston muodostamiseen.

4.2.1 Linkitystaulukko

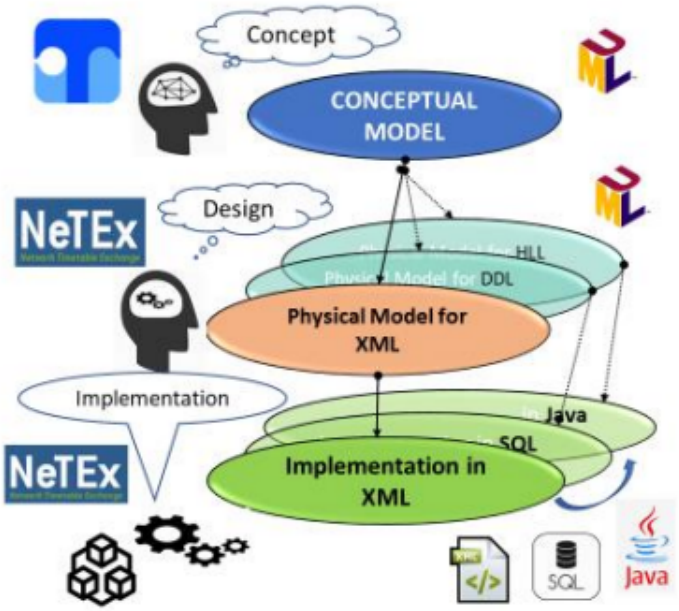
Data4PT-projektissa kartoitettiin formaalia menetelmää eri tietomallien sekä tietotyyppien linkittämiseksi [3]. Projektin työryhmä on tunnistanut, että MMTIS-asetuksen[2] vaatima yhdenmukaisen staattisen sekä dynaamisen datan jakaminen yhdenmukaisilla tietomalleilla ja -tyyppeillä määritetään XSD-määrittelydokumenteilla tai UML-malleilla. Laajojen tietoryhmien yhteensovittamiseksi sekä jakamiseksi tarvitaan menetelmä eri asiantuntijoiden yhteistyön mahdollistamiseksi, mihin Linkitystaulukon käytöllä tavoitellaan.

Työmenetelmä perustuu datakategorioiden tunnistamiseen, päällekkäisyyksien käsittelyyn viitemallien sekä tukevien tietomallien avulla ja viimeisenä datakategorioiden sisällä olevien tietotyyppien linkittämiseen toisiinsa. Beurèe ja Knowles toteavat, tietomallien ja -tyyppien linkitystä tarvitaan esimerkiksi saman kategorian tietotyyppejä kuvaavan kahden vastaavan tietomallin yhteensovittamiseksi. Lisäksi kahden tietomallin tietotyyppien linkittämisellä saadaan tuotettua määrittelyt esimerkiksi datan konversiotyökaluille. Kolmas käyttötapaus liittyy kahden eri tietomallin väliseen integraatioon, jossa on tarpeen löytää selkeä ja tehokas rajapinta kahden tietomallin välille johonkin rajattua käyttötarvetta varten ja tietomallia on tarpeen rikastaa sopivan esitystavan varmistamiseksi.

Kuten kuvassa 4.1 ja Transmodel- sekä NeTEx-tietomallien suhteesta on esitetty, Transmodel toimii konseptuaalisena viitekehyksenä ja auttaa UML-kaavioiden avulla hahmottamaan eri datakategorioita. NeTEx tarjoaa tarkempien XSD-dokumenttien ja attribuuttien avulla kyvykkyyden tehdä tarkempia linkityksiä tietotyyppien välillä sekä mahdollistaa toteutettavien tietokantaskeemojen johtamisen tietotyypeistä.

Tietomallien linkittämisessä tulee huomioida, että linkittäminen toteutetaan saman tason malleilla (vrt. kuvan 4.1 eri tasot) ja linkittämisessä tulisi välttää imple-

[2]Komission delegoitu asetus (EU) 2017/1926, annettu 31 päivänä toukokuuta 2017, Euroopan parlamentin ja neuvoston direktiivin 2010/40/EU täydentämisestä EU:n laajuisten multimodaalisten matkatietopalvelujen tarjoamisen osalta



Kuva 4.1: Tietomallien linkittäminen konseptista ja suunnittelusta implementaatioon [3].

mentaatiotason dataan ja mallinnukseen mahdollisesti liittyviä metatietoja. Tietomallien linkittäminen jakaantuu Beurèe ja Knowles esittämässä mallissa kolmeen vaiheeseen [3]:

1. Vertailtavien tietomallien datakategorioiden tunnistaminen sekä epämuodollinen ylätason termien linkittäminen toisiinsa.

2. Systemaattinen ylätason konseptuaalisten mallien semanttinen vertailu. Myös vertailtavien mallien ala-mallit huomiodaan semanttisessa määrittelyssä.

3. Tarkalla tasolla tehtävä yksittäisten tietotyyppien sekä attribuuttien vertailu linkitystaulukon avulla. Lähde- ja kohdemallin tarkat relaatiot.

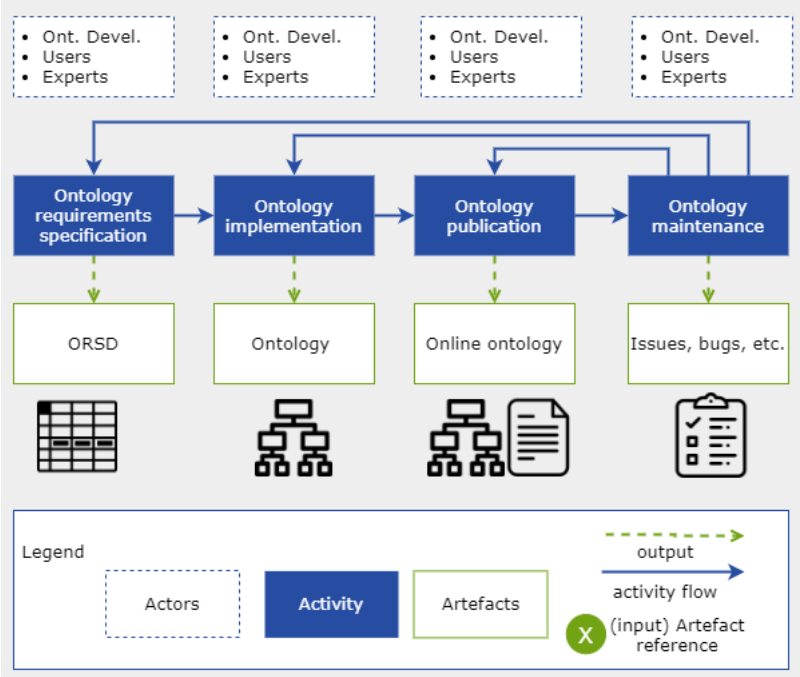
Liitteessä G on esitetty kohdassa 3 käytetyn linkitystaulukon sisältö.

4.2.2 LOT

Linked Open Terms (LOT) -menetelmä soveltuu tietomallien ontologioiden sekä sanastojen kehittämiseen [22]. Menetelmä sisältää määrittelyvaiheen, implementaa-

tiovaiheen, julkaisu- sekä ylläpitovaiheet, kuten kuvassa 4.2 on esitetty sinisissä laatikoissa.

Ensimmäisessä vaiheessa tarkastellaan tarkastelun kohteena olevien tietomallien suhdetta linkitettävään tietomalliin vertailemalla standardeja, käyttötapauksia sekä formaatteja, joista saadaan muodostettua substanssiosaajien ja käyttäjien avulla linkityksen onnistumisen validoinnissa käytettäviä kyselyitä. Ensimmäisen vaiheen lopputuotoksena on ORSD-dokumentti (Ontology Requirement Specification Document), jota tarvitaan ontologioiden linkittämisessä. Ensimmäisen vaiheen jälkeen implementaatiovaiheessa muodostetaan eri ontologiamallien linkityksiä, jotka soveltuvat kohteena olevan tietojoukon riippuvuuksien tunnistamiseen. Lähdetietomalleista luodaan RDF-dokumentit, joiden avulla lähteen tietotyypit on linkitetty referenssinä toimivaan tietomalliin. SPARQL-kyselykielellä suoritetaan ensimmäisessä vaiheessa muodostettuja kyselyitä RDF-aineistoa vasten, millä varmistetaan tietomallien linkityksien semanttinen sekä ontologinen yhdenmukaisuus. Julkaisuvaiheessa julkaistaan luettava dokumentaatio, jossa on kuvattuna mm. metadata, tietomallin kaaviot, luokkakuvaukset sekä mahdolliset rajoitukset niiden suhteissa. Ylläpitovaiheessa vastaava edellämainittu prosessi jatkuu iteratiivisesti päivittäen ja täydentäen tietomallia [22].

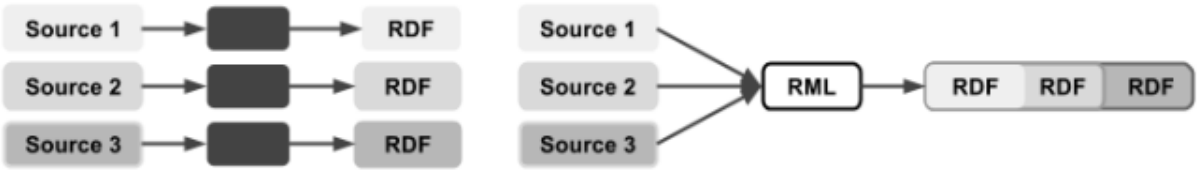


Kuva 4.2: LOT-menetelmän päävaiheet sekä iteratiivinen, jatkuva prosessi [18].

4.2.3 RML

RDF mapping language (RML) on geneerinen menetelmä, jonka avulla on mahdollista muodostaa tapauskohtaisesti heterogeenisten tietomallien sekä erilaista datan serialisaatiota hyödyntävien tietomallien välisiä linkityksiä [9]. RML pohjautuu W3C-standardoituun linkitysmenetelmään R2RML:ään ja täydentää sekä mahdollistaa tietomallien linkityksen toteuttamisen monipuolisemmin.

RML-menetelmässä lähdedatasta muodostetaan RDF-dokumentti (Resource Description Framework), jossa lähdedatan tietueet linkitetään viitekehyksen semanttisiin tietotyyppeihin. RML-menetelmällä saadaan linkitettyä eri lähteistä, mahdollisesti eri serialisoinnilla olevaa dataa toisiinsa RDF-dokumentissa toteutettujen linkityksien pohjalta [9]. Menetelmä hyödyntää R2RML-menetelmässä muodostettua Triples Graph -mallia, jossa tietotyypit linkitetään subjektin, predikaatin ja objektin avulla toisiinsa. Annettavan lähdedatan lisäksi RML-menetelmän prosessi vaatii määrittelydokumentissa kuvatut säännöt (Triples Graph), joilla linkittäminen toteutetaan viitteenä olevaan tietomalliin [9]. Kuvassa 4.3 on havainnollistettu, miten RML-menetelmän prosessi muodostaa lähdeaineistojen välillä linkitykset eri tietotyyppien osalta.



Kuva 4.3: Datalähteiden linkittäminen toisiinsa ilman sekä RML-menetelmää käyttäen [9].

RML-menetelmän avulla kyetään linkittämään dataa joka on tarjolla erilaisissa serialisaatiossa, kuten CSV, JSON, XML. Lähdedatan lisäksi vaaditaan linkityssäännöt jokaisen lähdedatan osalta. Liitteessä H on esimerkki RML-prosessin vaatimasta määrittelydokumentista. Määrittelydokumentin Logical Source -osiossa (`rml:logicalSource`) määritetään lähdedataan liittyviä tietoja, kuten lähdeosoite, serialisaatio, iteraattori sekä lähdedatan osalta tietotyyppien tunnisteet. Iteraattorin (`rml:iterator`) avulla määritetään tarkasteltavan datan osalta, millä tasolla linkittäminen toteutetaan jokaisen tietotyypin osalta. Lähdedatan osalta `rml:source` sekä `rml:iterator` ovat pakollisia [20]. Lähdedatan tietotyyppien tunnisteet osoitetaan `rr:subjectMap`

avulla. `rr:predicateObjectMap` -osassa määritetään lähdedatan tietotyyppien predikaatti (`rr:predicate`) sekä referenssimallin mukainen tietue. `rr:predicate-ObjectMap` sisältää myös mahdolliset RML-prosessissa tehtävät toimenpiteet, kuten esimerkiksi liitos-operaatio `rr:joinCondition`.

RML-määrittelydokumentin avulla RML-prosessori noutaa määrittelydokumentissa esitetyistä lähteistä datan ja muodostaa lähdedatan serialisaation mukaisesti oman aliprosessin. Aliprosessit käyvät silmukassa läpi määrittelydokumentissa esitetyn datalähteen iteraattorin osoittamat tietueet ja linkittävät ne määrittelydokumentin mukaisesti predikaattiin sekä objektiin. Lopputuloksena muodostuu RDF:n (Resource Description Framework) mukainen dokumentti, jossa on määrittelydokumentin mukaiset linkitykset eri tietomallien välillä [20]. RML-prosessorin eri komponentit ovat modulaarisia, joten tapauksesta riippuen tietotyyppien linkitys voidaan tehdä datalähteen tai määrittelydokumentin määrittämässä järjestyksessä [9, 20].

4.2.4 Chimera

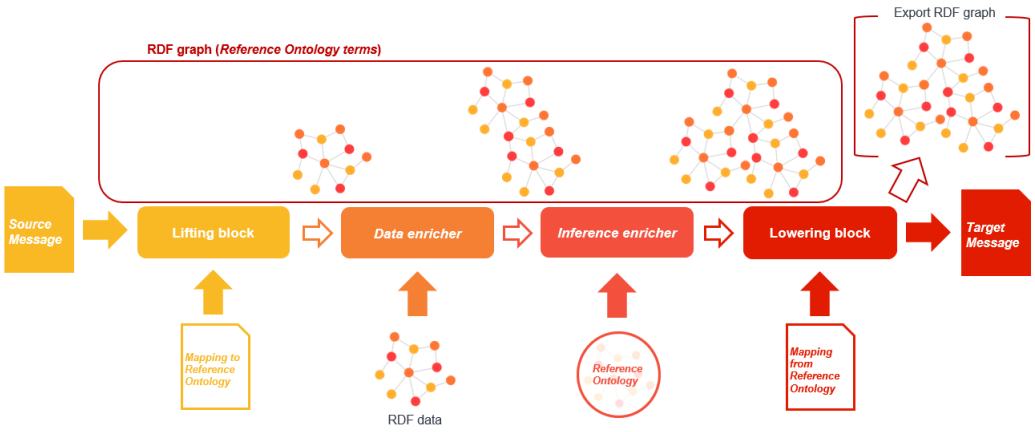
Chimera on ohjelmistokehityksen viitekehys, jonka tavoitteena on tarjota keskitetty, useasta tietomallista yhteen kohdemalliin, linkityksiä toteuttava ratkaisu [23]. Chimeran tarkoituksena on tarjota lähde- ja kohdemallien välinen dataputki, joka kykenee muodostamaan lähdemallien tietotyypeistä linkitetyn tietokannan, josta voidaan tapauskohtaisesti muodostaa kohdemallin mukainen data halutulla serialisaatiolla. Scrocca et al. painottavat [23], että Chimeran kehityksessa on käytetty inspiraationa Hohpe et al. Enterprise Integration Patterns -kirjan [11] mukaisia menetelmiä modulaarisesta integraatioputkesta ja erityisesti siinä esitetyä "a Message Translator system" -menetelmää.

Chimeran tietomallikonversiot alkavat lähdemallin tietotyyppien linkittämisestä referenssitietomalliin RDF-dokumentiksi, jonka jälkeen RDF-dokumentaatiota rikastetaan muiden lähdemallien tietotyypeillä (lifting). Viimeisessä vaiheessa RDF-dokumentaatioita vasten voidaan suorittaa kyselyitä oikeiden riippuvuuksien muodostamiseksi sekä tuottaa halutussa formaatissa dataa ulos (Lowering).

Kuva 4.4 havainnollistaa Chimeran eri vaiheet, joista ensimmäisessä muuntovaiheessa (Lifting block) lähdedata vastaanotetaan sekä referenssinä toimivan tietomallin avulla tietotyypit linkitetään RDF-dokumentiksi. Toisessa vaiheessa vastaanotetusta lähdedatasta muodostettua RDF-dokumenttia voidaan rikastaa muilla ontologia-aineistoilla tietotyyppien suhteiden monipuolistamiseksi (Data enric-

her block). Rikastaminen voi olla tarpeen, jos vastaanotetun datan ja siitä muodostetun RDF-dokumentin pohjalta ei ole mahdollista tuottaa vaadittavaa kohdedataa. Kolmannen vaiheen päättelyprosessissa (Inference enricher) jo muodostettujen ontologia-aineistojen avulla sekä niistä löydettyjen sääntöjen avulla voidaan rikastaa RDF-dokumenttia lisäinformaatiolla. Viimeisessä vaiheessa Chimera suorittaa kyselyitä RDF-graafia kohden ja muodostaa tietomallien linkityksien avulla kohdedatan tarvittavassa serialisaatiossa.

Chimera tukee suuren datamäärän prosessointia eräajona sekä yksittäisten datasanomien muuntamista datavirtana, sillä se hyödyntää Apache Camel[3] -alustaa [23]. Chimera on kehitetty multimodaalisen joukkoliikenteen tietomallien keskinäiseen linkittämiseen, minkä takia Chimera pystyy tuottamaan RDF-graafin, jonka pohjalta ontologia-aineistoja on mahdollista rikastaa iteratiivisesti sekä jo tuotettu graafi on hyödynnettävissä myöhemmin. Scrocca et al. [23] toteavat ratkaisunsa hyvänä puolena sen, että RDF-aineiston mukainen ontologiamalli kehittyy jatkuvasti paremmaksi ja sitä voidaan käyttää uudelleen.



Kuva 4.4: Chimera-menetelmän eri vaiheet [4].

4.3 Haasteet / kriittiset menestystekijät

Scrocca et al. [23] toteavat, että vaikka eri toimijoiden data olisikin yhteensopivassa NeTEx-tietomallissa, datan integrointi ja uudelleenkäytettävyys ei ole helppoa, sillä eri tietotyyppien tulkinnoissa on eroja lähteiden välillä. Lisäksi he nostavat esil-

[3]https://camel.apache.org/

le, että vaikka olisi hyviä, koneellisesti luettavia määrittelydokumentteja, ei näiden muodostaminen ole välttämättä helppoa domain-osaajalle nykyisin käytössä olevilla työkaluilla. Myös Dimou et al. [9] nostavat esille, että tietomallien geneeriset linkitysmenetelmät eivät välttämättä kykene käsittelemään kaikkia syötteenä saatavaa lähdedataa. Ruckhaus et al. [22] nostavat vastaavasti esille, että tietomallien linkittämisen yleisemmin esiintyviä haasteita ovat eri versioita sisältävät, keskeneräiset tai merkitykseltään implisiittiset referenssitietomallit, joita vasten linkityksiä tehdään.

5 Tutkimusvaihe

Luvussa käsitellään tutkimuksen toteutusvaiheeseen liittyviä asioita. Aliluvussa 5.1 esitetään tutkimuksen tarkoitus ja tutkimuskysymykset. Erityisesti havainnollistetaan joukkoliikenteen toimintakentän muutoksista johdettavia motivaatiotekijöitä, joiden pohjalta olen tunnistanut tutkimuskysymyksiä. Aliluvussa 5.2 käydään läpi tutkimusasetelmaan liittyvia asioita, kuten tutkimusmenetelmä, tiedonkeruumenetelmät ja datan analyysi. Tämä havainnollistaa teoriaa toteutettavan tutkimuksen sovellettavista menetelmistä ja eri vaiheista. Aliluvussa 5.3 kuvataan tutkimuksen eteneminen tutkijan näkökulmasta esitellen sidosryhmät, eri vaiheet, valitut linkitysmenetelmät ja kuvataan tutkimuksen eteneminen.

5.1 Tutkimuksen tarkoitus

Matkatietoon perustuvien palveluiden tarjoajat tarvitsevat laadukasta dataa, jossa tietosisältöön ja yhdisteltävyyden säännöt ovat määritelty eri toimijoiden välillä kansallisesti. Vaikka kansallisesti olisikin määrätty datan jakaminen tietyn standardin muotoisena, väljästi määritetyn standardin mukainen implementaatio voi vaihdella paljon, jolloin näennäisesti yhteensopiva data onkin yhteensopimatonta [19]. VTT:n tutkimusraportin mukaan koostetun matkatiedon tulisi olla laadukasta ja sisältää seuraavia ominaisuuksia:

- Semanttinen yhteensopivuus, jolloin datan hyödyntäjät voivat automaattisesti hyödyntää tarjottua dataa helposti
- Yhdenmukaisuus, jotta eri lähteistä tuleva data on yhdenmukaista
- Tarjotun datan oikeellisuus sekä täsmällisyys, jotta se luotettavaa hyödynnettäväksi
- Täydellisyys, jotta tarjottu data sisältää kaiken olennaisen tiedon hyödyntäjille

VTT:n raportin mukaan olennainen tarve olisi koostaa data eri lähteistä sekä kohdentaa resursseja datan laadunvarmistamiseen. Esimerkiksi nykyisin tarjolla olevaa pysäkkidataa on tarjolla eri rekistereistä toimijoittain ja eri pysäkkirekisterien

data eroaa toisistaan mm. nimeämiskäytänteiden tai koordinaattipisteiden osalta, kun eri rekisterit on toteutettu palvelemaan eri toimijoiden omia tavoitteita sekä näkökulmia. Lisäksi eri matkatiedon tuottamiseen käytettävät järjestelmät ovat heterogeenisiä ja niiden yhteentoimivuudessa on haasteita sekä prosessien kehittämistarpeita. Raportissa [19] esille nostettujen eri toimijoiden tavoitteet vaativat laadukkaan, semanttisesti yhteensopivan ja täydellisen datan tarjotakseen laadukkaita palveluita eri sidosryhmille.

Eri toimijoiden tavoitteet asettavat datan tuottamiselle subjektiivisia vaatimuksia omien tavoitteiden saavuttamiseksi ja samalla vaatimukset jaetun datan tuottamiseksi kasvaa. Näiden vaatimuksien täyttämiseksi tarvitaan menetelmä, jolla toimija kykenee tuottamaan omaan käyttöön parhaiten soveltuvaa dataa. Lisäksi mahdollisimman eksplisiittisellä menetelmällä voidaan varmistaa datan yhdistettävyys sekä ontologinen yhteensopivuus referenssitietomallin tietotyyppeihin.

Liitteessä I on esitetty tutkimusvaiheessa tarkasteltavien menetelmien rooli osana tietomallien ja -tyyppien linkittämistä referenssimallin dataan. VTT:n raportissa nostettujen laatuvaatimuksien lisäksi on merkityksellistä, miten laadukkaan kohdemallin mukaisen datan tuottamiseksi vaadittavia määrittelydokumentteja on mahdollista muodostaa domain-asiantuntijan toimesta käytössä olevilla ohjelmistoilla. Tutkimuksessa on tarkoitus vastata seuraaviin kysymyksiin:

- Miten pienen perehtymisen jälkeen domain-osaajien on mahdollista tuottaa linkittämiseen vaadittava määrittelydokumentti?
- Miten paljon teknistä osaamista määrittelydokumentin muodostaminen vaatii domain-osaajalta?
- Mikä on linkityksessä tuotetun lopputuloksen (määrittelydokumentti) jatkohyödynnettävyys toisissa linkityksissä?
- Mitkä tekijät määrittelydokumenteissa vaikuttavat yhdistetyn datan laatuun?

Luvussa 4 esitetyt menetelmät on kehitetty erilaisiin tarpeisiin ja erilaisista lähtökohdista. Osa malleista, kuten Beurèe ja Knowlesin esittelemä linkitystaulukko [3] sekä LOT-prosessi tuottavat määrittelydokumentin, joka toimii esimerkiksi integraatio-ohjelmiston määrittelynä. Näiden menetelmien muodostamia määrittelydokumentteja voidaan hyödyntää esimerkiksi RML-menetelmässä tai Chimerassa. Tutkimuskysymyksien kannalta on olennaista keskittyä eri tietomalleihin perustuvan datan

yhdistämisessä vaadittavan määrittelydokumenttien vertailemiseen sekä hyödynnettävyyteen eri sidosryhmien näkökulmasta.

5.2 Tutkimusasetelma

Tässä aliluvussa käymme läpi tutkimusmenetelmään, tiedonkeruumenetelmiin ja datan analyysiin liittyviä asioita. Käymme läpi tapaustutkimukselle olennaiset asiat. Kohdassa 5.2.1 esitetään tutkimusmenetelmän teoriaa laadullisen ja tapaustutkimuksen näkökulmasta. Vastaavasti kohdassa 5.2.2 esitetään menetelmiä, joita hyödynnetään myös tässä tutkimuksessa. Viimeisessä aliluvun kohdassa 5.2.3 kuvataan tapoja, joilla tutkimusaineistoa voidaan käsitellä ja jalostaa analyysia varten.

5.2.1 Tutkimusmenetelmä

Tutkimukset voidaan jakaa kahteen pääkategoriaan määrälliseen (kvantitatiivinen) sekä laadulliseen (kvalitatiivinen) tutkimukseen. Määrällinen tutkimus perustuu tutkimuksen kohteen kuvaamiseen ja tarkasteluun tilastollisten ja laskennallisten analyysimenetelmien avulla. Laadullisessa tutkimusmenetelmässä kohdetta tarkastellaan empiirisiin aineistoihin sekä erilaisien analyysimenetelmien avulla [32]. Hanna Vilkka toteaa kirjassaan [31], että laadullisen tutkimuksen erityispiirteenä on, ettei sen tavoitteena ole löytää yksiselitteistä totuutta. Tutkimuksen aikana tehtävän aineistoanalyysin pohjalta ratkaistaan asioita, jotka eivät ole välittömän havainnon näköpiirissä.

Tutkimuksen lähtökohtana voi olla johonkin teoriaan pohjautuva lähtökohta (deduktiivinen) tai vastaavasti aineistovetoinen (induktiivinen) lähtökohta. Usein aineistopohjainen lähtökohta tutkimukseen yhdistetään laadullisten tutkimuksien lähtökohdaksi, sillä se lähtee liikkeelle aineistosta, johon kohdistetaan erilaisia laadullisen tutkimuksen ja analyysimenetelmien teorioita [32].

Laadullisen tutkimuksen alalajeja on lukuisia ja Creswell toteaa kirjassaan [8] seuraavien alalajien toistuvan vuosien saatossa:

- Narratiivinen tutkimus, jossa yksilön kertomusta käytetään aineistona
- Etnografinen tutkimus, jossa tutkimuksen kohdetta tarkastellaan kulttuurisista sekä ympäristötekijöiden näkökulmista
- Grounded theory -tutkimus, jossa useamman henkilön kokemuksista pyritään

luomaan teoria

- Fenomenologinen tutkimus, jossa usean henkilön kokemia ilmiöitä käytetään aineistona
- Tapaustutkimus, jossa tutkitaan reaalimaailman tapausta

Toisin kuin määrällisessä tutkimuksessa tapaustutkimuksessa kohteena on tapahtumakulku tai ilmiö. Tapaustutkimus perustuu valitun kohdetapauksen kokonaisvaltaiseen perehtymiseen ja siinä usein yhdistetään useita eri aineistoja, kuten kirjallisuus, haastattelut ja havainnointi. Näiden avulla rajattuun kontekstiin liittyvää tapausta voidaan analysoida eri näkökulmista ja tarkastella lopputulosta tapauksen lähtötietojen sekä olosuhteiden muodostamana [32, 12]. Tapaustutkimus yrittää vastata kysymyksiin *miten* ja *miksi* ja sen avulla on tarkoitus selvittää, mitä voimme oppia tapauksesta. Se soveltuu erittäin hyvin monimutkaisten ja pitkäkestoisten tapauksien tutkimiseen ja olosuhteiden vaikutuksiin lopputuloksen osalta. Vaikka tapaustutkimus käsittelee tarkasti rajattua kontekstia, mahdollistaa se tapausta analysoimalla yleistettäviä lainalaisuuksia ja auttaa tunnistamaan niihin liittyviä tekijöitä [12].

Kuten Jokinen on nostanut esille artikkelissaan [12], että tapaustutkimuksessa on erilaisia paradigmoja, joiden näkökulmasta tutkimusta toteutetaan. Näkökulmia ovat mm. faktanäkökulma, kokemusnäkökulma sekä konstruktionistinen näkökulma [32].

5.2.2 Tiedonkeruumenetelmät

Tiedonkeruumenetelmien avulla muodostetaan tutkimuksen analyysia varten aineisto, joita voi olla useita erilaisia laadullisessa tutkimuksessa [32]. Laadullisessa tutkimuksessa voidaan hyödyntää aineistoina esimerkiksi haastatteluita, havainnointiaineistoja ja erilaisia dokumentteja. Haastattelut jaetaan strukturoituihin sekä semi-strukturoituihin haastatteluihin, joilla vaikutetaan siihen, kuinka paljon haastattelussa on tilaa ja vapautta poiketa varsinaisesta haastattelurungosta tai kuinka organisoituna haastattelu halutaan pitää [32]. Haastatteluita on monentyyppisiä, kuten teemahaastattelut, asiantuntijahaastattelut ja useita muita erityistilanteisiin soveltuvia haastattelumuotoja. Esimerkiksi teemahaastattelussa, joka on yksi yleisimmistä haastattelumuodoista, ei ole tarkasti sidottua haastattelurunkoa ja haastattelija voi soveltaa tai esittää kysymyksiä eri tavalla tilanteen ja haastateltavan

mukaan. Haastattelu ei aina etene alkuperäisen suunnitelman mukaan, mutta aina tuottaa aineistoa tutkimuksen analyysia varten [32, 31].

Tapaustutkimuksessa hyödynnetään usein sekä laadullisen sekä määrällisen tutkimuksen tiedonkeruumenetelmiä. Tapaustutkimuksessa voidaan hyödyntää valmiita, jo olemassa olevia aineistoja luonnollisina aineistoina. Luonnolliset aineistot ovat syntyneet ilman tutkijan omaa panosta [32]. Jokinen et al.[12] esittävät, että tapaustutkimuksen kohteeseen sovellettavista menetelmistä voidaan soveltaa triangulaation ideaa, jonka alkuperä on navigoinnissa ja maastomittauksessa. Triangulaation avulla suhteellinen sijainti voidaan selvittää kahden pisteen muodostaman kulman avulla. Vastaavasti tapaustutkimuksessa voidaan aineisto-, teoria-, menetelmäsekä tutkijatriangulaatiolla varmistaa eri näkökulmista tutkimuksen tuloksien varmuutta ja niistä tehtäviä johtopäätöksiä. Tällä tarkoitetaan sitä, että tutkija kykenee varmistamaan edellä mainittujen triangulaatioiden avulla, että tuloksien luotettavuuden varmistamiseksi aineistoa ja menetelmiä on hyödynnetty useista näkökulmista.

5.2.3 Datan analyysi

Tutkimuksessa käytettävän aineiston ja datan analyysi on monisyinen, iteratiivinen prosessi, johon ei ole yhtä oikeaa toteutustapaa [32]. Analyysissa tavoitteena on muodostaa aineiston, teorian sekä oman ajattelun avulla suurempaa informaatioarvoa tutkimustuloksien sekä johtopäätöksien tueksi. Laadullisessa analyysissa voidaan soveltaa erilaisia menetelmiä, kuten sisällönanalyysin muotoihin kuuluvia koodaamista, teemoittelua sekä tyypittelyä. Näiden lisäksi on monia muita datan analyysiin soveltuvia menetelmiä, jotka soveltuvat erilaisiin analyysin lähestymistapoihin.

Datan analyysi aloitetaan yleensä aineistoon tutustumalla sekä litteroimalla aineisto kirjalliseen muotoon, jotta sen analysointi ja tarkastelu on helpompaa [32]. Litteroinnilla tarkoitetaan esimerkiksi puhemuotoisen aineiston muuntamista tekstimuotoon. Tällä varmistetaan, että aineisto on tutkittavassa muodossa.

Litteroinnin jälkeen tutkija muodostaa kokonaiskäsityksen aineistosta ja edellä mainittuja menetelmiä hyödyntäen analysoi aineistoa ja löytäen erilaisia asioita, jotka vastaavat tutkimuskysymyksiin. Esimerkiksi aineiston koodaamisella eri aineiston osia yhdistellään tai muunnellaan eri tekijöiden perusteella. Vastaavasti tyypittelyssä aineistosta voidaan erotella erilaisia ilmiöitä ja niihin liittyviä aineiston osia. Teemoittelussa vastaavasti eri tutkimusongelmaan vastaavat ja soveltuvat teemat

erotetaan ja aineiston osia koostetaan teemojen alle. Analyysi ei itsessään nosta esille vastauksia tutkimuskysymyksiin, vaan auttaa jäsentämään aineistoa siten, että tutkijan on helpompaa löytää vastauksia tutkimuskysymyksiin.

5.3 Tutkimuksen toteutus

Tutkimuskysymyksiin vastaamiseksi vertailemme muutamaa valittua mallia tapaustutkimusta käyttäen. Eri linkitysmenetelmät ovat tapauksia, joissa eri tietomallien linkityksiä voidaan toteuttaa ja tarkastella niiden kehitystyön motivaationa toimivia reaalimaailman ongelmia sekä niiden hyödynnettävyyteen liittyviä tekijöitä. Tässä työssä tarkoituksena on tarkastella eri näkökulmista linkitysmenetelmien empiirisiä kokeiluja, joita tehdään yhdessä domain-asiantuntijoiden sekä integraattoreiden kanssa sekä kartoittaa haastattelututkimuksen avulla eri sidosryhmien edustajien näkemyksiä sekä kokemuksia teoriaosuudessa esitettyjen linkitysmenetelmien soveltuvuudesta joukkoliikenteen tietomallien linkittämiseksi toisiinsa. Samalla kartoitetaan muita olosuhdetekijöitä, jotka vaikuttavat tietomallien linkittämiseen sekä niiden toimivuuteen käytännössä. Haastattelutuloksia voidaan reflektoida teoreettiseen aineistoon ja muodostaa uusia havaintoja, jotka antavat vastauksia tutkimusongelmaan sekä -kysymyksiin.

Tämän työn tapaustutkimuksessa on hieman kokemusnäkökulman sekä suurilta osin konstruktionistisen näkökulman piirteitä. Oman taustani puolesta minulla on myös subjektiivista kokemusta työssä käsitellyistä asioista ja ilmiöistä, joten osittain tunnistan haastateltavien kokemuksia sekä niiden vaikutuksia eri ilmiöihin. Toisaalta lähestyn asiaa vahvasti konstruktionistisesta näkökulmasta, sillä haluan tunnistaa linkitysmenetelmien sekä niiden käyttöön liittyviä tekijöitä ja ajattelen kaiken olevan osa kokonaisuutta, joka muodostuu eri tekijöiden yhteistuloksena.

Tapaustutkimuksen suurin riski tämän työn toteuttamisen kannalta on tutkimusvaiheeseen osallistuvien asiantuntijoiden sitouttaminen linkitysmenetelmien käyttöön. Riskiä voidaan pienentää tutkimuksen tavoitteilla ja taustoittamalla tutkimutyön pohjalta tehtäviä potentiaalisia johtopäätöksiä käytettävien asiantuntijoiden reaalimaailman työn helpottamiseksi.

Kohdassa 5.3.1 esitetään taustaa tutkimukseen valituista sidosryhmistä. Kohta 5.3.2 käy läpi tutkimusprosessin vaiheistuksen ja käytettävät linkitysmenetelmät ovat esitetty kohdassa 5.3.3. Vastaavasti kohdissa 5.3.4 ja 5.3.5 on semi-strukturoidun haastattelun kysymysrungot.

5.3.1 Sidosryhmät

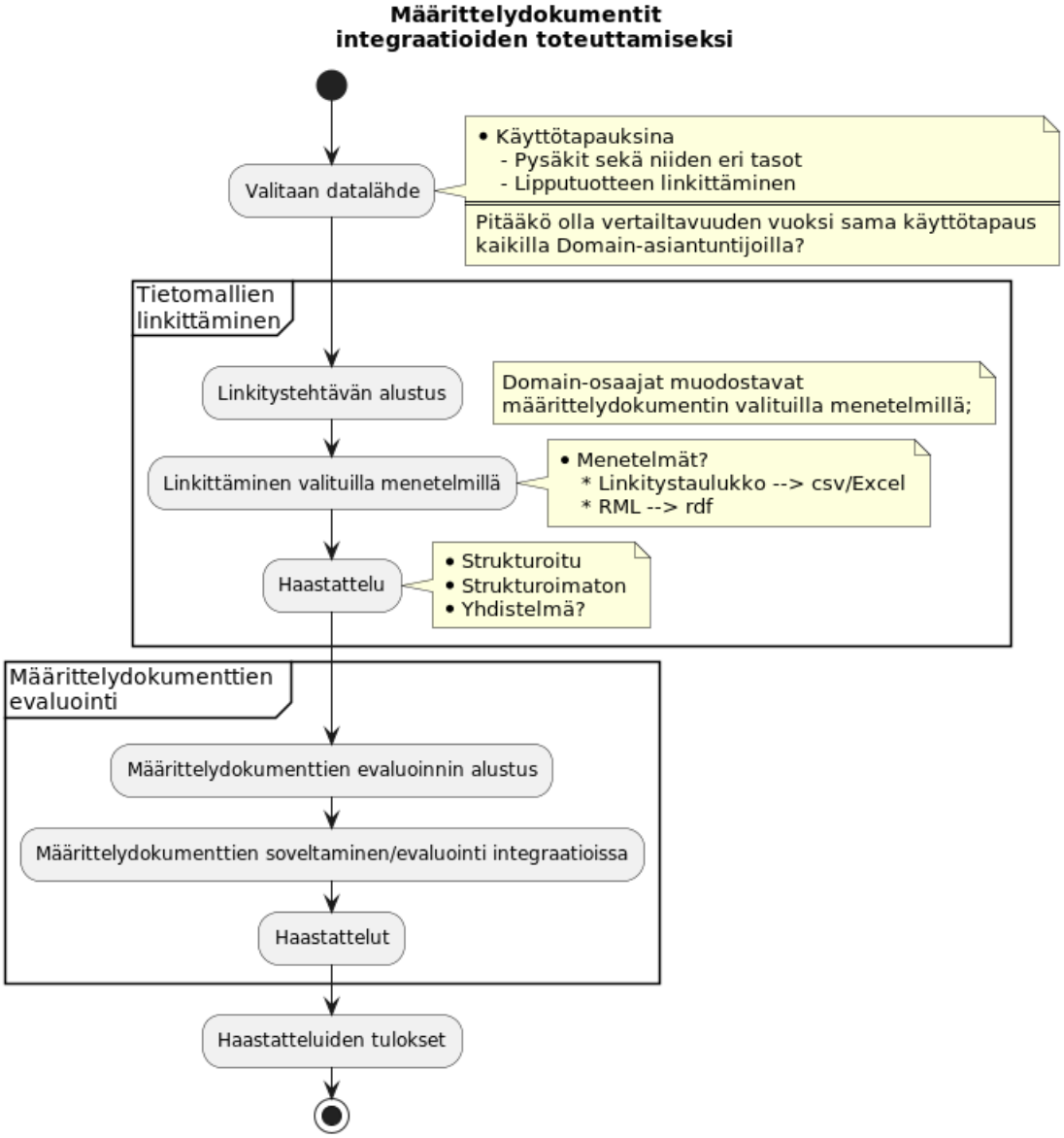
Aliluvussa 5.1 esiteltiin erilaisia havaintoja, jotka kohdentuvat domain-asiantuntijoiden mahdollisuuteen tuottaa laadukkaasti tietomallien yhdistämistä määrittelydokumenttien avulla sekä datan käsittelyvaiheessa integraattoreiden mahdollisuuteen uudelleenhyödyntää olemassaolevia määrittelydokumentteja. Havaintojen pohjalta voidaan tunnistaa kaksi sidosryhmää, jotka liittyvät olennaisesti yhdistetyn kohdemallin datan laatuun, domain-asiantuntijat sekä dataintegraattorit. Domain-asiantuntijat tuntevat substanssin sekä alan termistöt ja ovat kyvykkäitä tunnistamaan relaatiot eri tietomallien ontologioiden välillä. Domain-asiantuntijat eivät ole välttämättä tietotekniikan taidoiltaan kyvykkäimpiä, joten määrittelydokumenttien tuottamiseen vaadittavat tekniset taidot ovat yksi tarkasteltava näkökulma. Vastaavasti datan laadun kannalta on tarpeen tarkastella määrittelydokumentin vaikutusta loppudatan laatuun. Dataintegraattorit ovat merkittävässä roolissa transformoidun datan laadun kannalta, sillä he tekevät usein itse lähde- sekä kohdetyyppien välisiä linkityksiä, vaikka eivät välttämättä ole asiantuntijoita eri tietomallien ontologioiden osalta.

Haastattelututkimukseen valitut domain-osaajat valittiin sen perusteella, että heillä on lähde- ja kohdetietomallin asiantuntemusta ja he kykenevät osoittamaan eri tietotyyppien eroavaisuudet mallien välillä mahdollisimman tarkasti. Haastattelututkimuksen integraattorit valittiin siten, että he edustavat liikennealan keskeisiä toimijoita dataintegraatioissa ja kykenevät siten reflektoimaan eri linkitysmenetelmien vaikutuksia sekä potentiaalia tulevaisuuden toimintaprosessien sekä linkitysmenetelmien kannalta.

5.3.2 Tutkimuksen vaiheet

Haastattelututkimus jakaantuu kahteen eri vaiheeseen, joista toinen on kohdennettu domain-asiantuntijoille ja jälkimmäinen dataintegraattoreille. Haastattelututkimuksen ensimmäisessä vaiheessa on tarkoitus suorittaa domain-asiantuntijoilla lähdedatan tietotyyppien linkittämistä kohdemallin tietotyyppeihin valikoiduilla linkitysmenetelmillä. Tavoitteena on, että domain-asiantuntijat kykenevät tuottamaan menetelmien mukaiset määrittelydokumentit lopputuotoksena. Tämän jälkeen suoritetaan laadullinen semi-strukturoitu haastattelu, jonka avulla kartoitetaan henkilöiden kokemuksia menetelmistä domain-asiantuntijan näkökulmasta. Toisessa vaiheessa dataintegraattoreille annetaan ensimmäisessä vaiheessa tuotetut määrittely-

dokumentit arvioitavaksi sekä osoitetaan tarvittaessa ohjelmistot määrittelydokumenttien testaamiseksi. Arvioinnin ja testaamisen jälkeen suoritetaan semi-strukturoitu haastattelu dataintegraattoreiden näkemuksistä sekä havainnoista. Lopuksi haastatteluiden päähavainnot koostetaan tuloksiksi. Kuva 5.1 havainnollistaa työn eri vaiheita.



Kuva 5.1: Tutkimuksen eri vaiheet

Tutkimuksen toteuttamiseksi rajatuilla resursseilla sekä linkitysmenetelmien vertailtavuuden takia lähdedataksi valitaan pienikokoinen ja domain-asiantuntijoille

tuttu, jotta tutkimukseen osallistuvilta sidosryhmiltä vaadittava aika olisi mahdollisimman pieni.

5.3.3 Linkitysmenetelmien valintaperusteet

Tutkimusvaiheessa vertailtaviksi linkitysmenetelmiksi valittiin linkitystaulukko sekä RML. Menetelmiksi valittiin kaksi menetelmää, jotka ovat ominaisuuksiltaan erilaisia sekä domain-osaajien, että dataintegraattoreiden näkökulmasta. Määrältään kahden menetelmän vertailu on tutkimusvaiheen käytännön toteutuksen kannalta riittävä, sillä se vähentää tutkimukseen osallistuvien henkilöiden käytettävää aikaa ja on riittävä olennaisten havaintojen saavuttamiseksi ja tutkimuskysymyksiin vastaamiseksi. Linkitystaulukko on helposti ymmärrettävä toimintaperiaatteeltaan ja sen hyödynnettävyys vastaavasti vaatii dataintegraattorilta toimenpiteitä varsinaisen linkittämisen toteuttamiseksi. RML tuottaa dataintegraattorin näkökulmasta suoraan koneellisesti luettavan määrittelydokumentin ja se tukee useita eri lähdedatan serialisaatioita, mikä tekee siitä kiinnostavan menetelmän yleistettävyyden kannalta. Domain-asiantuntijoiden näkökulmasta on kiinnostavaa kartoittaa, miten RML vastaa tutkimuskysymyksienkin avulla nostettuihin haasteisiin.

5.3.4 Domain-osaajan haastattelurunko

Domain-osaajalle muodostettava haastattelurunko tulee muodostaa siten, että haastattelussa saadaan kartoitettua haastateltavan mahdolliset linkittämismenetelmän käyttöön vaikuttavat tekijät, kuten aikaisempi kokemus käytettävistä menetelmistä, tietotekninen taito, kokemus lähde-ja kohdetietomallista, motivaation taso linkittämiselle (tuntee käytännön ongelmia, jotka mahdollista ratkaista ko. menetelmällä). Toisessa haastattelun vaiheessa käydään läpi kokemuksia käytetyistä linkitysmenetelmistä, kartoitetaan linkitysmenetelmän käyttöön liittyviä haasteita sekä niiden kohdentumisesta johonkin tiettyyn linkitysmenetelmän ominaisuuteen. Kolmannessa vaiheessa kartoitetaan haastateltavan ajatuksia linkitysmenetelmien eroista domain-asiantuntijan näkökulmasta sekä ajatuksia siitä, näkeekö domain-asiantuntija linkitysmenetelmien mahdollisuuksia muodostua laajemman asiantuntijoiden menetelmäksi kansallisella tasolla. Viimeisessä vaiheessa haastateltava saa kertoa vapaasti omista kokemuksistaan sekä mahdollisista seikoista, joita haastattelussa ei muuten nostettu esille. Kysymysrunko:

- Taustatekijöiden ja kokemuksen kartoitus

  - Minkälainen kokemus sinulla lähdedatasta sekä sen tietomallista?
  - Minkälainen tietotekninen taito sinulla on?
  - Ovatko esitetyt linkitysmenetelmät tuttuja entuudestaan?
  - Voisiko käytetyillä menetelmillä ratkaista jotain tuntemaasi ongelmaa, joka liittyy tietomallien yhteensopivuuteen tai yhteiskäyttöön?
- Kokemukset suoritetuista linkityksistä
  - kuvaile linkitystaulukon käyttöä ja siihen liittyviä kokemuksia? Mistä pidit tai et pitänyt?
  - Oliko linkitystaulukon muodostamisessa ongelmia?
  - Oliko menetelmä soveltuva aineiston linkittämiseen?
  - kuvaile RML-menetelmän käyttöä ja siihen liittyviä kokemuksia?
  - Oliko RML:n avulla muodostuvat RDF-graafin muodostamisessa ongelmia?
  - Oliko menetelmä soveltuva aineiston linkittämiseen?
- Menetelmien laajempi käyttö sekä eroavaisuudet
  - Miten kuvailisit käyttämäsi menetelmien pääasiallisia eroja domain-asiantuntijan näkökulmasta?
  - Voisiko käyttämiä menetelmiä hyödyntää laajemman asiantuntijajoukon toimesta kansallisesti?
  - Voisiko tuotettuja määrittelydokumentteja jakaa ja ylläpitää yhdessä eri domain-asiantuntijoiden kesken?
- Muita esille nostettuja asioita haastattelusta?

5.3.5 Integraattorin haastattelurunko

Ensimmäisessä dataintegraattorin haastatteluvaiheessa kartoitetaan taustatiedot, jotka vaikuttavat määrittelydokumenttien arviointiin, kuten käytössä olevat menetelmät/ohjelmistot, joita integraattori hyödyntää. Lisäksi haastatellaan integraattorin kokemusta käsiteltävistä tietomalleista.

Haastattelun toinen vaihe pitää sisällään arvioinnin määrittelydokumenttien käytettävyydestä osana dataintegraatiota sekä tietomallien ja -tyyppien linkittämistä

(käytettävyys), määrittelydokumenttien sovellettavuus integraattorin käyttämissä menetelmissä sekä arviointia määrittelydokumenttien pohjalta muodostettujen dataobjektien oikeellisuudesta ja datan laadusta.

Kolmannessa vaiheessa tiedustellaan integraattorin ajatuksia määrittelydokumenttien keskinäisistä eroavaisuuksista sekä arviota siitä, missä tilanteissa ko. määrittelydokumenttia tulisi käyttää. Pyydetään myös haastateltavan ajatuksia määrittelydokumenttien iteratiivisestä rikastamisesta ja niiden soveltuvuudesta siihen. Viimeisessä vaiheessa haastateltava saa kertoa vapaasti omista kokemuksistaan sekä mahdollisista seikoista, joita haastattelussa ei muuten nostettu esille. Kysymysrunko:

- Integraattorin taustatiedot
  - Minkälainen kokemus teillä on käsiteltävistä tietomalleista?
  - Mitä teknologioita käytätte integraatioissa?
  - Minkälaisia kokemuksia teillä on erilaisista määrittelydokumenteista?
  - Onko teillä ennakko-odotuksia tai ajatuksia määrittelydokumenttien osalta?
- Määrittelydokumenttien laatu
  - Miten suoritatte määrittelydokumenttien arvioinnin?
  - Miten hyödynnettäviä määrittelydokumentit ovat mielestänne?
  - Miten pienellä vaivalla ne olisivat hyödyksi integraatioissa?
  - Miten alttiita määrittelydokumentit ovat domain-asiantuntijan virheille?
  - Ovatko linkitykset hyödynnettävissä mahdollisista virheistä huolimatta?
- Määrittelydokumenttien eroavaisuudet
  - Miten määrittelydokumentit eroavat toisistaan? Mistä näkökulmista katsoen?
  - Onko tunnistettavissa käyttökohteita, joihin pitäisi erityisesti käyttää tiettyä linkitysmenetelmää?
  - Miten hyvin määrittelydokumentit ovat rikastettavissa ja inkrementaalisesti kehitettävissä esimerkiksi kansallista käyttöä varten?

  - Olisiko laajemmalle yhteistyölle tarvetta jaettujen määrittelydokumenttien muodostamiseksi? Mitä hyötyä/haittaa siitä olisi?

- Muita esille nostettuja asioita haastattelussa?

5.4 Tutkimuksen kulku

Tässä aliluvussa kuvataan tutkimuksen eri vaiheiden kulkua ja kuvataan vapaammin havaintoja eri vaiheista, jotta lukijan on helpompi huomioida tutkimuksen eri vaiheissa esille nousseita asioita. Aliluvun kohdissa esitetään kronologisessa järjestyksessä etenevää tutkimusta ja sen vaiheiden toteutumista. Kohdassa 5.4.1 kuvataan tutkimuksen lähtötilannetta, jossa domain-asiantuntijat saavat linkitysharjoitukset tehtäväkseen. Kohdassa 5.4.2 kuvaa havaintoja tietomallien linkittämisestä valituilla menetelmillä. Kohdat 5.4.3 ja 5.4.4 kuvaavat havaintoja tuotetuista määrittelydokumenteista ja niiden arvioinnista. Viimeisessä kohdassa 5.4.5 pohditaan tutkimuksen validiteettia ja reliabiliteettia.

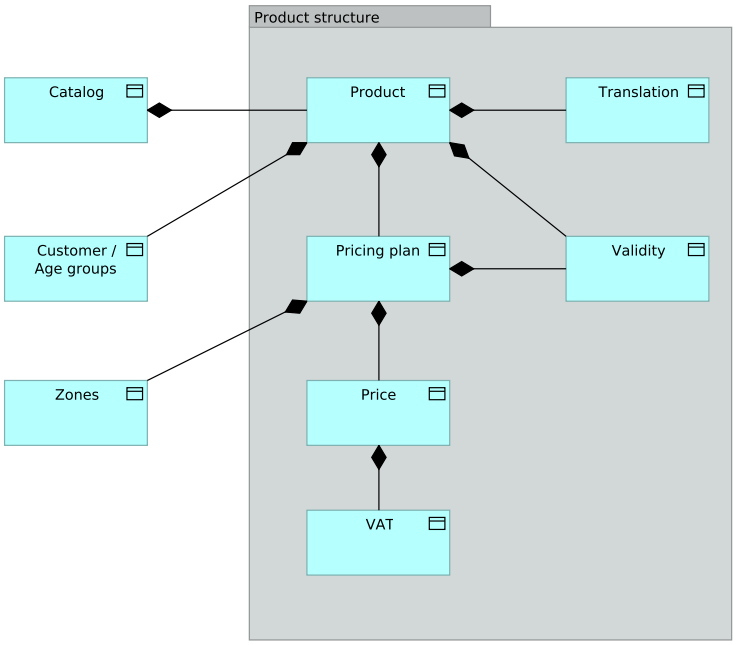
5.4.1 Lähtötilanne

Tutkimuksen lähtötilanteessa domain-asiantuntijoille on kerrottu tutkimuksen tarkoitus sekä esitelty kuvan 5.1 mukaiset vaiheet ja heidän työpanoksensa merkitys tutkimuskysymyksien ratkaisemisen kannalta. Domain-asiantuntijaa pyydetään suorittamaan lähdedatan pohjalta linkittäminen NeTEx-tietomallin mukaisiin objekteihin. Lähdedataksi valitaan domain-asiantuntijalle tuttu lähdedata.

Domain-asiantuntijan A:n datalähteeksi valikoitui lippu- ja maksujärjestelmän tuote- ja tariffipalvelun sisäinen tietomalli, joka mallintaa joukkoliikenteessä myytävien tuotteiden perustietoja, tariffeja sekä myyntiin ja käyttöön liittyviä sääntöjä. Tuote- ja tariffipalvelun vastuulla on myös tarjota MMTIS-asetuksen[1] mukaisesti tuote- ja hintatietoja NeTEx part 3 mukaisesti.

Kuvassa 5.2 on esitetty ylätasolla Tuote- ja tariffipalvelun sisäinen tietomalli, jossa tuote (Product) sisältää hinnoittelusuunnitelmia erilaisiin tarpeisiin. Hinnoittelusuunnitelma vastaavasti sisältää tarvittavat hintatiedot ja niiden ehdot, kuten voimassaolotiedot (Validity). Tuotteen käyttöä ohjaavat mm. asiakas- ja ikäryhmät

[1]Komission delegoitu asetus (EU) 2017/1926, annettu 31 päivänä toukokuuta 2017,Euroopan parlamentin ja neuvoston direktiivin 2010/40/EU täydentämisestä EU:n laajuisten multimodaalisten matkatietopalvelujen tarjoamisen osalta



Kuva 5.2: Tuote- ja tariffipalvelun sisäinen tietomalli.

(Customer / Age groups) sekä vyöhykkeet (Zones). Catalog on joukko tuotteita, jotka tarjotaan myyntikanavalle myynnin toteuttamiseksi.

Lähtötilanteen tehtävänannossa domain-asiantuntija A tunnisti tutkimusongelmaan liittyvät haasteet käytännön työssä sekä koki tutkimusvaiheessa toteutettavan linkittämisharjoitteen motivoivana sen tuottaman jatkohyödyn kannalta. Tehtävänannossa hänelle esitettiin sekä linkitystaulukko [3] että RML -menetelmä [9] ja siinä käytettävä RMLEditor -työkalu[2]. Lähtötilanteen tehtävänantokeskustelussa linkitystaulukko oli domain-asiantuntija A:n mielestä yksinkertaisempi ymmärtää, kun vastaavasti RML-menetelmän käyttäminen vaatii enemmän tukea ja selostusta alkuvaiheessa käyttöperiaatteen ymmärtämiseksi. RML-menetelmän tuottaman määrittelydokumentin mahdollisuudet koneelliseen tietomallien ja datan linkitykseen herättivät erityistä kiinnostusta menetelmää kohtaan, vaikka alussa sen käyttäminen vaatisi enemmän perehtyneisyyttä domain-asiantuntijalta.

Domain-asiantuntija B:n osalta datalähteenä käytettiin GTFS-datapaketin stops-tiedoston tietotyyppejä, joiden avulla voidaan muodostaa yksinkertainen pysäkkiverkoston hierarkia ja rakenne , kuten asema-, pysäkki- ja sijaintitiedot. Tämän lisäksi GTFS-pysäkkitietojen avulla mallinnetaan joukkoliikenteen palvelutietoja, ku-

[2]https://app.rml.io/rmleditor/

ten esteettömyys- ja vyöhyketietoja. Tällä hetkellä usean joukkoliikennetoimijan tietomalli perustuu GTFS-tietomalliin, jolloin MMTIS-asetuksen mukaisten minimivaatimuksien täyttämiseksi toteutetaan GTFS-tietomallin linkittäminen NeTEx-tietomallin tietotyyppeihin sekä tehdään tarvittava datan rikastaminen NeTEx-datan muodostamiseksi. Domain-asiantuntija B:n suhtautuminen tutkimusongelmaan käytännönläheistä, sillä erityisesti NeTEx-tietomalli herättää kiinnostusta ja miten sen avulla voidaan tuottaa parempia ja laadukkaampia joukkoliikennepalveluita asiakkaille.

Lähtötilanteen tehtävänannossa domain-asiantuntija B:lle oli nopeasti selvää, kuinka linkitystaulukkoa [3] käytetään. Vastaavasti RML-menetelmän [9] periaatteen ymmärtäminen on haastavampaa ja vaatii enemmän perehtymistä ja tukea. Domain-asiantuntijoille annettavan tehtävänannon osalta on tunnistettavissa, että tietomallien linkitysmenetelmien omaksumisessa on teoriaosaamisen kannalta eroavaisuuksia. Linkitystaulukko on helpommin lähestyttävä, sillä moni käyttää arkityössään taulukko-ohjelmia hyödykseen, jolloin varsinainen tietomallien linkittäminen on helpompaa. Vastaavasti RML-menetelmän käytön aloittaminen on helpompaa, mikäli henkilöllä on vahva tietomallien ja datan käsittelyn pohjaosaaminen. Mikäli ontologioiden ja graafien käsitteisiin joutuu käyttämään enemmän aikaa, menee varsinaisen tehokkaan työn aloittamiseen enemmän aikaa. Toisin sanoen domain-asiantuntija ei välttämättä ole tietomallien ja datan käsittelyn asiantuntija.

5.4.2 Tietomallin linkittäminen

Tietomallien linkittämistä varten domain-osaajille jaettiin linkki Excel-taulukkoon, johon muodostettiin Data4PT- työryhmän [3] esimerkin mukaisesti lähdedatan tietotyypit sekä attribuutit. Domain-osaajia autettiin linkitystaulukon täyttämisessä alussa, jonka jälkeen linkitystaulukon täyttäminen sujui omatoimisesti. Linkitystaulukon osalta oli tarkennettava ja sovittava mahdollisten luokkatietueiden sekä attribuuttien hierarkia, jonka mukaisesti taulukko täytetään. Joidenkin kohdemallin tietotyyppien sekä attribuuttien etsinässä autettiin linkitysmenetelmän käyttämiseksi.

RML-menetelmän käyttämiseksi vaadittiin enemmän perehtymistä kyseisen menetelmän sekä käytettävän RMLEditor-ohjelman [20] toimintaan. Vaikka ohjelma on yksinkertainen, sen käytettävyydessä on puutteita, jotka aiheuttivat hämmennystä domain-osaajissa. Esimerkiksi tallennustoiminnossa sekä virhelokit eivät toimineet toivotulla tavalla.

5.4.3 Linkittämisen lopputulos

Lopputuloksena linkitystaulukoista saatiin muutamia attribuutteja lukuunottamatta linkitykset lähde- sekä kohdemallien välillä. Näiden tuotoksien avulla on mahdollista toteuttaa ylätason konseptuaalinen linkittäminen, mutta tarkempien attribuuttien linkittäminen vaatisi iteraatiota tietomallien asiantuntijoiden kanssa. RML-menetelmällä ei kyetty muodostamaan ehyttä määrittelydokumenttia johtuen käytetyn RMLEditor-ohjelman [21] keskeneräisyydestä, kuten eri tallennustoiminnallisuuksien sekä virheenkäsittelyn puutteista.

Tietomallien linkittämisessä sekä linkitystaulukon että RML-menetelmän avulla oli omat haasteensa. Linkitystaulukon osalta kyettiin nopeasti muodostamaan linkitykset ylätason rakenteiden välillä, kuten domain-asiantuntija A:n kanssa suoritetussa tuoterakenteen linkityksessä, jossa myyntiehdot ja säännöt sisältävä tuote muodostuu NeTEx:n `SalesOfferPackage`-tietotyypiksi. Haastavinta on löytää jokaiselle attribuutille eksplisiittinen vastine ja vaatii lukuisia kommentteja sekä tarkentavia selitteitä linkityksien välille.

5.4.4 Määrittelydokumenttien arviointi

Johtuen tuotettujen määrittelydokumenttien puutteellisuudesta, dataintegraattoreille anntetaan arvioitavaksi yksi linkitystaulukolla tehty tietomallien linkitysdokumentti sekä esimerkkiaineisto, joka antaa parhaimman mahdollisuuden arvioida eri määrittelydokumentteja sekä niiden hyödynnettävyyttä dataintegraattorin näkökulmasta ja mahdollistaa tutkimuskysymyksiin vastaamisen. Lisäksi osoitetaan määrittelydokumentin arviointiin soveltuvat menetelmät tai vastaavasti pyydetään käyttämään integraattorin jo käytössä olevia integraatiotyökaluja. Dataintegraattorit käyvät läpi määrittelydokumentit sekä niihin liittyviä tietomalleja kontekstin ymmärtämiseksi.

5.4.5 Tutkimuksen validiteetti ja reliabiliteetti

Joseph Maxwell esittää artikkelissaan [14] laadulliseen tutkimukseen sovellettavia näkökulmia validiteetin arvioimiseksi:

- Kuvaileva validiteetti, jossa arvioidaan tutkimushavaintojen objektiivisuutta sekä kuvaamisen tarkkuutta

- Teoreettinen validiteetti kuvaa teorian soveltuvuutta tutkittavaan tutkimuskohteeseen
- Yleistettävyys kuvaa sitä, miten hyvin tutkimuksen tuloksia voidaan yleistää koskemaan tutkimuksen sisä- ja ulkopuolelle jääviä asioita. Esimerkiksi tuloksien hyödyntäminen toiseen kohderyhmään tai toisessa kontekstissa.
- Tulkinnallinen validiteetti, jolla arvioidaan tutkimuksessa tehtyjen tulkintojen täsmällisyyttä ihmisistä tai asioista.

Kuvailevan validiteetin näkökulmasta tapaustutkimuksessa on pyritty tarkastelemaan eri näkökulmista tietomallien linkitysmenetelmiä tarkastelevia lähteitä, jotta voidaan tunnistaa pääasialliset tietomallien linkitysmenetelmät ja niitä voidaan luokitella toisiinsa nähden tapaustutkimuksen rajaamiseksi. Tapaustutkimukseen valittujen linkitysmenetelmien osalta on suoritettu domain-asiantuntijoiden linkitykset sekä dataintegraattoreiden perehtyminen määrittelydokumentteihin, mikä auttoi lisäämään ymmärrystä käytetyistä menetelmistä ennen suoritettuja haastatteluita. On huomioitava, että validiteetin näkökulmasta riskinä on, että tutkijalla on rajalliset mahdollisuudet vaikuttaa haastateltavien henkilöiden kykyyn perehtyä tarvittaviin menetelmiin tai dokumentteihin. Tällä voi olla vaikutusta siihen, miten haastateltavat ovat vastanneet ja mistä näkökulmasta he vastaavat haastattelukysymyksiin. Esimerkiksi dataintegraattorien määrittelydokumenttien arvioinnin osalta olisi voitu tarkemmin kartoittaa, mitä menetelmiä dataintegraattorit käyttivät ja kuinka paljon aikaa käytettiin. Se olisi lisännyt läpinäkyvyyttä haastatteluissa esitetyille näkemyksille, kuinka paljon määrittelydokumentteihin on perehdytty. Tätä riskiä on pienennetty sillä, että haastateltavien kanssa on pidetty tiiviisti yhteyttä sekä linkitysharjoituksien yhteydessä varmistettu, että tehtävänanto on selkeä haastatteluun osallisuvilla asiantuntijoilla. Lisäksi haastatteluiden litteroinnissa on tehty luokittelua sekä tunnistettu yhtäläisiä tekijöitä useammasta haastattelusta, mikä vähentää yksittäisessä haastattelussa esitetyn asian painoarvoa kokonaisuudessa. Haastatteluiden litteroinnissa hyödynnettiin nauhoituksia haastatteluista, mikä auttoi tarvittaessa tarkistamaan haastattelutuloksien ja keskeisten asioiden koostamisessa. Haastatteluissa noudatettiin pääosin suunniteltua kysymysrunkoa, mutta ajoittain keskustelu erkaantui sivuun varsinaisesta kysymyksestä, mikä vaikeutti litteroinnin toteuttamista. Toisaalta haluttiin antaa haastateltavalle mahdollisuus kuvata näkemyksiään vapaammin, jos aiheeseen liittyen nousee esille uusia näkökulmia, joita tulisi huomioida ja pohtia analyysivaiheessa.

Teoreettisen validiteetin näkökulmasta työssä on hyödynnetty laadullista tapaustutkimusta, joka soveltuu tutkimuskysymyksien vastaamiseen sekä mahdollistaa tutkimuksen kohdentamisen tutkimusongelman ympärillä olevan ilmiön tarkastelemiseen. Työssä on hyödynnetty sekä linkitysmenetelmiin liittyvää sekä erityisesti kohdennettu joukkoliikenteen tietomallityöhön keskittyneitä lähteitä. Tämän avulla on ollut tarkoitus hyödyntää ja rajata tutkimuksessa käytettyjen menetelmien soveltumista erityisesti joukkoliikenne-kontekstissa hyödynnettyihin linkitysmenetelmiin. Työssä olisi voinut tarkastella laajemminkin linkitysmenetelmiä, erityisesti yleiskäyttöisempiä työkaluja, mutta niiden merkitystä ja suhdetta linkitysprosesseihin ja soveltamiseen olisi ollut haastavaa sitoa ilman, että tutkijana olisin tuottanut omaa mallia.

Yleistettävyyden näkökulmasta työssä hyödynnettiin erityisesti joukkoliikenteen tietomallien linkitysmenetelmiin liittyvää teoriaa sekä niihin sovellettavien menetelmien empiirisiä kokeiluja, jolloin havainnot, haastattelututkimuksen tulokset ja analyysi muodostavat kokonaisuuden, joka on muodostunut tämän tapaustutkimuksen reunaehtojen puitteissa. On todettava, että työssä esitetyt pääasialliset havainnot, jotka nousivat esille useammassa haastattelussa, voidaan yleistää tutkimuksen sisäpuolelle jäävissä asioissa.

Tulkinnallisen validiteetin osalta voidaan haastatteluiden osalta pohtia, minkälaisia vaikutuksia on ollut siinä, että tutkija sekä haastatteluihin osallistuneet ovat toisilleen tuttuja toimialalta. Onko haastateltavat mahdollisesti jättäneet sanomatta jotain sellaista, mikä olisi haastatteluiden litteraation kannalta oleellista, kun tiedostavat tutkijan olevan joukkoliikenteen tietojärjestelmien asiantuntija? Tätä riskiä voidaan pienentää sillä, että analyysissa reflektoidaan tutkimuksen tuloksien, teorialähteiden sekä oman pohdinnan avulla asioita. Vastaavasti haastateltavat sekä heidän syvä kontekstin ymmärrys sekä kokemus tutkimuksen aiheesta lisäävät tutkimuksen havaintojen validiteettia.

Laadullista tutkimusta ja sen reliabiliteettia voidaan arvioida kolmella tavalla [32]. Tutkimusmetodin reliabiliteetin arviointi, ajallinen reliabiliteetti sekä johdonmukaisuus tuloksissa. Laadullisen tapaustutkimuksen osalta reliabiliteetti on ongelmallinen arvioitava, sillä tapaustutkimus on tehty tietyistä lähtökohdista sekä tietyillä reuna-ehdoilla. Tutkimuksen tulokset riippuvat kaikista taustatekijöistä, jotka vaikuttavat kokonaisuuteen. Tutkimuksen haastattelut ovat tallenteina, jolloin toisen tutkijan on mahdollista arvioida kyseinen aineisto sekä muodostaa oma analyysi niiden perusteella. Ajallisesta näkökulmasta linkitysmenetelmät kehittyvät sekä

mahdollisesti tekoälypohjaiset ratkaisut muuttavat menetelmiä ja niihin sovitettavia prosesseja, jolloin ei voida taata, että samoista lähtökohdista toteutettava tutkimus tuottaisi vastaavia tuloksia. Johdonmukaisuuden näkökulmasta on todettava, että tutkimuksessa nousi esille fundamentaaleja lainalaisuuksia, jotka todennäköisesti johtopäätöksinä pitäisivät paikkansa, vaikka taustatekijöissä tapahtuisikin muutoksia.

6 Tutkimuksen tulokset

Tässä luvussa esitetään tutkimuksen tulokset. Aliluvuissa 6.1 ja 6.2 esitetään keskeisimmät havainnot domain-asiantuntijoiden ja dataintegraattorien haastatteluista. Näissä aliluvuissa on litteroituna haastatteluiden olennaiset sisällöt. Aliluvussa 6.3 on haastatteluiden tuloksien yhteenveto sekä niitä peilataan tutkimuskysymyksiin.

6.1 Haastattelut (domain-asiantuntijat)

Domain-asiantuntijoiden haastattelut suoritettiin aliluvussa 5.3.4 esitetyn haastattelurungon mukaisesti. Kohdassa 6.1.1 esitetään haastateltavien taustaa ja kartoitetaan heidän kokemusta, mikä havainnollistaa heidän lähtötasoaan. Kohdassa 6.1.2 käsitellään haastateltavien kokemuksia linkitysmenetelmistä, joita käytettiin linkitysharjoituksissa. Viimeisessä kohdassa 6.1.3 kartoitetaan haastateltavien mielipiteitä ja ajatuksia menetelmien laajemmista käyttömahdollisuuksista ja eroavaisuuksista.

6.1.1 Taustatekijöiden ja kokemuksen kartoitus

Domain-asiantuntija A:lla on erityisen vahva osaaminen joukkoliikenteen lippu- ja maksujärjestelmien toiminnasta sekä niiden tietorakenteista, minkä takia eri järjestelmissä käytetyt tietomallit ja niiden eroavaisuudet ja yhtäläisyydet ovat helposti hahmotettavissa. Hänellä on pitkä kokemus joukkoliikenteen lippu- ja maksujärjestelmistä sekä niihin liittyvistä tuoterakenteista aina 90-luvulta lähtien. Hänellä on kokemusta useista eri järjestelmäsukupolvista ja niiden keskinäisistä erovaisuuksista. Hän korostaa sitä, että perusasiat eri järjestelmien ja mallien välillä ovat samoja ja suurimmat muutokset liittyvät käteismyyntiin ja siihen liittyviin muutoksiin. Domain-asiantuntija A:n tietotekninen osaaminen on domain-asiantuntijaksi hyvällä tasolla. Osaaminen kattaa taulukko-ohjelmistojen sujuvan käytön, relaatiotietokantojen toiminnan sekä kyselykielien peruskäytön. Näiden lisäksi hänellä on kokemusta erilaisista ohjelmointirajapinnoista sekä niiden hyödyntämisestä palveluiden kehittämisessä ja testaamisessa. Linkitystaulukko tai sen erilaiset variaatiot ovat

tuttuja jo pitkältä ajalta, kun RML-menetelmä on vastaavasti täysin uusi menetelmä tietomallien linkityksessä. Domain-asiantuntija A:n mielestä käytettyjä linkitysmenetelmiä voidaan hyödyntää erityisesti tietomallien migraatioissa, kuten tuote- ja tariffitietoja kuvaavien mallien välillä. Se helpottaisi eri toimijoiden kykyä hyödyntää joustavasti eri tietomallien dataa esimerkiksi integraatioalustan kautta. Domain-asiantuntija A pohtii, että kertaluontoisessa linkityksessä tuskin tulisi käyttämään RML-menetelmää, mutta vastaavasti se voisi olla hyödyllinen käytettäväksi esimerkiksi integraatioalustalla tietomallien muunnoksissa.

Domain-asiantuntija B:llä on noin kymmenen vuoden kokemus käsiteltävästä pysäkkidatasta sekä niihin liittyvistä tietomalleista. Domain-asiantuntija B kokee olevansa keskivertoa parempi tietoteknisten taitojen osalta. Vaikka hänellä ei ole varsinaista tietoteknistä koulutusta, hän on erittäin kiinnostunut erilaisia järjestelmistä sekä niiden toiminnan perehtymiseen myös omalla ajallaan. Järjestelmien sekä palveluiden lisäksi myös erilaiset laitteet sekä niiden toiminta kiinnostaa. Domain-asiantuntija B:llä ei ole aikaisempaa kokemusta tutkimusvaiheessa käytetyistä linkitysmenetelmistä, mutta muita samankaltaisia menetelmiä on tullut aikaisemmin vastaan. Esimerkiksi hän koki RML-menetelmän olleen idealtaan vastaavanlainen kuin Power BI:ssa käytössä oleva editorinäkymä. Domain-asiantuntija B kokee, että molempia linkitysmenetelmiä voidaan hyödyntää tutkimuksessa tehtävään GTFS-tietomallin linkittämiseen NeTEx-tietomallin mukaiseksi. Aikaisemman kokemuksen pohjalta hän aloittaisi mieluummin RML-menetelmällä linkittämisen, sillä se muistuttaa PowerBI:n käyttöä.

6.1.2 Kokemukset suoritetuista linkityksistä

Domain-asiantuntija A:n kokemus linkitysmenetelmistä nosti esille, että kaksi jokseenkin samankaltaista ja samankaltaiseen tietomalliin tehtävää linkitystä on helpompaa ja suoraviivaisempaa hyödyntää linkitystaulukkoa kuin RML-menetelmää. Kuitenkin hän koki, että täysin uuden tietomallin (kohdemallin) muodostaminen tai monimutkaisia relaatioita sisältävien linkityksien tekeminen olisi helpompaa RML-menetelmällä. Linkitystaulukossa on haasteita määritellä eksplisiittisesti muut relaatiot kuin yhden suhde yhteen -relaatiot. RML-menetelmä auttaa havainnollistamaan paremmin monimutkaisemmat linkitykset eri tietomallien välillä. Domain-asiantuntija A:n mielestä haastavinta eri tietotyyppien ja käsitteiden linkittämisessä on monimutkaiset kokonaisuudet, jotka vaativat hyvää perehtymistä sekä lähde- että kohdemallin osalta. Oman linkitystehtävän (tuotetiedot) osalta ne olivat suhteel-

lisen yksinkertaisia ja linkitystaulukon käyttö tuntui luontevalta. Mielenkiintoisena näkökulmana Domain-asiantuntija A nosti esille, että miten eri tietomallien linkittäminen sekä mahdollisten datamuunnoksien vastuut jakautuvat eri toimijoiden kesken, mutta esitetyt havainnot koskettivat enimmäkseen implementaatiovaiheessa mahdollisesti esiintyviä väärinymmärryksiä, kuten tietotyypit, aikavyöhykkeet ja valuuttatietojen esittäminen. RML-menetelmän käyttöä häiritsi käytetyn RMLeditor -ohjelmaan liittyneet ongelmat, mikä vaikutti linkitysmenetelmän kokemukseen. Domain-asiantuntija A:n mukaan RML-menetelmässä on valtavasti potentiaalia pitkäjänteisemmän ja monimutkaisempien tietomallien linkittämiseen ja niiden ylläpitämiseen.Epäilyksiä kuitenkin herättää, onko menetelmä kuinka laajasti käytettyä ja onko tämä vain kapeamman käyttäjäkunnan käytössä? Toistaiseksi hän ei "hehkuttaisi" tämän menetelmän puolesta, sillä se vaatisi vielä enemmän käyttöä ja kokemusta, jotta voisi tehdä syvällisen arvion asiasta. Tämän takia kerran tehtäviin linkityksiin hän ei menetelmää käyttäisi, sillä enemmän menisi aikaa menetelmän ja työkalun opettelemiseen.

Domain-asiantuntija B:n mielestä tutkimuksessa tehtävä GTFS-aineiston linkittäminen NeTEx-tietomalliin oli haastavaa. NeTEx-tietomalli on paljon rikkaampi kuin GTFS, minkä takia linkitystaulukossa muodostuu kardinaalisuudeltaan "yhdestä moneen-relaatiota. Tämä aiheuttaa sen, että linkitystaulukkoon joutuu kirjaamaan epätäsmällisiä kommentteja. Domain-asiantuntija B:n mielestä herää epäilys mahdollisista väärinymmärryksistä linkitystaulukon eri käyttäjien välillä, kun tietotyyppien linkittäminen ei ole eksplisiittistä. Toinen esille nostettava haaste oli useita eri tasoja sisältävän NeTEx-mallin asemointi taulukkoon, jossa piti jakaa luokkaobjekti attribuuttien kanssa kahteen eri sarakkeeseen.

RML-menetelmä vaikuttaa domain-asiantuntija B:n mielestä potentiaaliselta sekä kehityskelpoiselta linkitysmenetelmältä, vaikka tutkimusvaiheessa käytetyssä RMLeditorissa oli ongelmia, mikä vaikutti linkittämisen toteuttamiseen. Erityisesti domain-asiantuntija B piti RML editorin visuaalisesta lähestymistavasta, mikä teki monimutkaisempien linkityksien tekemisestä selkeämpää. Lisäksi RML editor mahdollisti sääntöjen asettamisen, millä ehdoilla lähdemallin tietotyyppi linkitetään kohdemallin tietotyyppeihin. Domain-asiantuntija B näki RML-editorin muodostaman RDF-graafin olevan jatkossa helpommin hyödynnettävissä myös toisiin tietomallien linkityksiin ja siten mahdollistaa laajempien linkityksien toteuttamisen.

Suurimpina eroina linkitysmenetelmien välillä Domain-asiantuntija B piti RMLmenetelmässä käytettyä visuaalisempaa käyttöliittymää ja selkeyttä tietotyyppien

linkityksien hahmottamisessa verrattuna linkitystaulukkoon. Toisaalta linkitystaulukko on perinteisempi ja useamman käyttäjän omaksuttavissa käyttöön, sillä ei ole vaatimuksia opetella uuden ohjelmiston käyttöä. Domain-asiantuntijan mielestä perehtymällä RML-menetelmään sekä sen käyttöön alussa, saadaan pideämmällä aikavälillä hyötyjä monimutkaisempien linkityksien toteuttamisessa.

6.1.3 Menetelmien laajempi käyttö sekä eroavaisuudet

Domain-asiantuntija A:n mielestä pääasialliset erot eri linkitysmenetelmien välillä liittyvät RML-menetelmän parempaan soveltuvuuteen pitkäjänteisessä linkitystyössä sekä erilaisten rajapinta-integraatioden kehitystyössä, kun taas linkitystaulukko soveltuu paremmin suoraviivaisiin, yksinkertaisiin käyttötapauksiin. Erityisesti RML-menetelmän potentiaali yhdessä eri asiantuntijoiden kanssa tehtävään määrittelytyöhön herättää kiinnostusta ja vähentäisi päällekkäistä työtä eri asiantuntijoiden välillä. Linkitystaulukon ylläpitämisessä domain-asiantuntija A:n mielestä on tärkeää sopia eri toimijoiden välillä säännöt ja reunaehdot, miten eri tietomallien välisiä linkityksiä/määrittelydokumentteja ylläpidettäisiin. Domain-asiantuntija A nosti haastattelun lopussa pohdittavaksi ajatuksen, miten tekoälyä voitaisiin hyödyntää tietomallien linkityksissä ja eri linkitysmenetelmien käytössä?

Menetelmästä riippumatta linkitysmenetelmiä käyttävät henkilöt joutuvat tekemään saman ajatustyön ja molemmat menetelmät sopivat tietomallien linkittämiseen. Linkitystaulukon käytössä ja yhteistyön mahdollistamiseksi tulisi sopia yhtenäiset käytänteet linkitystaulukon täyttämiseksi, kuten millä kielellä kommentoidaan, miten kirjataan mahdolliset huomiot. Linkitystaulukko antaa toisaalta mahdollisuuden vapaammin täyttää tietoja, mutta riskinä on mahdolliset implisiittiset kirjaukset sekä tulkinnanvaraisuus eri asiantuntijoiden välillä. RML-menetelmä soveltuu muodostuvien RDF-graafien takia hyviä yhteistyöhön eri asiantuntijoiden välillä, sillä jo tehtyjä linkityksiä tietotyyppien välillä on muiden asiantuntijoiden mahdollista hyödyntää, mikä auttaa kehittämään tarkempia linkityksiä eri tietomallien välillä. RML-menetelmä ohjaa myös yhtenäisempään linkitystapaan, sillä se vaatii määrittämään linkitykset tietyllä tavalla. Myös ”yhdestä moneen” -kardinaalisuuden sisältävät linkitykset ovat helpommin tulkittavissa eri asiantuntijoiden välillä, kun työkalu mahdollistaa visuaalisen näkymän tarkastella tietoja. Linkitystaulukkoa voi joutua lukemaan läpi enemmän ennen kuin pystyy hahmottamaan linkitykset eri tietotyyppien välillä. Domain-asiantuntija B nosti esille ajatuksen, että RMLEditor-työkalussa esiintyneet virheet saattavat aiheuttaa sen, että

koko linkitysmenetelmää ei haluta käyttää, vaikka pohjalla oleva tietomallien linkitykset määrittävä RDF-dokumentti täyttäisi vaatimukset yhteistyössä ja jatkuvasti tarkentuvien tietomallien linkityksien tekemiseksi.

6.2 Haastattelut (dataintegraattorit)

Dataintegraattorien haastattelut suoritettiin aliluvussa 5.3.5 esitetyn haastattelurungon mukaisesti. Kohdassa 6.2.1 kartoitetaan dataintegraattorien taustatietoja ja kokemuksia joukkoliikenteen tietomalleista ja linkitysmenetelmistä. Kohdissa 6.2.2 ja 6.2.3 on esitetty dataintegraattorien havaintoja määrittelydokumenttien piirteistä ja laatuun vaikuttavista tekijöistä.

6.2.1 Integraattorin taustatiedot

Dataintegraattori A:lla on syvällinen kokemus käsiteltävistä tietomalleista, sillä hän on sekä hyödyntänyt tietomallien mukaista dataa eri projekteissa sekä tuottanut dataa eri ratkaisuissa. Tietomallien välisissä dataintegraatioissa hän on usein hyödyntänyt valmiita ohjelmistokomponentteja, joita on ollut käsiteltävien tietomallien osalta saatavilla, mutta usein tapauskohtaisesti on tarvittu itse toteutettuja ratkaisuja. Käytettäviä teknologioita erityisesti tietokantojen osalta määrittävät olemassa olevat sovelluskirjastot ja tietomallin rakenne. Esimerkiksi GTFS-tietomallin mukainen data on helpommin relaatiotietokannassa ylläpidettävää strukturoitua dataa, kun vastaavasti NeTEx-tietomallin mukainen data on dataintegraattori A:n mukaan helpommin NoSQL-tietokannassa ylläpidettävää dataa. Eri integraatioiden osalta kuitenkin tapauskohtaiset vaatimukset määrittävät käytettävät teknologiat. Aikaisemmissa tapauksissa dataintegraattori A on hyödyntänyt sekä kohdannut linkitystaulukoita sekä UML-kaavioita tietomallien linkityksien määrittelemiseksi.

Dataintegraattori B:llä on noin kuuden vuoden ajalta kokemusta joukkoliikenteen eri tietomalleista, erityisesti tuote- ja tariffihallinnan sekä tuotteiden myyntiin liittyvien osa-alueiden tietomallit ovat tuttuja. Hänellä on pitkä kokemus ja ammattitaito erilaisista tietomalleista, jolloin ajan kanssa osaaminen tapauskohtaisesti käsiteltävistä tietomalleista syventyy, kun niitä työstetään syvällä tasolla. Hän hyödyntää erilaisia standardidokumentteja datalähteenä, mikä ei kuitenkaan korvaa ajan saatossa muodostettua ammattitaitoa soveltaa standardeja käytännön ratkaisuissa. Dataintegraatioissa he hyödyntävät mm. Quarkus, Java, Camel, Kafka, Minio tek-

nologioina sekä työkaluina.

Määrittelydokumentteina käytetään niitä dokumentteja, joita saadaan domain-asiantuntijoilta, ei ole selkeästi yhtä ainoaa tapaa tuottaa määrityksiä. Kokemus on osoittanut, että kaikista tärkein on iteraatio eri sidosryhmien kesken ja domain-asiantuntijoiden kanssa yhteistyössä muodostetaan eri tietotyyppien väliset mappaykset tai muodostetaan kokonaan uusi tietomallin skeema. Dataintegraattori B:n mielestä linkitys tai määrittelydokumenttien menetelmää on hankala sovittaa kaikkien sidosryhmien tarpeisiin, minkä takia on parempi panostaa iteratiiviseen määrittelytyöhön, jossa sekä domain-asiantuntijat että dataintegraattori käyttävät heille parhaiten soveltuvia ja tutuimpia menetelmiä. Dataintegraattori B toteaa:

> "Määrittelyissä mukana olevat sidosryhmät voidaan jakaa Excel-kerrokseen sekä koodareiden tapauskohtaisesti sovellettaviin määrittelymenetelmiin. Aivan ylin johto kommunikoi pääsääntöisesti jopa Powerpointia käyttäen."

Lisäksi hän painottaa, että edellä mainittujen ryhmien välillä on oltava yhdistävä taho ja kaikista arvokkaimmat ja kriittisimmät henkilöt määrittelyiden lopputuloksien kannalta ovat ne henkilöt, jotka kykenevät liikkumaan kaikilla tasoilla ja kommunikoimaan sidosryhmän ymmärtämällä tavalla. Näitä on usein ainoastaan yhden käden sormilla laskettavissa domain-kohtaisesti.

6.2.2 Määrittelydokumenttien laatu

Molemmat dataintegraattorit vertailivat linkitysmenetelmillä luotuja määrittelydokumentteja siitä kulmasta, miten hyödynnettäviä ne ovat kontekstin ymmärtämisen sekä varsinaisen toteutuksen näkökulmasta. Dataintegraattori A:n näkemyksen mukaan RDF-dokumentti on ihmissilmälle vaikeampi määrittelydokumenttina, mutta vastaavasti sille löytyy valmiita parsereita, jolloin se on syntaktiset virheet ovat vältettävissä dataintegraatioissa. Lisäksi RDF-määrittelydokumentti on hyödynnettävissä joissakin integraatioviitekehyksissä. Hän korostaa sitä, että jokaisen käyttötapauksen osalta joutuu perehtymään kontekstiin sekä integraatiossa käsiteltävään dataan, jolloin linkitystaulukko on helpompi asiaan perehtymisen kannalta, sillä siinä on mahdollista lisätä myös sanallisia tarkennuksia ja huomioita. Toisaalta linkitystaulukko ei ole täsmällisin erityisen monimutkaisissa linkityksissä. Dataintegraattori A nostaa esille, että linkitystaulukolla muodostettu määrittely voi olla hieman puutteellinen ja silti hyödyllinen käytettäväksi integraation toteuttamisessa.

Dataintegraattori B:n mukaan määrittelydokumenttien laatuun vaikuttaa keskeisesti iteraatio domain-asiantuntijan sekä dataintegraattorin välillä, mikä tuottaa varsinaisen teknisen määrittelydokumentin, mutta varmistaa myös lopullisen määrittelydokumentin laadun. Tämä mahdollistaa sen, että dataintegraattori kykenee varmistamaan domain-asiantuntijan määrittämät tietotyypit ja saa tarvittavat tiedot muodostaessaan laadukkaan lopputuotoksen. Linkitystaulukossa riski väärinymmärryksille kasvaa tietotyyppien linkityksissä, joissa kardinaalisuus on jotain muuta kuin yhden suhde yhteen. Näissä tilanteissa linkitystaulukon pohjalta muodostettavan linkityksen laatu voidaan varmistaa toimivalla iteraatiolla eri sidosryhmien välillä.

6.2.3 Määrittelydokumenttien eroavaisuudet

RML-menetelmällä tuotetut RDF-määrittelydokumentit voivat olla asiaan perehtymättömän integraattorin osalta hankalasti tulkittava, joten dataintegraattori A pääsisi helpommin liikkeelle linkitystaulukon avulla, jossa on vapaammin selitteitä linkitettävien tietueiden välillä. Vaikka RDF-dokumenttia voisi koneellisesti lukea parserin avulla, hän kaipaa mahdollisuutta käydä aineistoa ihmisen luettavassa muodossa läpi. Dataintegraattori A hyödyntäisi yksinkertaisemmissa integraatioissa linkitystaulukkoa sekä muita tukevaa dokumentaatiota, kuten arkkitehtuurikuvauksien tai esimerkkitapauksen avulla. Erityisesti monimutkaisemmat relaatiot eri tietueiden välillä puoltaisivat tukimateriaalin käyttämistä. RML-menetelmä voisi soveltua laajoihin, erittäin monimutkaisiin integraatioihin, mutta hänen mielestään tiukassa, aikarajatussa projektissa ei ole välttämättä aikaa uuden toteutusmallin käyttöönotolle. Dataintegraattori A pyrkii itse toteuttamaan integraatioissa käytettävän koodin itse, sillä luottaa omaan kykyyn ja oman koodin tuomaan joustavuuteen erilaisissa tapauksissa. Hänen mielestään RML-menetelmä soveltuisi paremmin laajempaan tietomallin yhteiskehittämiseen, sillä se olisi helpommin versioitavissa sekä mahdollistaisi dokumentaation automaattisen generoinnin luontevammin.

Dataintegraattori B painottaa, että linkitysmenetelmästä huolimatta tarvitaan iteraatio eri osapuolien välillä, joten varsinainen linkitysmenetelmä on ainoastaan tapauskohtaisesti valittu työkalu. Hän painottaa, että joka tapauksessa domain-asiantuntija kuvaa asiat tietyllä abstraktiotasolla, josta puuttuu mahdollisesti tietyt dataintegraattoria kiinnostavat asiat. Viime kädesssä työkalu ei ole tärkeää, vaan saumaton yhteistyö domain-asiantuntijan sekä dataintegraattorin kanssa ratkaisevat lopputuloksen laadun. Hänen mielestään ruutupaperille tehty määrittely voi olla

erinomainen, jos siihen on kuvattu selkeästi tietotyyppien linkitykset. Kun vertaillaan RML-menetelmää sekä linkitystaulukkoa toisiinsa, on tärkeässä roolissa linkitysmenetelmän elinkaari ja yleisyys, jotta jatkuvuus linkityksien toteuttamisessa pysyy eheänä. Dataintegraattori B toteaa valitun linkitysmenetelmän vaikuttavan koodaamisen työmäärään, joten RML-menetelmällä voidaan tapauskohtaisesti vähentää tarvittavan ohjelmoinnin määrää, kun linkitykset muodostuvat suoraan koneluettaviksi.

Dataintegraattori B näkee eri linkitysmenetelmien välillä eroavaisuuden tietomallien yhteiskehittämisen näkökulmasta. Hänen mielestään laajassa tietomallikehittämisessä korostuu formaalien työkalujen käyttäminen ja yksityiskohtaisesti kuvattavien tietotyyppien tarkkuus, jolloin linkitystaulukko ei ole hänen mielestään riittävä. Tietomallien yhteiskehittämisessä kansallisella tai laajemmalla tasolla on todennäköisesti kyse standardin määrittelytyöstä, jolloin käytettävän menetelmän tulee soveltua laajojen tietomallilinkityksien muodostamiseen. Hän myös korostaa, että laajalle yhteistyölle tietomallien kehittämisessä on perusteita, sillä se mahdollistaa määrittelyn tason nostamisen ja todennäköisyys sen yleiselle hyväksynnälle kasvaa.

6.3 Yhteenvetoa tutkimuksen tuloksista

Tapaustutkimuksessa valittiin teoriaosuudessa käsitellyistä linkitysmenetelmistä kaksi vertailtavaa menetelmää, jotka poikkeavat toisistaan eri tekijöiden pohjalta. Valituissa menetelmissä päädyin Linkitystaulukkoon sekä RML-menetelmään, sillä ne poikkesivat toisistaan merkittävästi. Linkitystaulukko on teknisesti varsin yksinkertainen ja helposti omaksuttavissa, eikä siihen tarvitse taulukkotyökalujen yleisyyden vuoksi merkittävää perehtymistä domain-asiantuntijoilta. Vastaavasti RML-menetelmässä on useita etuja, kuten tietomallien ja datan validointimahdollisuudet sekä iteratiivisesti kasvavan graafitietokannan muodostuminen RDF-dokumenttien muodossa. Tavoitteenani oli selvittää, onko löydettävissä käsiteltävän tapauksen osalta linkitysmenetelmää, joka tukee sekä domain-asiantuntijoiden, että dataintegraattorien tarpeita. Lisäksi halusin selvittää, mitkä tekijät muodostavat linkitysprosessissa haasteita ja onko löydettävissä asioita, jotka voivat parantaa tietomallien linkityksien laatua samalla säilyttäen linkitysmenetelmän eri sidosryhmien tarvitseman tuen linkitysprosessissa. Tutkimuskysymykset, jotka esitettiin työn aikaisemmassa vaiheessa:

- Miten pienen perehtymisen jälkeen domain-osaajien on mahdollista tuottaa linkittämiseen vaadittava määrittelydokumentti?
- Miten paljon teknistä osaamista määrittelydokumentin muodostaminen vaatii domain-osaajalta?
- Mikä on linkityksessä tuotetun lopputuloksen (määrittelydokumentti) jatkohyödynnettävyys toisissa linkityksissä?
- Mitkä tekijät määrittelydokumenteissa vaikuttavat yhdistetyn datan laatuun?

Mielestäni ensimmäiseen tutkimuskysymykseen saimme vastauksia ja vaikuttavia tekijöitä aliluvussa 5.4, jossa pääsin tarkastelemaan domain-osaajien perehtymistä valittuihin linkitysmenetelmiin. Lisäksi haastatteluvaiheessa molemmat sidosryhmät nostivat esille havaintoja ja kokemuksiaa vertailtavista linkitysmenetelmistä. Tutkimuskysymykseen voidaan vastata, että erityisesti domain-osaajan aikaisempi kokemus sekä tietotekninen taitotaso vaikuttavat siihen, miten helposti he kykenevät tuottamaan valitulla menetelmällä määrittelydokumentin.

Vastaus tutkimuskysymykseen liittyen domain-osaajalta vaadittavaan tekniseen osaamiseen liittyen määrittelydokumentin muodostamiseksi ei ole suoraviivainen. Mikäli tavoitellaan ideaalia tilannetta, jossa domain-asiantuntija tuottaisivat pitkälle valmiin, eheän määrittelydokumentin dataintegraattoreita varten, linkitystaulukko on yleisesti helpommin lähestyttävä ei-teknisen henkilön toimesta. Taulukkoohjelmat ovat yleisesti käytettyjä ja sen takia linkitystaulukon käyttäminen sisältää vähemmän uusia opeteltavia asioita suhteessa vertailtavaan RML-menetelmään. RML-menetelmän osalta nousee erilaiset mahdollisuudet hyödyntää visuaalisia käyttöliittymäohjelmistoja, kuten tässä työssä käytetty RMLeditor-ohjelma, joka mahdollistaa RML-menetelmän hyödyntämisen ei-teknisemmän henkilön toimesta. Kuitenkin nousi esille haastatteluissa, että käytetyllä ohjelmistolla sekä sen toimintavarmuudella on suuri merkitys käyttäjän käyttökokemuksen kannalta sekä mahdollisuuksiin tuottaa valmis määrittelydokumentti. Domain-asiantuntijoiden haastatteluiden pohjalta nousi esille, että domain-asiantuntija A koki RML-menetelmän käytön vieraaksi ja nosti esille tarpeen paremmalle perehtymiselle käytetyn menetelmän osalta. Vastaavasti domain-asiantuntija B, jolla oli aikaisempaa kokemusta mm. PowerBi-ohjelmistossa tehtävistä tietomallien linkityksistä, piti RMLeditor-työkalulla aloittamista helpompana lähestymisenä tehtäviin tietotyyppien linkityksiin. Käytännössä käyttökokemus käytettävästä linkitysmenetelmästä on ratkaisevassa roolissa. On huomioitava, mikäli RML-menetelmän kanssa olisi hyödynnetty

toista työkalua kuin RMLeditor, sillä olisi ollut vaikutusta linkitysmenetelmän käyttökokemukseen domain-asiantuntijan näkökulmasta. Tähän tutkimukseen valitsin RMLeditor-työkalun käytettäväksi, sillä se oli ennakko-odotuksien pohjalta yksinkertaisimman oloinen editor-ohjelma ja muut vastaavat ohjelmistot olisivat vaatineet enemmän varsinaisen RDF-dokumentin syntaksin ymmärtämistä. Dataintegraattorien haastatteluiden pohjalta voidaan todeta, että ainakaan paikallisen tason tietomallien linkityksissä ei heidän mukaan ole merkitystä, miten domain-asiantuntija tuottaa määrittelydokumentin ja kyseisen dokumentin laadun edelle voidaan nostaa tiiviimpi iteraatio eri osapuolien välillä, mikä vähentää riskejä mahdollisesti puutteellisen määrittelydokumentin osalta. On kuitenkin huomioitava, että kansainvälisen tason tietomallien linkittäminen on todennäköisesti enemmän domain-osaajien keskenään tekemää jatkuvaa työtä, mutta he tekevät määrittelytyötä aktiivisesti ja omaavat kokemusta myös käytettävistä linkitysmenetelmistä.

Kolmantena tutkimuskysymyksenä tarkasteltiin, mikä on tuotettujen määrittelydokumenttien jatkohyödynnettävyys. Jatkohyödynnettävyyttä voidaan tarkastella eri näkökulmista, joita ovat dokumentin syntaktinen laatu, määrittelydokumentin hyödynnettävyys iteraatiossa sekä hyödynnettävyys jatkuvassa iteratiivisessa määrittelyssä. Syntaktisen laadun osalta RML-menetelmän tuottama RDF-dokumentti on validoitavissa mahdollisten virheiden osalta, mikä antaa paremmat edellytykset määrittelydokumentin hyödyntämiseksi dataintegraatioissa sellaisenaan. Tämän asian myös dataintergraattorit toivat esille haastatteluissaan. RML-menetelmää voidaan hyödyntää erilaisilla työkaluilla ja esimerkiksi käytetty RMLeditor-työkalu mahdollisti lähde- ja kohdetietomallin mukaisen datan lataamisen suoraan datana, jolloin käyttäjälähtöisten virheiden mahdollisuus on huomattavasti pienempi. Linkitystaulukko täytetään käyttäjän toimesta kokonaan, mikä mahdollistaa käyttäjän tekemät syntaktiset virheet taulukossa. Mikäli linkitystaulukkoa haluaisi koneellisesti lukea, se vaatisi erikseen tehtävän validointimenetelmän sekä tapauskohtaisesti sovittavan tavan täyttää taulukkoa. Toisaalta myös Beurèe ja Knowles esittävät[3], että linkitystaulukko soveltuu pohjaksi dataintegraatioita varten, mutta eivät esitä sen suoraan sellaisenaan sovellettavaksi dataintegraatioissa. Linkitystaulukko mahdollistaa haastatteluiden pohjalta helpommin lähestyttävän tavan tehdä tietueiden linkityksiä toisiinsa ja mahdollistaen myös implisiittisemmän kommentoinnin ja selitteiden kirjaamisen, mikä toisaalta lisää iteraation tarvetta, mutta auttaa myös ymmärryksen lisäämisessä erilaisissa rajatapauksissa eri sidosryhmien välillä. Haastatteluiden pohjalta nousi esille, että iteraation merkitys on suuri erityisesti datainte-

graattoreiden näkökulmasta. He nostivat esille tarpeen perehtyä kontekstiin ja linkitystaulukko on luontevampi ja helpommin lähestyttävä määrittelydokumentti johtuen mahdollisuudesta lisätä kommentteja sekä selitteitä linkityksien osalta. RML-menetelmän muodostama RDF-dokumentti ei ole haastatteluiden perusteella mielekästä ihmissilmin luettavaksi ja sen pohjalta haastateltavat kaipaavat rinnalle jotain muuta tukimateriaalia. On huomioitava, että RML -menetelmälle on ja on toteutettavissa eri tarpeisiin sovellettavia ratkaisuja visuaaliseen esittämiseen. Jatkohyödynnettävyyden näkökulmasta ja pitkäjänteiseen tietomallien määrittely ja linkitystyöhön haastateltavat totesivat, että RML-menetelmässä on selvästi enemmän potentiaalia verrattuna linkitystaulukkoon. Se mahdollistaa useiden eri RDF-graafien pohjalta tehtävät linkitykset ja muiden yhteisöjen tuottamien tietomallilinkityksien hyödyntämisen tai rikastamisen.

Viimeisenä tutkimuskysymyksenä tarkasteltiin, mitkä tekijät vaikuttavat yhdistetyn datan laatuun. Yhdistetyn datan laatu muodostuu usean eri tekijän, joita tässä työssä on aikaisemmin nostettu esille, muodostamasta kokonaisuudesta. Tämän takia on vaikea määrittää tarkasti, mitkä tekijät juuri määrittelydokumentissa vaikuttavat yhdistetyn datan laatuun. Tutkimuskysymykseen voidaan haastatteluiden sekä havaintojen pohjalta nostaa kaksi pääasiallista tekijää riippuen määrittelydokumentin roolista osana kokonaisprosessia. Ensimmäinen tekijä on määrittelydokumentin syntaktinen laatu, jonka painoarvo on suurempi linkitysprosesseissa, joissa määrittelydokumenttia käytetään suoraan teknisessä dataintegraatiossa. Tällöin erityisesti RML-menetelmän kaltainen validoitava dokumentti on kattavasti toteutettuna paremmin soveltuva. Toinen pääasiallinen tekijä on määrittelydokumentin kyky tukea eri sidosryhmien iteraatiota. Käytännössä tämä tarkoittaa sitä, miten dokumentin avulla domain-asiantuntijat voivat dokumentoida tietotyyppeihin liittyviä yksityiskohtia mahdollisimman tarkasti toisille domain-asiantuntijoille tai tukea dataintegraattoria muodostamaan täsmällinen kohdemallin mukainen data.

Tutkimuskysymyksien sekä työssä saatujen havaintojen sekä haastatteluiden perusteella määrittelydokumentin paremmuutta on mahdotonta sanoa yksiselitteisesti. Kyseessä on luvuissa 6 sekä 7 nostettujen tekijöiden, valitun kokonaisprosessin, linkitysmenetelmän sekä määrittelydokumentin muodostama kokonaisuus, jotka vaikuttavat kokonaisuuden lopputulokseen.

7 Analyysia ja reflektointia

Tässä luvussa analysoidaan teorian sekä empiirisessä tutkimusvaiheessa saatujen tuloksien pohjalta tietomallien linkitykseen liittyviä tekijöitä. Aliluvussa 7.1 analysoidaan eri linkitysmenetelmien eroavaisuuksia ja niiden ominaisuuksia toisiinsa. Aliluku 7.2 analysoi määrittelyprosessien suhdetta linkitysmenetelmiin. Erityisesti aliluvussa pyritään tunnistamaan lopputuloksen kannalta olennainen riippuvuus linkitysmenetelmän ja toiminta- sekä applikaatioprosessien välillä. Aliluvussa 7.3 analysoidaan koko toimintakentän riippuvuuksia toisiinsa sekä isojen osakokonaisuuksien vaikutusta tutkimuksen havaintoihin sekä linkitysmenetelmiin.

7.1 Linkitysmenetelmien eroavaisuudet haastatteluiden pohjalta

Tässä aliluvussa käsitellään linkitysmenetelmien eroavaisuuksia toisiinsa erilaisista näkökulmista tarkastellen. Linkitysmenetelmiä on tarkasteltu kohderyhmien ja hyödynnettävyyden näkökulmista. Kohdassa 7.1.1 pohditaan linkitysmenetelmien kohderyhmiä, joiden näkökulmasta ne ovat kehitettyjä. Vastaavasti kohdassa 7.1.2 analysoidaan, miten eri linkitysmenetelmät tuottavat lisäarvoa ja minkälaisissa tapauksissa niitä kannattaa hyödyntää.

7.1.1 Linkitysmenetelmien kohderyhmät

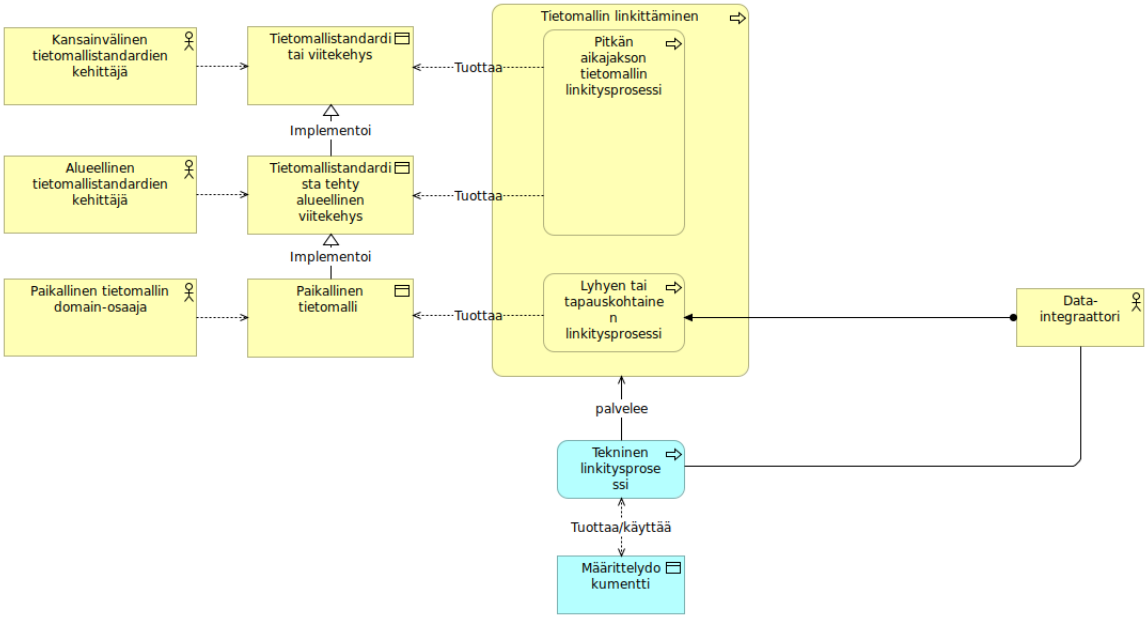
Linkitystaulukko on toteutettu MMTIS-asetuksen asettamien tiedon jakamiseen liittyvien vaatimuksien aiheuttamien laajojen tietoryhmien linkittämiseen toisiinsa ja tietomallien kehittämisessä tehtävän yhteistyön apuna eri domain-asiantuntijoiden välillä [3]. Beurèe ja Knowles esittävät menetelmän hyödyntämisen eri määrittelyprosessien osana ja kohderyhmänä ovat domain-asiantuntijat, jotka tuntevat tietomallien yksityiskohdat ja kontekstin tarkasti, mutta eivät todennäköisesti toteuta teknisiä integraatioita. Esimerkiksi datan konversiotyökalujen osalta he nostavat esille tekijöitä, joita domain-asiantuntijoiden tulisi huomioida tietomallien linkityksien laatuun sekä tekniseen toteutukseen vaikuttavina tekijöinä [3, s.26]. Haastatteluiden perusteella Linkitystaulukko oli kaikkien haastateltavien mielestä helposti ja nopeasti lähestyttävä.

Dimou et al. [9] esittävät, että RML-menetelmä on tarkoitettu heterogeenisen datan ja progressiivisesti kehittyvien eri tietomallien yhteensopivuuden varmistamiseen myös implementaatiotasolla, jossa helposti muodostuu päällekkäisiä konsepteja sekä konflikteja eri datojen tunnisteiden välillä. Menetelmä on tarkoitettu sekä domain-asiantuntijoille sekä dataintegraattoreille ja sen kanssa hyödynnettävät ohjelmistot, kuten tässä työssä hyödynnetty RMLEditor [21] täydentävät eri osapuolien tarpeita tietomallilinkityksien tekemiseksi. Domain-asiantuntijoille tarjotaan visuaalisia sekä vaihtoehtoisia työkaluja tuottaa RDF-dataa, kun vastaavasti RML-menetelmä pyrkii mahdollistamaan dataintegraattoreille syntaktisesti laadukasta dataa. Vaikka lähdemateriaaleissa Linkitystaulukko oli vahvasti sidottu joukkoliikenteen kontekstiin, molemmat menetelmät soveltuvat hyvin geneerisinä tapoina sovellettavaksi.

Domain-asiantuntijat aloittaisivat tietomallien linkityksien tekemisen menetelmällä, jonka kaltaisesta lähestymistavasta heillä on aikaisempaa kokemusta. Esimerkiksi domain-asiantuntija A:lle Linkitystaulukon sekä sen variaatioiden käyttäminen on tuttua jo pidemmältä ajalta, mikä nostaa kynnystä hyödyntää hänelle vierasta menetelmää. Vastaavasti Domain-asiantuntija B:llä oli aikaisempaa kokemusta PowerBI:llä tehtävistä relaatiosta data-entiteettien välillä ja siksi hän löysi tuttuja piirteitä RML-menetelmästä.

Dataintegraattorit nostivat haastatteluissan esille useita tekijöitä, jotka korostavat itse linkitysmenetelmien olevan sivuroolissa, käytännössä ainoastaan menetelmiä, jotka valitaan tapauskohtaisesti. Kaikista tärkeintä on ymmärtää domain-asiantuntijaa ja iteraation merkitys eri sidosryhmien välillä on erittäin tärkeää dataintegraation onnistumisen kannalta. Toisaalta voidaan ajatella, että tietomallien linkitysmenetelmiä hyödynnetään erilaisissa käyttötapauksissa, mikä vaikuttaa siihen, onko iteraatiota domain-asiantuntijan ja dataintegraattorin välillä olemassa muuten kuin määrittelydokumentissa olevat linkitykset. Tämä vaikuttaa olennaisesti, minkälaisia ominaisuuksia linkitysmenetelmältä vaaditaan. Jos domain-asiantuntija ja dataintegraattori kykenevät tiiviimpään iteraatioon dataintegraation toteuttamiseksi, on itse linkitysmenetelmä pienemmässä roolissa lopputuloksen kannalta. Vastaavasti tilanteissa, joissa integraatioissa vaadittava datan tietomallin tunteva domain-asiantuntija ei ole käytettävissä, määrittelydokumentin laatu sekä käytetty menetelmä korostuvat. Näiden pohjalta linkitysmenetelmien osalta yksi näkökulma on tietomallien linkityksien taso, tehdäänkö linkityksiä kansainvälisten viitekehysten välillä vai paikallisella tasolla. Tästä hyvänä esimerkkinä on Transmodel-viitekehyksen

sekä siitä tehty NeTEx-implementaatio sekä sen alueelliset profiilit. Näissä tapauksissa paikalliset sekä tapauskohtaiset vaatimukset ohjaavat myös tietomallien linkitystyötä. Kansainvälisellä tasolla tehtävät tietomallistandardien määrittelyt sekä mahdolliset linkitykset toisiin standardeihin ovat laajoja kokonaisuuksia, joiden elinkaari on pidempi sekä muutokset tapahtuvat hitaammin. Vastaavasti paikallisella tasolla tarpeet vaihtuvat useammin ja tietomalleja voi olla useampia linkitettävänä, kuten kuva 7.1 havainnollistaa. Lisäksi tämän työn liitteessä I on hyvä esimerkki tietomallien linkittämisprosessista, jossa eri toimijoilla on erilaiset motivaatiotekijät, joiden pohjalta he suhtautuvat tietomallien määrittelytyöhön sekä tehtäviin dataintegraatioihin.



Kuva 7.1: Tietomallien linkittämistä sekä määrittelytyötä tehdään eri tasoilla, jotka ohjaavat käytettäviltä menetelmiltä toivottavia ominaisuuksia.

Dataintegraattori A korosti haastattelussa sitä, että linkitystaulukko on helpommin lähestyttävä myös dataintegraattorin näkökulmasta, sillä eri linkityksien osalta on helpompaa ja sujuvampaa lukea mahdollisia selitteitä tai kommentteja, joita domain-asiantuntija on linkitystaulukkoon lisännyt. Varsinaisen määrittelydokumentin lisäksi molemmat dataintegraattorit nostivat esille, että tukimateriaali, kuten arkkitehtuurikuvaukset ja muu dokumentaatio auttavat kontekstin ymmärtämisessä. Dataintegraattorit vaikuttivat varsin avoimilta sen suhteen, minkälainen itse määrittelydokumentti on, sillä ammattitaito ja kokemus auttavat tarvittaessa hakemaan lisätietoa standardidokumenteista tai muista lähteistä. Dataintegraattorit olivat sitä mieltä, että ideaalitilanteessa RML-menetelmällä muodostettu RDF-määrittelydokumentti vähentää syntaktisia virheitä tietomallien linkityksissä sekä suoraan koneluettavana vähentää potentiaalisesti tarvittavan ohjelmoinnin määrää integraatioissa. Datan laatu sekä yhteiskäyttöisyys tarjottavan tiedon osalta ovat myös olleet tavoitteita kansainväliselle yhteistyölle tietomallien kehittämisessä. Myös Lusikka T. et al. toteavat tutkimusraportissaan[19], että tarjottavan matkatiedon on oltava laadukasta, semanttisesti yhteensopivaa ja laadukasta dataa, mikä vahvistaa RML-menetelmän ideaalitilanteessa tarjoamia mahdollisuuksia.

7.1.2 Linkitysmenetelmien hyödynnettävyys eri tietomallien osalta

Eri menetelmien hyödynnettävyyttä tarkastelen domain-asiantuntijoiden sekä dataintegraattorien haastatteluissa sekä lähdemateriaaleissa esitettyjen ominaispiirteiden pohjalta. Hyödynnettävyys ei ole yksiselitteistä, joten pohdin asiaa eri sidosryhmien näkökulmasta, tietomallilinkityksien skaalautuvuuden sekä evoluution näkökulmasta.

Linkitystaulukko oli haastateltavien mielestä RML-menetelmää yksinkertaisempi erityisesti yksinkertaisten linkityksien osalta. Tähän osittain vaikuttaa varmasti taulukko-ohjelmistojen yleisyys ja tuttu käyttöliittymä, mikä vähentää perehtymisen tarvetta itse ohjelmiston käyttöön.

Linkitettävien tietomallien monitasoisuus vaikuttaa haastatteluiden perusteella linkitysmenetelmien hyödynnettävyyteen. Domain-asiantuntijat nostivat esille, että yksinkertaisemmat tietotyyppien linkitykset sekä tietomallit on suoraviivaisempaa suorittaa linkitystaulukkoa käyttämällä. Linkitettävien tietomallien monitasoisuus sekä konseptuaalisten tietotyyppien jakautuminen eri tasoille osoittautui domain-haastattelijoille haastavaksi erityisesti linkitystaulukon osalta, kun tietotyyppien osalta tehtiin sisennyksiä taulukkorakenteeseen NeTEx-tietomallin osalta. Beurèe ja Know-

les esittävät linkitystaulukon ohessa käytettäväksi erilaisia UML-kaavioita sekä muita määrittelyä tukevia materiaaleja, mikä vastaa haastatteluissa esiintynyttä näkemystä tarvittavista tukimateriaaleista, jotka auttavat hahmottamaan monimutkaisia kokonaisuuksia. Linkitystaulukon osalta ongelmana nähtiin se, että se jättää tietyiltä osin avoimeksi sen, miten eri tasoja tulisi kuvata yhdenmukaisesti. Haastatteluiden sekä käytettyjen linkitysmenetelmien pohjalta voidaan päätellä, että linkitystaulukkoa on helpompaa hyödyntää jonkin tietyn kontekstin tietomallien linkittämisessä toisiinsa ja vaatii määrittelevien asiantuntijoiden tiivistä yhteistyötä linkityksien toteuttamiseksi. Linkitystaulukko on suurimmalle osalle lähestyttävämpi menetelmänä, joten muun tukimateriaalin, vahvan domain-osaamisen sekä yhteistyön avulla tietomallien linkitykset on suoraviivaista toteuttaa. Dataintegraattorit nostivat myös esille sen, että määrittelydokumentin muodolla ei heidän osaltaan ole suurta merkitystä, sillä tapauskohtaisesti he joutuvat rakentamaan integraatiota varten teknisen toteutuksen erikseen. Tämän osalta voidaan ajatella, että ei ole suurta merkitystä, jos linkitystaulukon määrittelyn pohjalta ei pystytä suoraan tekemään varsinaista teknistä implementaatiota dataintegraatioissa.

RML-menetelmä ja siihen liittyvät työkalut, kuten RMLEditor [21] auttavat hahmottamaan asioita visuaalisen työkalun avulla, mikä pienentää haastetta tietomallien eri konseptitasojen yhteensovittamisessa. Hyödynnettävyyden näkökulmasta haastateltavat nostivat esille erityisesti laajojen, monimutkaisten sekä pitkällä aikavälillä toteutettavien tietomallilinkityksien kehittämisen, jossa RML-menetelmällä on omat erityispiirteensä. Sen ehdottomiin vahvuuksiin kuuluu syntaktisesti eheän määrittelydokumentin muodostaminen, jota voidaan validoida. RML kykenee yhdistämään useita tietomalleja sekä niiden tietotyyppien välisiä linkityksiä toisiinsa, mikä lisää sen hyödynnettävyyttä tietomallien muuttuessa tai uusien tarvittavien linkityksien tekemisessä. Sen avulla on mahdollista myös löytää konseptuaalisen tason yhtäläisyyksiä eri tietomallien välillä, mikä nopeuttaa uusien linkityksien muodostamista, kun käytettävissä on laaja, aikaisemmin muodostettu graafitietokanta määrittelyistä. Linkitystaulukon osalta useamman tietomallin vertailu keskenään keskittyi lähinnä siihen, mikä tietomalleista toimii referenssinä muille tietomalleille, jotka viittaavat kyseiseen referenssitietomallin tietotyyppeihin. Varsinainen taulukossa suoritettu linkittäminen toteutetaan kahden tietomallin välillä.

Tietomallien linkityksien näkökulmasta suurimpana erona voidaan havaita, että RML-menetelmän lähtökohtana on jatkuvasti lisääntyvät linkitykset eri tietomallien RDF-dokumenteissa esitettyjen tietotyyppien välillä, mikä muodostaa edellytykset

tietomallikehityksen tehokkaalle evoluutiolle. Aina kehittyvän "tietämyksen"pohjalta on helpompi jatkaa muita linkityksiä, eikä ole tarvetta lähteä alusta alkaen liikkeelle. Vaikuttaisi siltä, että haastatteluidenkin osalta on tulkittavissa, että suurin kynnys on aloittaa kyseisen menetelmän käyttö, jos linkitystaulukon avulla on suoraviivaisempaa toteuttaa tarpeeseen sopiva tietomallien linkittäminen. Erityisesti paikallisella tasolla, omiin tarpeisiin tehtävät tietomallien linkitykset tehdään todennäköisesti ratkaisukeskeisemmin, jolloin ei välttämättä ole mielekästä käyttää aikaa linkitysmetelmään perehtymiseen.

7.2 Määrittelyprosessin suhde käytettyihin linkitysmenetelmiin

Tässä aliluvussa käydään läpi tutkimuksen havaintojen ja teorian pohjalta esille tuotujen asioiden vaikutuksia valittavan linkitysmenetelmän ja määrittelyprosessin keskinäiseen riippuvuuteen. Kohdassa 7.2.1 pohditaan määrittelyprosesseja ja kohdassa 7.2.2 pohditaan linkitysmenetelmän valintaa ohjaavia tekijöitä. Viimeisessä kohdassa 7.2.3 käsitellään yhteisön toiminnassa huomiotavia tekijöitä.

7.2.1 Määrittelyprosessit

Linkitystaulukkoa käytettiin lähdemateriaaleissa osana prosessia, jossa ylätasolla hahmotetaan vertailtavien tietomallien määritelmiä sekä konsepteja toisiinsa, jotta voidaan hahmottaa osa-alueet, joiden osalta voidaan toteuttaa tarkempia linkityksiä. Tarkemmalla tasolla tehtävät tietotyypit sekä niiden attribuuttien vertailu tehdään linkitysmenetelmän määrittämän prosessin mukaisesti, joten linkitystaulukko on olennainen osa esitettyä määrittelyprosessia. Vastaavasti RML-menetelmä keskittyy tekniseen tietomallien linkittämiseen, eikä ota kantaa eri sidosryhmien väliseen toimintaan. Itse linkitysmenetelmän osalta tekninen prosessi on varsin suoraviivainen sisältäen mahdollisen lähdedatan lukemisen tai valmiin RDF-dokumentin luennan, jonka jälkeen toteutetaan linkittäminen sekä uuden määrittelydokumentin tuottaminen.

Mielenkiintoista on se, ettei kumpikaan haastateltavien käyttämistä linkitysmenetelmistä ota varsinaisesti kantaa itse sidosryhmien väliseen yhteistyöhön sekä heidän väliseen yhteistoimintaan. Erityisesti haastatteluissa dataintegraattorit korostivat iteraation merkitystä domain-asiantuntijoiden kanssa, mikä oli haastateltujen dataintegraattoreiden mukaan tärkeämpää kuin itse käytettävä linkitysme-

netelmä. Linkitysmenetelmien esittelyosioissa esiteltiin LOT-menetelmän[22], jossa keskiössä oli kokonaisprosessi tietomallien linkittämiseksi ja käsitteli eri sidosryhmien roolit, työvaiheet sekä tuotokset eri vaiheissa. Dataintegraattorien kommenttien ja perusteluiden pohjalta LOT-menetelmää tulisi tarkastella vastaisuudessa tarkemmin. LOT-menetelmässä hyödynnettiin RDF-dokumentteja määrittelyaineistona, joten mahdollisesti RML-menetelmää voisi täydentää LOT-menetelmän määrittelyprosessin vaiheistuksilla.

7.2.2 Kriittisimmät tekijät linkitysmenetelmän valinnassa

Linkitysmenetelmän valitsemiseksi tulisi tarkastella lähdemateriaalin, haastatteluiden sekä omien havaintojen perusteella seuraavia asioita:

- Millä tasolla (kansainvälinen stadardityö, paikallinen linkitys) tietomallien linkitystä toteutetaan?
- Linkitettävien tietomallien määrä
- Määrittelytyöhön osallistuvien sidosryhmien osaamisprofiili
- Tietomallien väliseen linkittämisprosessin kesto

Tietomallien linkittämisen taso vaikuttaa siihen, minkälaisia ominaisuuksia linkitysmenetelmältä toivotaan ja minkälaiseen kokonaisprosessiin sitä sovelletaan. Esimerkiksi kohdennetun domain-alueen kansainvälinen tietomallistandardityö voidaan toteuttaa hyvin linkitystaulukkoa käyttämällä, kun työhön osallistuvien domain-osaajien määrä on rajattu ja linkitystaulukossa mahdollisesti esiintyvät implisiittiset kuvaukset kyetään selittämään ymmärrettävästi. Mikäli kansainvälisellä tasolla tehtäisiin laaja-alaista eri domaineja yhdistävää tietomallien linkittämistä, olisi selkeästi tarvetta useampaa lähdemallia tukevalle linkitysmenetelmälle. Lisäksi todennäköisyys sille, että eri tietomallien välillä on merkitykseltään vastaavia tietotyyppejä tietomallien risteymäkohdissa (päällekkäisyys) kasvaa. Näin ollen linkitysmenetelmälle kohdistuu tarpeita sopeutua modulaarisesti ja tukea inkrementaalista tietomallien linkittämistä. Vastaavasti paikallisella tasolla tehtävässä tietomallien linkittämisessä on erilaiset tavoitteet, jotka kohdistuvat todennäköisesti lähdetietomallien yhteensovittamiseksi standardi- tai viitekehysmallia vasten tai paikallisen tason tarpeisiin toteutetun tietomalliimplementaation linkittämiseksi.

On myös huomioitava, että tietomallien määrä sekä sidosryhmien laajuus puoltavat sitä, että linkitysmenetelmää on mahdollista käyttää modulaarisesti sekä hyödyntää mahdollisimman paljon automatisoituja koneellisia menetelmiä. Esimerkiksi on haastavampaa toteuttaa sadan toimijan käyttämien tietomallien linkitys linkitystaulukolla kuin pyytää jokaista toteuttamaan RDF-dokumentti valitulla työkalulla, jonka jälkeen näistä muodostettaisiin graafitietokanta esimerkiksi RML-menetelmää käyttäen.

Tietomallien linkitysmenetelmään vaikuttavat myös osallistuvien sidosryhmien osaamisprofiili sekä taustatekijät. Kuten haastatteluissa dataintegraattori A totesi, tekijöiden ja työhön osallistuvien sidosryhmien tulisi itse tapauskohtaisesti määritellä käytettävät linkitysmenetelmät. Todennäköisesti kansainvälistä tietomallien määrittely- ja linkitystyötä toteuttavat kokeneet domain-osaajat, joilla on eri osaamisprofiili verrattuna paikallisen tason dataintegraatioon, jossa vastaavasti näkökulma on huomattavasti teknisempi. Voidaan siis todeta, että erityisesti sillä on merkitystä, lähestytäänkö määrittelyprosessia ylätason konseptuaalisella tasolla vai lähempänä implementaatiotasoa.

Haastatteluissa dataintegraattori B nosti esille, että aina ei ole aikaa miettiä uutta tietomallien linkitysmenetelmää, vaan yksinkertaisesti käytettävissä oleva aika ja resurssit määrittävät etenemistavan. Mikäli paikallisen tason integraatiota varten tulee tuottaa eri tietomallien linkittäminen, usein valitaan osallistuville sidosryhmille tutuin menetelmä. Kuten domain-asiantuntijat totesivat, että aikaisemman kokemuksensa pohjalta jompi kumpi linkitysmenetelmä tuntui heistä tutummalta ja helpommin omaksuttavalta menetelmältä. Vastaavasti pitkäjänteisemmän määrittely- ja linkitystyön toteuttamisessa on todennäköisesti enemmän aikaa huomioida käytettävät menetelmät ja prosessit sekä halutaan välttää pitkäjänteisessä työssä esille mahdollisesti esille tulevat linkitysmenetelmien rajoitteet.

7.2.3 Yhteisön toiminnassa huomioitavat tekijät

Tietomallien linkittämistä tehdään eri tasoilla, kuten kuvassa 7.1 on esitetty. Riippuen siitä, tehdäänkö paikallisen tason linkitystä vai onko kyseessä laajemman kansainvälisen standardien välisestä linkityksestä eri asiat määrittävät tekijät, jotka tulisi huomioida linkitykseen osallistuvien keskuudessa onnistuneen lopputuloksen saavuttamiseksi. Aliluvussa 7.2.2 esitettyjen tekijöiden lisäksi dataintegraattorit nostivat esille iteraation merkityksen laadukkaan lopputuloksen saavuttamiseksi. Varsinaisella linkitysmenetelmällä on pienempi rooli kuin toimivalla iteraatiolla eri si-

dosryhmien kesken. Lyhytkestoisessa tai kertaluonteisessa projektissa iteraatio ja osallistuvien sidosryhmien tottumukset ohjaavat menetelmän valintaa. Vastaavasti pitkäkestoisessa, esimerkiksi standardien ja viitekehyksien välisissä linkityksissä, toimivan prosessin sekä prosessin muuntautumiskyvyn merkitys kasvaa. Näin saavutetaan tilanne, jossa pitkäkestoisessa työssä voidaan jatkuvasti kehittää ja parantaa linkityksessä käytettävää prosessia. Valittavan linkitysmenetelmän tulisi mukautua linkitysprosessin mukaan ja mahdollisesti linkitysmenetelmässä esiintyviä heikkouksia voidaan paikata tehokkaalla ja toimivalla iteraatiolla. Yhteenvetona voidaan todeta, että laadukkaan tietomallien linkityksen muodostamiseksi projektin alussa on huomioitava aikaisemmin esitetyt linkitysmenetelmään vaikuttavat tekijät, tehokas iteraatio eri sidosryhmien välillä sekä linkitysprosessille asetettava tavoite.

7.3 Yhteistyö, standardit ja iteraatio eri sidosryhmien välillä

Toimintakentän muutokset joukkoliikenteessä, jossa kasvaneet asiakastarpeet sekä palvelutason laatu ohjaavat palveluntarjoajien toiminnan kehittämistä laadukkaammaksi ja sitä kautta parempien palveluiden tarjoamisessa loppuasiakkaalle, asettavat korkeampia vaatimuksia joukkoliikenteen tietojärjestelmien käyttämälle datalle. Kuten aikaisemmin on esitetty MMTIS-asetus velvoittaa eri joukkoliikennepalveluiden toimijoita jakamaan dataa asetuksen mukaisesti ja käytännössä tarkoitusta varten muodostetulla NeTEx-tietomallin mukaisella datalla. On kuitenkin huomionarvoista, että NeTEx-tietomalli ei itsessään vielä tuota eri toimijoiden käyttämien tietomallien pohjalta muunnettua NeTEx-tietomallin mukaista dataa. Tarvitaan määritetyn tietomallin mukaisesti myös datan migraatioon soveltuvia muuntimia, jotka vastaavasti tarvitsevat linkitykset tietomallien tietotyyppien välillä. Tässä työssä käytetyt menetelmät sekä niiden lähdemateriaali tarkastelee tätä näkökulmaa. On tunnistettavissa seuraavat osakokonaisuudet:

- Tietomallien kehitystyö
- Tietomallien linkitysmenetelmät
- Tietomallien linkitysprosessit (Toiminta sekä applikaatiokerroksen prosessit)
- Iteraatio eri sidosryhmien välillä
- Sidosryhmien tavoitteet sekä visio

Edellä olevan listan jokainen asia on oma kokonaisuutensa, joilla on omat tavoitteensa, näkökulmansa sekä omat sidosryhmänsä, jotka osallistuvat kyseisen osakokonaisuuden edistämiseen. Esimerkiksi tietomallien kehitystyössä domain-asiantuntijat mallintavat sekä muodostavat tarkasteltavasta näkökulmasta reaalimaailman tilaa kuvaavan tietomallin. Käytännössä edellisellä tarkoitan sitä, että esimerkiksi aikaisemmin esitetty GTFS- ja NeTEx-tietomalli tarkastelevat samaa kontekstia hyvin erilaisista näkökulmista ja eri sidosryhmien tarpeita ja tavoitteisiin pyrkien. Tietomallien linkitysmenetelmät sisältävät myös erilaisia näkökulmia, kuten olen tässä työssä havainnut esimerkiksi Linkitystaulukon sekä RML-menetelmien välillä. Tavoitteet ovat samankaltaisia, mutta linkitysmenetelmien näkökulmat ja periaatteet voivat erota merkittävästi. Linkitysprosessit hyödyntävät linkitysmenetelmiä ja niihin osallistuvat sidosryhmät voivat olla toisiinsa nähden päällekkäisiä tai erota toisistaan. Voidaan ajatella, että tietomallit keskittyvät johonkin tietyn sidosryhmän asettaman tavoitteen mukaiseen kontekstiin ja näkökulmaan ja tietomallien linkitysprosesseissa (hyödyntäen linkitysmenetelmiä) haetaan vastaavuuksia joko merkitykseltään päällekkäisille tietotyypeille tai eri tietomallien "reunojen välillä". Mielestäni tätä asiaa konseptuaalisella tasolla kuvattiin erinomaisesti Beurèe ja Knowles toimesta Linkitystaulukon menetelmäosuudessa [3]. Haastatteluissa nousi esille iteraation merkitys osana linkitysprosessia ja mielestäni kaikissa asioissa sen merkitys on suuri. Esimerkiksi EU:n tasolla muodostettu MMTIS-asetus on syntynyt joukkoliikenteen eri käyttäjäryhmien tarpeet täyttävien ratkaisujen muodostamiseksi sekä sen valmistelu on sisältänyt iteraatiota eri sidosryhmien kesken, joten käytännössä iteraatiota sekä vaikuttamista tapahtuu kaikkialla. Sidosryhmien tavoitteet ja visio eri asioihin liittyen ohjaavat muutostarpeita niin toimintaympäristön kuin niitä tukevien tietojärjestelmienkin osalta.

Tietomallien linkitysmenetelmät sekä niihin liittyvät linkitysprosessit ovat osa isoa kokonaisuutta, jossa kaikki tekijät vaikuttavat toisiinsa. On mahdotonta löytää täydellistä yksittäistä ratkaisua, joka soveltuisi jokaiseen käyttötapaukseen. Linkitysmenetelmien lisäksi haastatteluissa sekä käytännön linkityksissä nousi esille havaintoja, että linkitysprosessi on myös suuressa roolissa ja tulisi tarkastella osana itse linkitysmenetelmän käyttöä. Aikaisemmin mainituista menetelmistä LOT-menetelmässä oli huomioitu täsmällisimmin eri prosessin vaiheet, menetelmän käyttö sekä sidosryhmät rooleineen. Erityisesti dataintegraattorien kanssa tehdyissä haastatteluissa esille tuotu iteraatio eri sidosryhmien välillä sekä tarve "generalistille", joka kykenee kommunikoimaan sekä domain-asiantuntijoiden sekä dataintegraatto-

rien kanssa heidän ymmärtämällään tavalla. Tämä "generalisti", jolla dataintegraattori B toimijan nimesi, vastaa rooliltaan juuri LOT-menetelmän esittämää "Ontology developer-roolia, joka kykenee toimimaan tässä työssä haastateltujen sidosryhmien välillä. Erityisesti pitkäkestoisempien kansallisen tai kansainvälisen tietomallityön sekä linkityksien toteuttamisessa tulisi tarkastella LOT-menetelmässä käytettyä prosessia, sillä se sisältää RML-menetelmässäkin käytettyjä RDF-dokumentteja sekä mahdollistaa RML-menetelmänkin tuottaman hyödyn jatkuvalle graafitietokannan muodostamiselle. Erityisenä huomiona myös se, että LOT-menetelmässä on esitetty iteraatio sekä toisteisuus, jossa menetelmässä opitaan iteraation kautta, mikä muodostaa jatkuvan parantamisen sekä kehittämisen edellytykset.

8 Yhteenveto ja johtopäätökset

Joukkoliikenteen palvelutasovaatimuksien kasvu sekä strategisten tavoitteiden saavuttaminen vaativat yhteensopivia järjestelmiä, jotka tukevat monipuolisia jakamistalouden palveluita. Tässä työssä selvitettiin eri joukkoliikenteen tietomallien linkitysmenetelmien eroavaisuuksia ja niiden ominaisuuksien merkitystä tietomallien linkittämisen prosessissa, laadussa ja semanttisen merkityksen täsmällisyydessä, mikä on edellytyksenä yhteensopivan dataekosysteemin ja palveluiden muodostumiselle.

Tutkimuskysymyksinä esitettiin:

- Miten pienen perehtymisen jälkeen domain-osaajien on mahdollista tuottaa linkittämiseen vaadittava määrittelydokumentti?
- Miten paljon teknistä osaamista määrittelydokumentin muodostaminen vaatii domain-osaajalta?
- Mikä on linkityksessä tuotetun määrittelydokumentin jatkohyödynnettävyys toisissa linkityksissä?
- Mitkä tekijät määrittelydokumenteissa vaikuttavat yhdistetyn datan laatuun?

Työn teoriaosuudessa käytiin läpi joukkoliikenteen palvelutietoja kuvaavien yleisten tietomallistandardien ja -implementaatioiden ominaispiirteitä. Tietomallien osalta esitettiin myös näkökulmia, joiden pohjalta tietomallit ovat kehittyneet. Tietomallien lisäksi esitettiin kirjallisuuden perusteella joukkoliikenteen tietomallien linkityksissä käytettyjä menetelmiä.

Työn empiirisessä vaiheessa suoritettiin joukkoliikenteen domain-asiantuntijoiden kanssa kahdella valitulla linkitysmenetelmällä harjoitukset. Linkitysmenetelmiksi valittiin linkitystaulukko ja RML-menetelmä, sillä ne ovat menetelminä erilaisista näkökulmista kehitettyjä. Dataintegraattorit tutustuivat domain-haastattelijoiden kanssa tuotettuihin määrittelydokumentteihin ja valittuihin linkitysmenetelmiin. Tämän jälkeen molemmat ryhmät haastateltiin.

Tutkimusmenetelmänä käytettiin tapaustutkimusta, josssa tapauksina olivat valitut linkitysmenetelmät ja niitä sovellettiin joukkoliikenteen kahden tietomallin vä-

listen entiteettien linkityksissä. Tiedonkeruumenetelminä toimivat tehdyt linkitysharjoitukset sekä semi-strukturoidut haastattelut domain-asiantuntijoiden ja dataintegraattorien kanssa. Haastatteluista nauhoitettiin tallenteet, jonka pohjalta data analysoitiin litteroimalla ja luokittelemalla keskeiset havainnot.

Päätuloksina todettiin, että erityisesti domain-osaajan aikaisempi kokemus sekä tietotekninen taitotaso vaikuttavat siihen, miten helposti valittavalla linkitysmenetelmällä saadaan tuotettua määrittelydokumentti. Toisen tutkimuskysymykseen ei ole yksiselitteistä vastausta. Linkitystaulukko on yleisesti helpommin lähestyttävä ei-teknisen henkilön toimesta, mikäli halutaan varmemmin tuotettu määrittelydokumentti dataintegraattorien käytettäväksi. RML-menetelmän käyttökokemusta voidaan parantaa yhteensopivilla käyttöliittymäeditoreilla. Dataintegraattorit korostivat, että paikallisen tason linkityksissä linkitysmenetelmällä ei ole suurta merkitystä, vaan tärkeintä yhteisen ymmärryksen muodostamiseksi on iteraatio domain-asiantuntijan ja dataintegraattorin välillä. Kolmannen tutkimuskysymyksen osalta todettiin, että RML-menetelmällä tuotettu määrittelydokumentti on helpommin tarkistettavissa syntaktisten virheiden varalta ja lisäksi se mahdollistaa paremman jatkohyödynnettävyyden erilaisissa graafitietokantojen avulla tehtävissä linkityksissä. Yleisesti haastateltavat kokivat, että linkitystaulukko tarjoaa paremmin mahdollisuuden lisätä implisiittisiä kommentteja linkityksien osalta, mikä tukee iteraatiota eri sidosryhmien välillä. Viimeisen tutkimuskysymyksen löydettiin kaksi tekijää, jotka vaikuttavat yhdistetyn datan laatuun. Määrittelydokumentin syntaktinen laatu on tärkeää, kun määrittelydokumenttia käytetään suoraan teknisessä dataintegraatiossa. Toisena tekijänä tunnistettiin määrittelydokumentin kyky tukea sidosryhmien iteraatiota. Vertailtavista linkitysmenetelmistä on mahdotonta arvioida paremmuutta yksiselitteisesti, sillä usean asian kokonaisuus määrittelee, mikä on paras menetelmä käytettäväksi.

Validiteetin näkökulmasta tapaustutkimuksessa pyrittiin tarkastelemaan eri näkökulmista tietomallien linkitysmenetelmiä tarkastelevia lähteitä, jotta tunnistettiin pääasialliset linkitysmenetelmät ja niitä voitiin luokitella menetelmien rajaamiseksi. Tutkimuksen kohde rajattiin käsittelemään joukkoliikenteen tietomalleja sekä niissä hyödynnettyjä linkitysmenetelmiä. Domain-asiantuntijoiden ja dataintegraattorien haastetteluilla pyrittiin tuomaan eri tarkastelunäkökulmia linkitysmenetelmien arvioimiseksi.

Tutkimuksen päätuloksissa nostettiin esille havainto, että iteraatio linkitysprosessiin osallistuvien sidosryhmien osalta on suuressa roolissa lopputuloksen laa-

dun kannalta. Työssä esitettiin LOT-menetelmä, joka olisi havaintojen perusteella huomioinut iteraatiota eri sidosryhmien välillä ja sisälsi myös RDF-graafeihin perustuvan määrittelydokumentin hyödyntämisen. Tämän työn perusteella LOT-menetelmä voisi tarjota valmiin menetelmän laajempien tietomallien linkittämiseksi ja huomioiden työn tuloksissa esitetyt havainnot.

Lähteet

[1] BELLINI, P., BILOTTA, S., COLLINI, E., FANFANI, M., JA NESI, P. Data sources and models for integrated mobility and transport solutions. *Sensors 24*, 2 (2024).

[2] BEYOND TRANSPARENCY. Pioneering open data standards: The GTFS story. URL `https://beyondtransparency.org/chapters/part-2/pioneering-open-data-standards-the-gtfs-story/`, viitattu 13.8.2022.

[3] BOURÈE, K., JA KNOWLES, N. Methodology for comparing data standards. URL `https://data4pt-project.eu/wp-content/uploads/2021/03/Data4PT-Methodology-for-comparing-data-standards.pdf`, viitattu 28.8.2022.

[4] CEFRIEL. Chimera: Composable semantic data transformation. URL `https://github.com/cefriel/chimera`, viitattu 9.12.2022.

[5] CEN TS 16614-1. *Public transport - Network and Timetable Exchange (NeTEx) - Part 1: Public transport network topology exchange format*, 2020.

[6] CEN TS 16614-2. *Public transport - Network and Timetable Exchange (NeTEx) - Part 2: Public transport scheduled timetables exchange format*, 2020.

[7] CEN TS 16614-3. *Public transport - Network and Timetable Exchange (NeTEx) - Part 3: Public transport fares exchange format*, 2020.

[8] CRESWELL, J. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. SAGE Publications, Thousands Oaks, 2012.

[9] DIMOU, A., SANDE, M. V., COLPAERT, P., VERBORGH, R., MANNENS, E., JA WALLE, R. V. D. RML: A generic language for integrated RDF mappings of heterogeneous data. Julkaisusarjassa *Proceedings of LDOW2014 – Linked Data on the Web* (Seoul, Korea, Elokuu 2014), CEUR-WS, 1 – 5.

[10] GTFS.ORG. GTFS schedule overview. URL `https://gtfs.org/schedule/`, viitattu 5.9.2022.

[11] HOHPE, G., JA WOOLF, B. *Enterprise integration patterns: Designing, building, and deploying messaging solutions*. Addison-Wesley Professional, Boston, 2004.

[12] LAINE, M., BAMBERG, J., JA JOKINEN, P. *Tapaustutkimuksen käytäntö ja teoria.* Gaudeamus, 2007, ss. 9–38.

[13] MARTIN DAVIS. Data model diagrams for GTFS. URL http://lin-ear-th-inking.blogspot.com/2011/09/data-model-diagrams-for-gtfs.html, viitattu 28.8.2022.

[14] MAXWELL, J. Understanding and validity in qualitative research. *Harvard Educational Review 62* (01 1992), 279–300.

[15] MYLONAS, C., MITSAKIS, E., AYFANTOPOULOU, G., STAVARA, M., TZANIS, D., YANNIS, G., JA LAIOU, A. Harmonization of national access points to intelligent transport systems data: A data content and added value perspective. *Transportation Research Procedia 72* (2023), 2928–2935.

[16] NETEX. NeTEx part 3. URL https://netex-cen.eu/model/conceptual/part3/index.htm, viitattu 25.10 .2022.

[17] NETEX. Overview. URL https://netex-cen.eu/, viitattu 6.9.2022.

[18] ONTOLOGY ENGINEERING GROUP. Linked open terms. URL https://lot.linkeddata.es/, viitattu 22.11.2022.

[19] PIHLAJAMAA, O., LAHTI, J., HEINO, I., JA LUSIKKA, T. *Joukkoliikenteen matkatietopalveluiden digitaalinen infrastruktuuri: Selvitys kehittämistarpeista ja -toimista*. No. VTT-R-01216-20 in VTT Research Report. VTT Technical Research Centre of Finland, Finland, Oct. 2020.

[20] RML.IO. RML documentation. URL https://rml.io/specs/rml/, viitattu 27.11.2022.

[21] RML.IO. RMLEditor. URL https://rml.io/tools/rmleditor/, viitattu 24.10.2023.

[22] RUCKHAUS, E., ANTON-BRAVO, A., SCROCCA, M., JA CORCHO, O. Applying the LOT methodology to a public bus transport ontology aligned with trans-model: Challenges and results. *Semantic Web vol. 14*, no. 4 (2023), 639–657.

[23] SCROCCA, M., COMERIO, M., CARENINI, A., JA CELINO, I. Turning transport data to comply with eu standards while enabling a multimodal transport knowledge graph. Julkaisusarjassa *International Semantic Web Conference* (Athens, Greece, Marraskuu 2020), Springer, 411 – 429.

[24] TRANSMODEL. Data model diagrams. URL https://www.transmodel-cen.eu/model/, viitattu 21.9.2022.

[25] TRANSMODEL. Documentation faq. URL https://transmodel-cen.eu/, viitattu 31.8.2022.

[26] TRANSMODEL. Transmodel tutorial part 1-3. URL https://www.transmodel-cen.eu/wp-content/uploads/2015/01/TUTORIAL-Part1-3-v0.2-1.pdf, viitattu 21.9.2022.

[27] TRANSMODEL. Transmodel tutorial part 4. URL https://www.transmodel-cen.eu/wp-content/uploads/2015/01/TUTORIAL_Part4_v2.1-1.pdf, viitattu 21.9.2022.

[28] TRANSMODEL. Transmodel tutorial part 5. URL https://www.transmodel-cen.eu/wp-content/uploads/2019/10/TUTORIAL_Part5_v2.3-1.pdf, viitattu 21.9.2022.

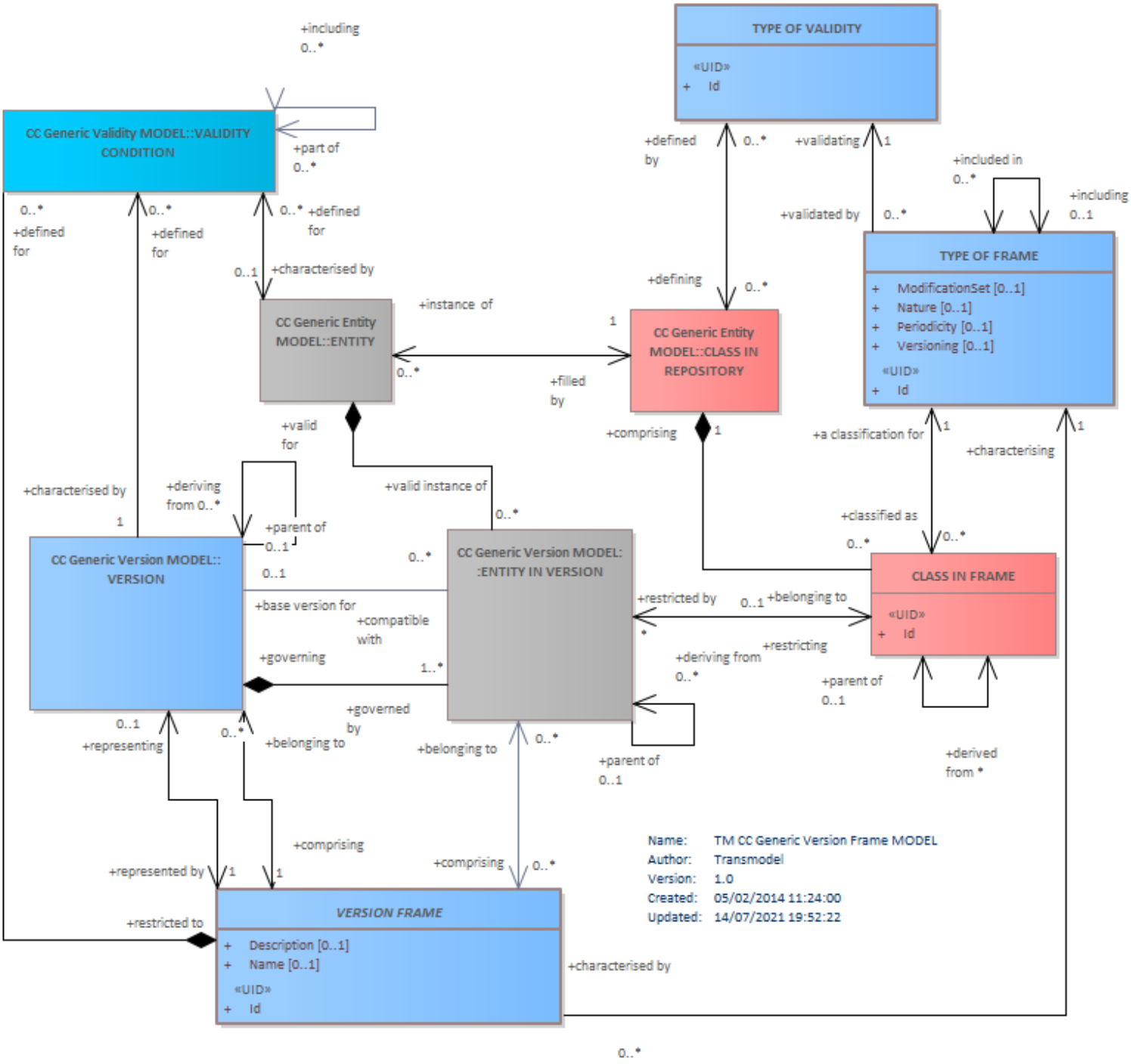
[29] TRANSMODEL. Transmodel tutorial part 6. URL https://www.transmodel-cen.eu/wp-content/uploads/2019/10/TUTORIAL_Part6_v2.2-1.pdf, viitattu 30.9.2022.

[30] TRANSMODEL. Transmodel tutorial part 7. URL https://www.transmodel-cen.eu/wp-content/uploads/2015/01/TUTORIAL_Part7_v2.1-1.pdf, viitattu 30.9.2022.

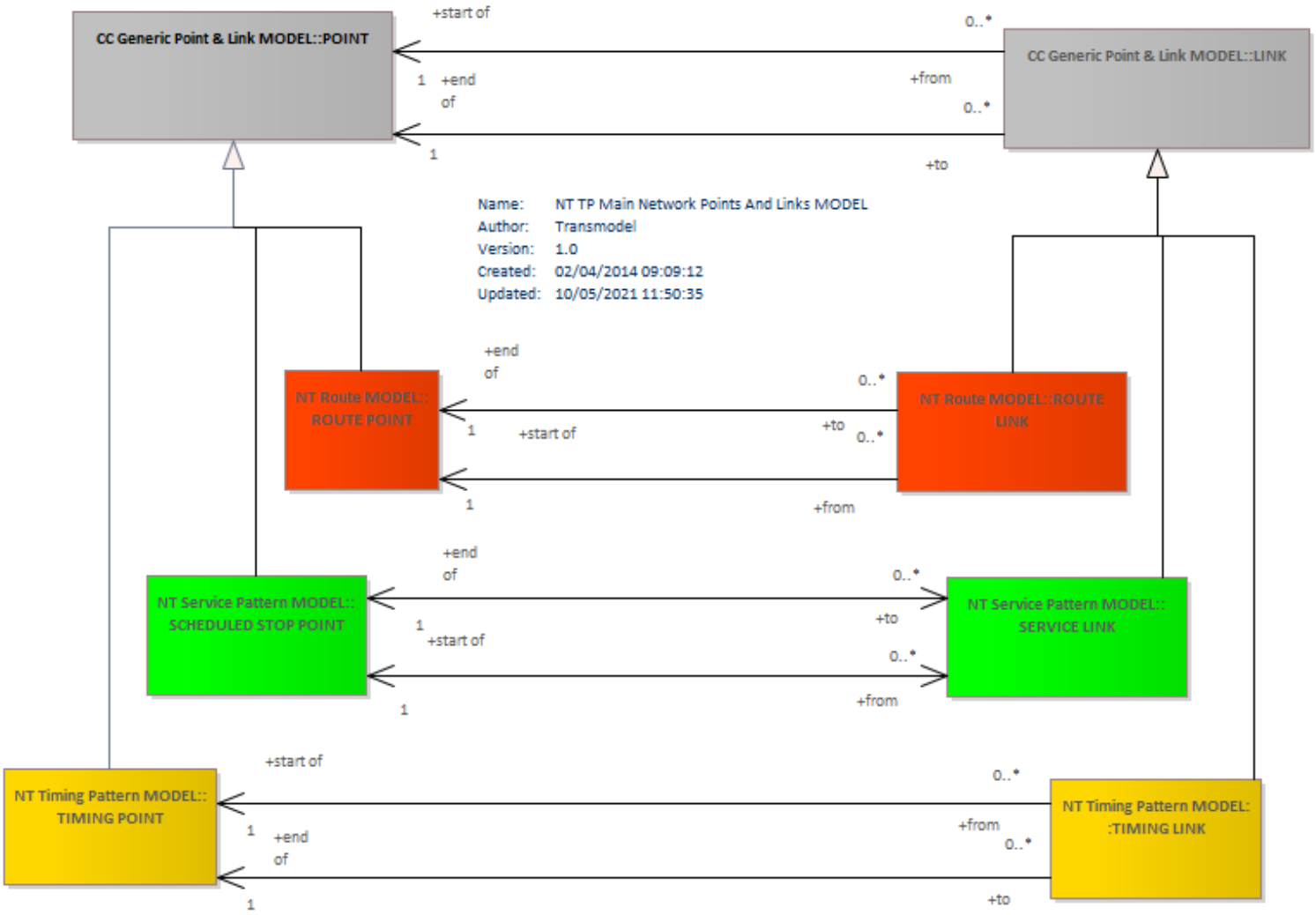
[31] VILKKA, H. *Tutkija ja kehitä*. Santalahti-kustannus, Jyväskylä, 2021.

[32] VUORI, J. Tapaustutkimus. URL https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/tutkimusasetelma/tapaustutkimus/, viitattu 29.4.2024.

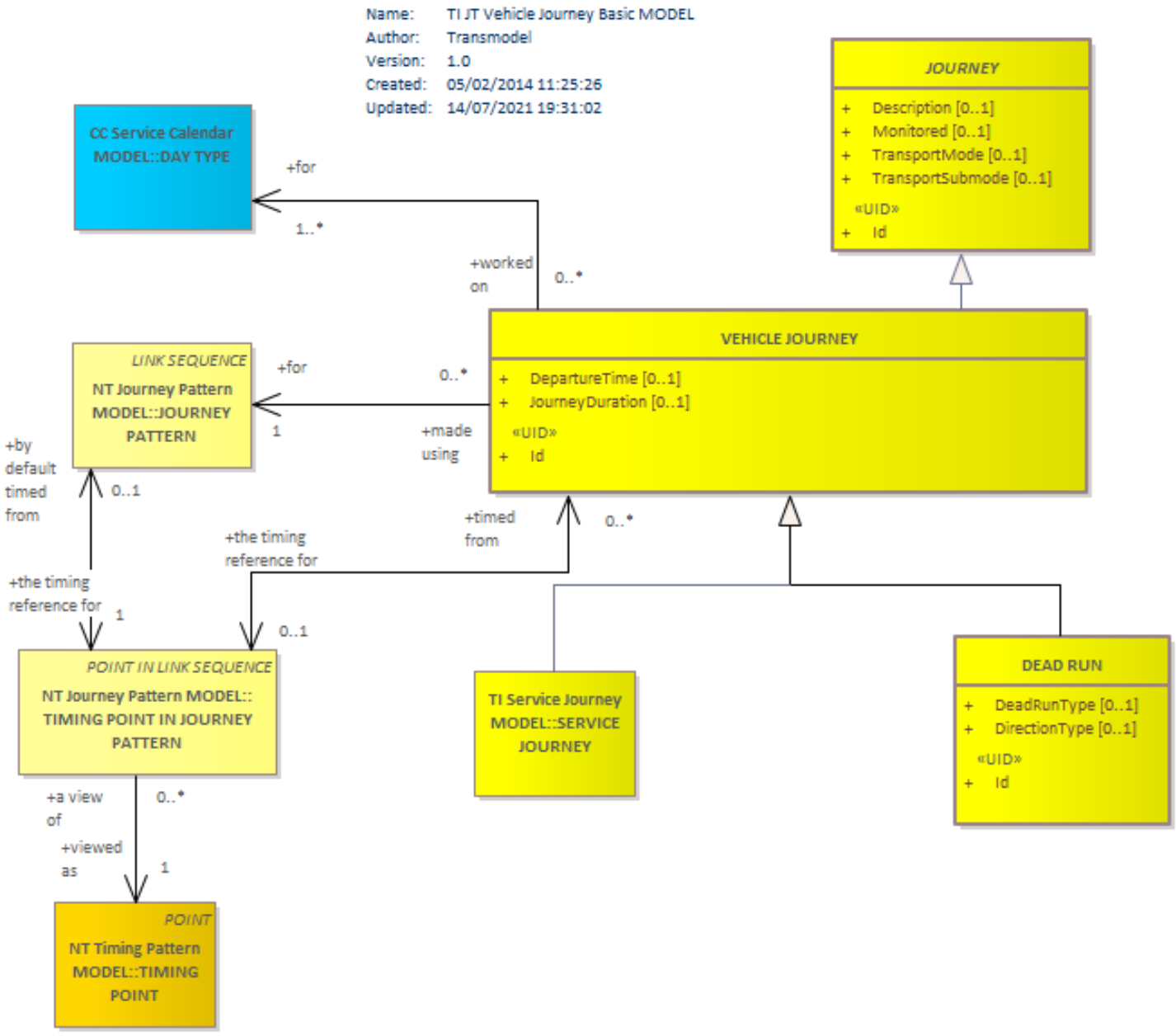
A Transmodel osa 1 geneerinen versiointimalli


TYPE OF VALIDITY
«UID»
+ Id
CC Generic Validity MODEL::VALIDITY CONDITION
+including
0..*
+part of
0..*
+defined by
0..*
+validating
1
+included in
0..*
+including
0..1
+validated by
0..*
TYPE OF FRAME
+ ModificationSet [0..1]
+ Nature [0..1]
+ Periodicity [0..1]
+ Versioning [0..1]
«UID»
+ Id
+defined for
+characterised by
+defining
+instance of
CC Generic Entity MODEL::ENTITY
CC Generic Entity MODEL::CLASS IN REPOSITORY
+filled by
+valid for
+comprising
+a classification for
+characterising
+characterised by
+deriving from 0..*
+valid instance of
+classified as
+parent of
CC Generic Version MODEL::VERSION
CC Generic Version MODEL::ENTITY IN VERSION
CLASS IN FRAME
«UID»
+ Id
+base version for
+compatible with
+restricted by
+belonging to
+governing
+restricting
+deriving from
+governed by
+parent of
+representing
+belonging to
+derived from *
+comprising
+represented by
+restricted to
VERSION FRAME
+ Description [0..1]
+ Name [0..1]
«UID»
+ Id
+characterised by
Name: TM CC Generic Version Frame MODEL
Author: Transmodel
Version: 1.0
Created: 05/02/2014 11:24:00
Updated: 14/07/2021 19:52:22

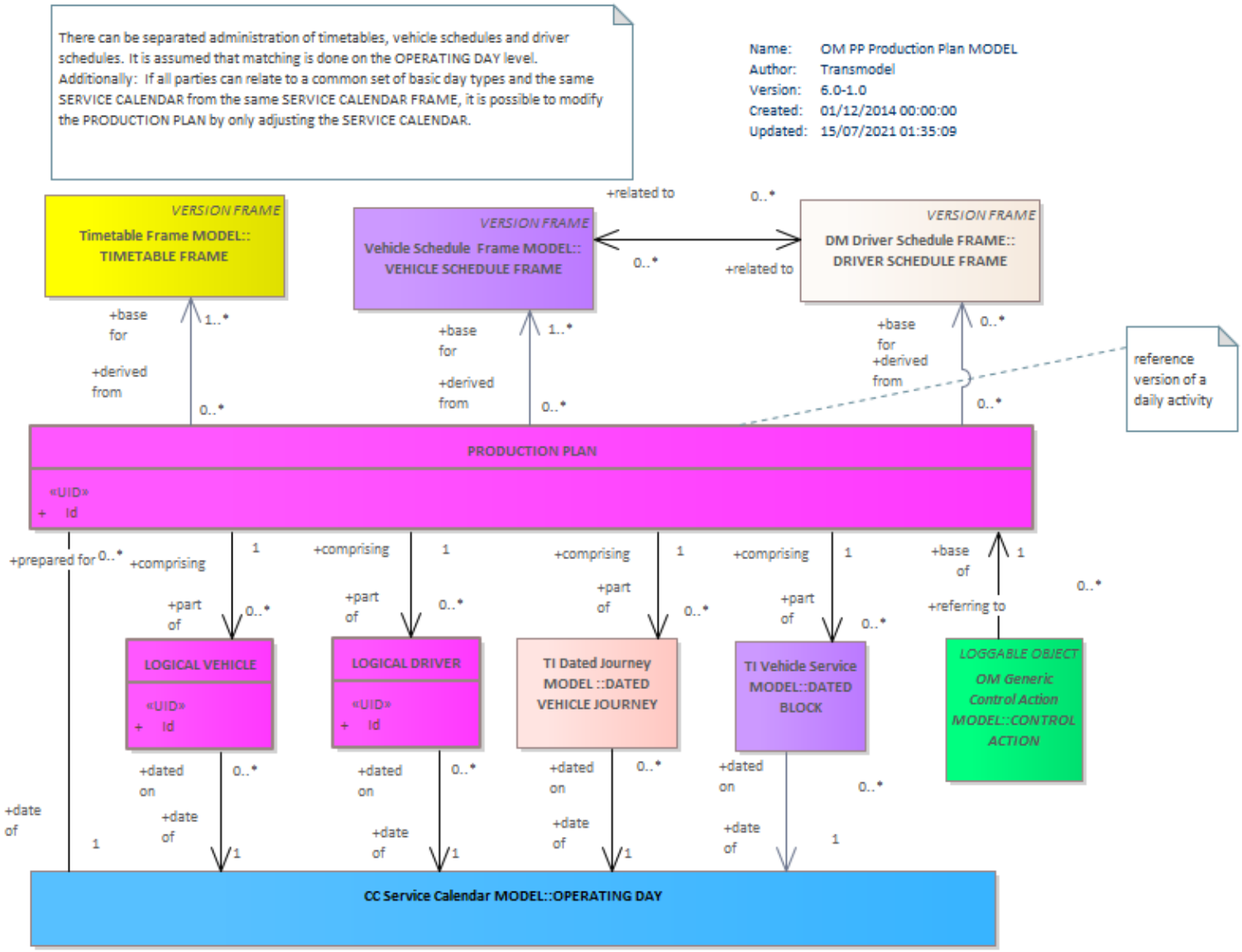
B Transmodel osa 2 useamman pisteen sekä linkin perivät tietotyypit


CC Generic Point & Link MODEL::POINT
+start of
0..*
1
+end of
+from
0..*
CC Generic Point & Link MODEL::LINK
1
+to
Name: NT TP Main Network Points And Links MODEL
Author: Transmodel
Version: 1.0
Created: 02/04/2014 09:09:12
Updated: 10/05/2021 11:50:35
+end of
NT Route MODEL:: ROUTE POINT
0..*
NT Route MODEL::ROUTE LINK
1
+start of
+to
0..*
1
+from
+end of
0..*
NT Service Pattern MODEL:: SCHEDULED STOP POINT
1
+start of
+to
NT Service Pattern MODEL:: SERVICE LINK
0..*
1
+from
+start of
0..*
NT Timing Pattern MODEL:: TIMING POINT
1
+end of
+from
NT Timing Pattern MODEL: :TIMING LINK
0..*
1
+to

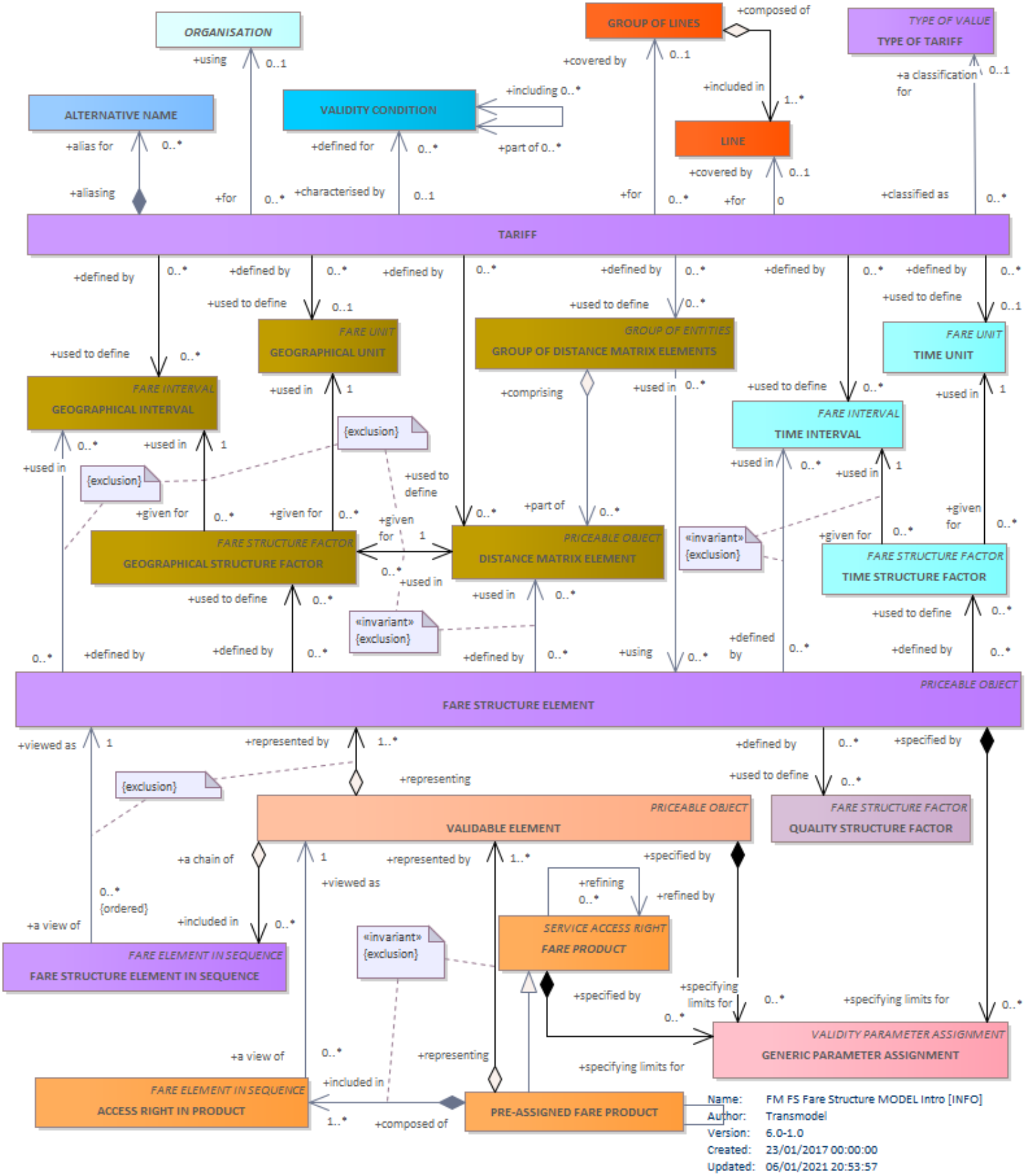
C Transmodel osa 3 määrittelemä `VEHICLE JOURNEY` -malli


Name: TI JT Vehicle Journey Basic MODEL
Author: Transmodel
Version: 1.0
Created: 05/02/2014 11:25:26
Updated: 14/07/2021 19:31:02
JOURNEY
+ Description [0..1]
+ Monitored [0..1]
+ TransportMode [0..1]
+ TransportSubmode [0..1]
«UID»
+ Id
CC Service Calendar MODEL::DAY TYPE
+for
1..*
+worked on
0..*
VEHICLE JOURNEY
+ DepartureTime [0..1]
+ JourneyDuration [0..1]
«UID»
+ Id
LINK SEQUENCE
NT Journey Pattern MODEL::JOURNEY PATTERN
+for
0..*
1
+made using
+by default timed from
0..1
+timed from
0..*
+the timing reference for
1
+the timing reference for
0..1
POINT IN LINK SEQUENCE
NT Journey Pattern MODEL:: TIMING POINT IN JOURNEY PATTERN
TI Service Journey MODEL::SERVICE JOURNEY
DEAD RUN
+ DeadRunType [0..1]
+ DirectionType [0..1]
«UID»
+ Id
+a view of
0..*
+viewed as
1
POINT
NT Timing Pattern MODEL::TIMING POINT

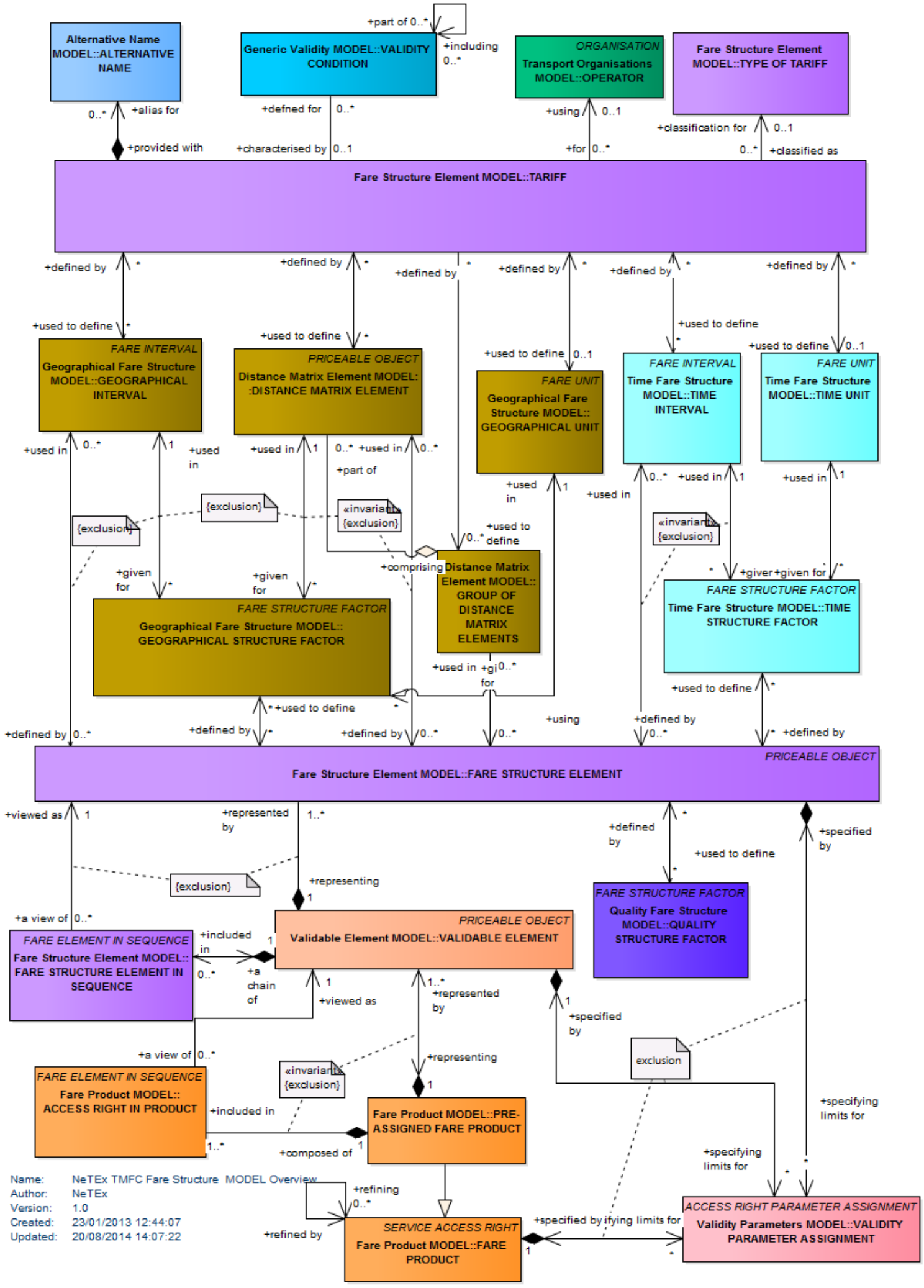
D Transmodel osa 4 tuotantosuunnitelma (`PRODUCTION PLAN`)


There can be separated administration of timetables, vehicle schedules and driver schedules. It is assumed that matching is done on the OPERATING DAY level. Additionally: If all parties can relate to a common set of basic day types and the same SERVICE CALENDAR from the same SERVICE CALENDAR FRAME, it is possible to modify the PRODUCTION PLAN by only adjusting the SERVICE CALENDAR.
Name: OM PP Production Plan MODEL
Author: Transmodel
Version: 6.0-1.0
Created: 01/12/2014 00:00:00
Updated: 15/07/2021 01:35:09
VERSION FRAME
Timetable Frame MODEL:: TIMETABLE FRAME
VERSION FRAME
Vehicle Schedule Frame MODEL:: VEHICLE SCHEDULE FRAME
VERSION FRAME
DM Driver Schedule FRAME:: DRIVER SCHEDULE FRAME
+related to
0..*
+related to
0..*
+base for
+derived from
1..*
0..*
+base for
+derived from
1..*
0..*
+base for
+derived from
0..*
0..*
reference version of a daily activity
PRODUCTION PLAN
«UID»
+ Id
+prepared for 0..*
+comprising
1
+part of
0..*
+comprising
1
+part of
0..*
+comprising
1
+part of
0..*
+comprising
1
+part of
0..*
+base of
1
0..*
+referring to
LOGICAL VEHICLE
«UID»
+ Id
LOGICAL DRIVER
«UID»
+ Id
TI Dated Journey MODEL ::DATED VEHICLE JOURNEY
TI Vehicle Service MODEL::DATED BLOCK
LOGGABLE OBJECT
OM Generic Control Action MODEL::CONTROL ACTION
+dated on
0..*
+date of
1
+dated on
0..*
+date of
1
+dated on
0..*
+date of
1
+dated on
0..*
+date of
1
+date of
1
CC Service Calendar MODEL::OPERATING DAY

E Tietomalli taksan muodostumisesta ja sen liitoksesta tuoterakenteeseen


ORGANISATION
+using
0..1
GROUP OF LINES
+composed of
+covered by
0..1
+included in
1..*
TYPE OF VALUE
TYPE OF TARIFF
+a classification for
0..1
ALTERNATIVE NAME
+alias for
0..*
+aliasing
VALIDITY CONDITION
+including 0..*
+part of 0..*
+defined for
0..*
LINE
+covered by
0..1
+for
0..*
+characterised by
0..1
+for
0
+classified as
0..*
TARIFF
+defined by
0..*
+used to define
0..1
FARE UNIT
GEOGRAPHICAL UNIT
FARE INTERVAL
GEOGRAPHICAL INTERVAL
+used in
1
GROUP OF ENTITIES
GROUP OF DISTANCE MATRIX ELEMENTS
+comprising
+used in
FARE UNIT
TIME UNIT
FARE INTERVAL
TIME INTERVAL
{exclusion}
+used to define
+given for
FARE STRUCTURE FACTOR
GEOGRAPHICAL STRUCTURE FACTOR
+part of
PRICEABLE OBJECT
DISTANCE MATRIX ELEMENT
«invariant»
{exclusion}
FARE STRUCTURE FACTOR
TIME STRUCTURE FACTOR
+defined by
+using
PRICEABLE OBJECT
FARE STRUCTURE ELEMENT
+viewed as
1
+represented by
1..*
+representing
{exclusion}
+specified by
+used to define
PRICEABLE OBJECT
VALIDABLE ELEMENT
FARE STRUCTURE FACTOR
QUALITY STRUCTURE FACTOR
+a chain of
0..*
{ordered}
+a view of
+included in
+refining
+refined by
SERVICE ACCESS RIGHT
FARE PRODUCT
FARE ELEMENT IN SEQUENCE
FARE STRUCTURE ELEMENT IN SEQUENCE
«invariant»
{exclusion}
+specifying limits for
VALIDITY PARAMETER ASSIGNMENT
GENERIC PARAMETER ASSIGNMENT
FARE ELEMENT IN SEQUENCE
ACCESS RIGHT IN PRODUCT
PRE-ASSIGNED FARE PRODUCT
+composed of
Name: FM FS Fare Structure MODEL Intro [INFO]
Author: Transmodel
Version: 6.0-1.0
Created: 23/01/2017 00:00:00
Updated: 06/01/2021 20:53:57

F NeTEx-mallin matkustusoikeuden määrittäminen taksatekijöiden avulla


Alternative Name MODEL::ALTERNATIVE NAME
Generic Validity MODEL::VALIDITY CONDITION
+part of 0..*
+including 0..*
ORGANISATION
Transport Organisations MODEL::OPERATOR
Fare Structure Element MODEL::TYPE OF TARIFF
0..* +alias for
+defined for 0..*
+using 0..1
+classification for 0..1
+provided with
+characterised by 0..1
+for 0..*
0..* +classified as
Fare Structure Element MODEL::TARIFF
+defined by *
+used to define *
FARE INTERVAL
Geographical Fare Structure MODEL::GEOGRAPHICAL INTERVAL
PRICEABLE OBJECT
Distance Matrix Element MODEL::DISTANCE MATRIX ELEMENT
+used to define 0..1
FARE UNIT
Geographical Fare Structure MODEL::GEOGRAPHICAL UNIT
FARE INTERVAL
Time Fare Structure MODEL::TIME INTERVAL
FARE UNIT
Time Fare Structure MODEL::TIME UNIT
+used in 0..*
+used in 1
+part of
{exclusion}
«invariant» {exclusion}
+used to define 0..*
+comprising
Distance Matrix Element MODEL::GROUP OF DISTANCE MATRIX ELEMENTS
+given for *
FARE STRUCTURE FACTOR
Geographical Fare Structure MODEL::GEOGRAPHICAL STRUCTURE FACTOR
FARE STRUCTURE FACTOR
Time Fare Structure MODEL::TIME STRUCTURE FACTOR
+used to define *
+using
+defined by 0..*
PRICEABLE OBJECT
Fare Structure Element MODEL::FARE STRUCTURE ELEMENT
+viewed as 1
+represented by 1..*
+representing 1
+specified by
FARE STRUCTURE FACTOR
Quality Fare Structure MODEL::QUALITY STRUCTURE FACTOR
+a view of 0..*
FARE ELEMENT IN SEQUENCE
Fare Structure Element MODEL::FARE STRUCTURE ELEMENT IN SEQUENCE
+included in 1
0..* +a chain of
PRICEABLE OBJECT
Validable Element MODEL::VALIDABLE ELEMENT
exclusion
FARE ELEMENT IN SEQUENCE
Fare Product MODEL::ACCESS RIGHT IN PRODUCT
+included in 1..*
Fare Product MODEL::PRE-ASSIGNED FARE PRODUCT
+composed of 1
+specifying limits for
+refining 0..*
+refined by
SERVICE ACCESS RIGHT
Fare Product MODEL::FARE PRODUCT
+specified by 1
+specifying limits for *
ACCESS RIGHT PARAMETER ASSIGNMENT
Validity Parameters MODEL::VALIDITY PARAMETER ASSIGNMENT
Name: NeTEx TMFC Fare Structure MODEL Overview
Author: NeTEx
Version: 1.0
Created: 23/01/2013 12:44:07
Updated: 20/08/2014 14:07:22

G Tietomallien linkitystaulukko

#	Source Class	A (O=Own; R=Relationship)	Source Attribute Relationship (blue)	Source Attribute type Simple Type Complex type Enumeration	Source multiplicity	Description (as in the Source)

Target elements = result of mapping

Target correspondence indication; comments	Corresponding Target class/attribute	Exact corresp. to Target class 1:1	Exact corresp. to Target attribute 1:1	Source specific - new (class.) 1:0	Source Additional attribute without contradiction 1:0	Belonging to a group of Source elements corresponding to a Target class N:1	Source element derived from several Target attributes 1:N	Other

H Esimerkki RML-määrittelydokumentista

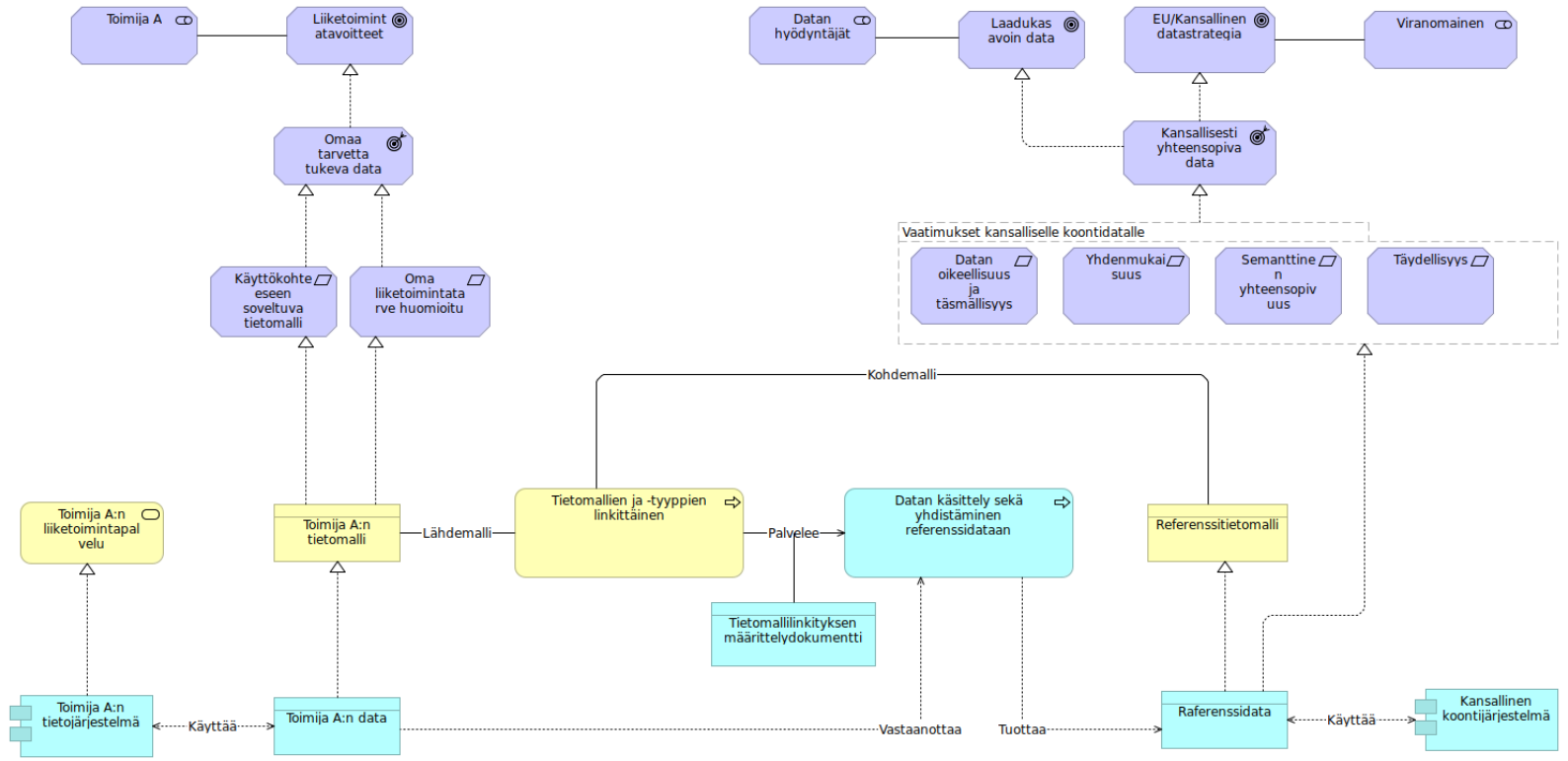
```
<#PerformancesMapping>
 rml:logicalSource [
  rml:source "http://ex.com/performances.json";
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$.Performance.[*]" ];
  rr:subjectMap [ rr:template "http://ex.com/{Perf_ID}" ];
 rr:predicateObjectMap [ rr:predicate ex:venue;
  rr:objectMap [ rr:parentTriplesMap <#VenueMapping> ] ];
 rr:predicateObjectMap [ rr:predicate ex:location;
   rr:objectMap [ rr:parentTriplesMap <#LocationMapping> ] ].

<#VenueMapping>
 rml:logicalSource [
  rml:source "http://ex.com/performances.json";
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$.Performance.Venue.[*]" ];
 rr:subjectMap [ rr:template "http://ex.com/{Venue_ID}" ].

<#LocationMapping>
 rml:logicalSource [ ......... ];
 rr:subjectMap [ rr:template "http://ex.com/{lat},{long}" ];
 rr:predicateObjectMap [ rr:predicate ex:long;
   rr:objectMap [ rml:reference "long" ] ]
 rr:predicateObjectMap [ rr:predicate ex:lat;
   rr:objectMap [ rml:reference "lat" ] ] .

<#ExhibitionMapping>
 rml:logicalSource [
  rml:source "http://ex.com/exhibitions.xml";
  rml:referenceFormulation ql:XPath;
  rml:iterator "/Events/Exhibition" ];
 rr:subjectMap [ rr:template "http://ex.com/{@id}" ];
  rr:predicateObjectMap [ rr:predicate ex:location;
   rr:objectMap [ rr:parentTriplesMap <#LocationMapping> ] ];
 rr:predicateObjectMap [ rr:predicate ex:venue;
   rr:objectMap [ rr:parentTriplesMap <#VenueMapping>;
    rr:joinCondition [
    rr:child "$.Performance.Venue.Name";
    rr:parent "/Events/Exhibition/Venue" ] ] ] .
```

I Linkittämisprosessi eri toimijoiden datan linkittämiseksi referenssimalliin


Toimija A
Liiketoimint atavoitteet
Datan hyödyntäjät
Laadukas avoin data
EU/Kansallinen datastrateqia
Viranomainen
Omaa tarvetta tukeva data
Kansallisesti yhteensopiva data
Vaatimukset kansalliselle koontidatalle
Datan oikeellisuus ja täsmällisyys
Yhdenmukaisuus
Semanttinen yhteensopivuus
Täydellisyys
Käyttökohteeseen soveltuva tietomalli
Oma liiketoimintatarve huomioitu
Kohdemalli
Toimija A:n liiketoimintapalvelu
Toimija A:n tietomalli
Lähdemalli
Tietomallien ja -tyyppien linkittäinen
Palvelee
Datan käsittely sekä yhdistäminen referenssidataan
Referenssitietomalli
Tietomallilinkityksen määrittelydokumentti
Toimija A:n tietojärjestelmä
Käyttää
Toimija A:n data
Vastaanottaa
Tuottaa
Raferenssidata
Käyttää
Kansallinen koontijärjestelmä