Tampereen yliopisto

Teemu Kontro

DATAN LAADUN HALLINTA SÄHKÖ-VERKON KÄYTTÖTOIMINNASSA

Diplomityö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastaja: Samuli Pekkola
Tarkastaja: Pasi Raatikainen
Syyskuu 2021

TIIVISTELMÄ

Teemu Kontro: Datan laadun hallinta sähköverkon käyttötoiminnassa
Diplomityö
Tampereen yliopisto
Tietojohtamisen DI-ohjelma
Syyskuu 2021

Datan määrä on kasvanut jatkuvasti datalähteiden määrän ja tallennusratkaisujen kehityksen myötä. Samalla datan laatu on muuttunut entistä tärkeämmäksi liiketoiminnassa, sillä virheellinen data voi aiheuttaa merkittäviä kustannuksia, heikentää organisaation mainetta tai hankaloittaa strategiatyötä. Lisäksi työntekijät käyttävät nykyään merkittävän osan ajastaan erilaisten poikkeamien korjaamiseen. Myös tietojärjestelmäpohjaiset automaatiohankkeet vaativat toimiakseen laadukasta dataa, sillä ihmiset eivät enää valvo kaikkia prosessin vaiheita.

Tämän tutkimuksen tarkoituksena oli perehtyä Fingrid Oyj:n kantaverkon käyttö- ja tiladatan laatuongelmiin sekä tarjota niihin kehitysehdotuksia aiemman kirjallisuuden pohjalta. Työ koostuu kirjallisuuskatsauksesta ja empiirisestä osiosta. Kirjallisuuskatsauksessa käytiin läpi datan laadun määrittelyä ulottuvuuksien kautta, datan laadun arviointimenetelmiä, datan hallinnointia, datan laatuun liittyviä haasteita sekä kehitysmenetelmiä. Empiirinen osuus koostui laadullisesta haastattelututkimuksesta, jonka aineistoa tarkasteltiin sisällönanalyysillä. Haastattelurungon pohjana hyödynnettiin AIMQ-arviointimenetelmän laatu-ulottuvuuksia ja väittämiä. Aineistossa korostuvat ongelmat jaoteltiin sisällön perusteella viiteen pääteemaan, joiden sisällä tarkasteltiin myös ongelmien keskinäisiä suhteita. Havaittuja ongelmia peilattiin kirjallisuudessa aiemmin tunnistettuihin ongelmiin, ja kohdeorganisaation datan laadun kypsyyttä arvioitiin kirjallisuudessa esitettyjen mallien avulla. Lopulta organisaatiolle muotoiltiin viisi erillistä kehitystoimenpide-ehdotusta.

Haastatteluaineistosta nousi esiin viisi pääteemaa ongelmien aiheuttajana: hajautettu järjestelmäarkkitehtuuri, tietovaraston ja raportoinnin vajaakäyttö, mittarien ja valvonnan puute, datan hallinnointi ja yleiset käytännöt sekä verkonhallinnan toiminnanohjausjärjestelmä. Hajautetut järjestelmät vaikeuttavat datan saatavuutta pilkkomalla sitä useaan eri järjestelmään sekä pakottamalla siirtelemään tietoa järjestelmien välillä, jolloin tiedonsiirtokatkokset aiheuttavat ongelmia. Tietovaraston laajamittaisempi hyödyntäminen helpottaisi saatavuutta ja tiedon visualisointia sekä raportointia, mutta työntekijät eivät ole tietoisia tästä mahdollisuudesta tai tietovaraston päivitystahti on liian hidas käyttäjien tarkoituksiin. Mittarien ja valvonnan puute luo epätietoisuutta datan käyttäjien keskuudessa, sillä datan laadusta ei ole takeita ja virheitä on usein huomattu viiveellä. Datan hallinnointia varten kohdeorganisaatiossa on käytössä datanhallintamalli, jonka jalkautus on kuitenkin vielä kesken ja nimetyt tietovastaavat eivät ota proaktiivisesti vastuuta datan laadusta. Verkonhallinnan tietojärjestelmä on käyttäjien mielestä sekava ja hidas, ja lisäksi sen sisältämissä tiedoissa on puutteita ainakin henkilö- ja laitetietojen osalta.

Datan laadun kehittämistä voi lähestyä joko proaktiivisesti tai reaktiivisesti. Proaktiivisessa strategiassa virheiden syntyminen pyritään ehkäisemään ennalta havaitsemalla ja poistamalla ongelmien juurisyitä. Mikäli tämä ei ole mahdollista, voidaan turvautua reaktiiviseen strategiaan eli virheiden korjaamiseen jälkikäteen. Fingridin käyttötoiminnan tapauksessa liiketoimintakriittinen tieto on valtaosin automaattisten mittausten tuottamaa sekä poikkeuksellisen nopeasti uusiutuvaa, joten perinteiset korjaus- ja kehitysmenetelmät eivät ole aina tarkoituksenmukaisia. Kehitystoimenpiteiksi kohdeorganisaatiolle ehdotetaan aktiivisempaa datan laadun valvontaa, datan keskittämistä, tietovirtojen dokumentointia, datanhallintamallin jalkautusta sekä verkko-omaisuusdatan puutteiden korjaamista.

Avainsanat: datan laatu, datan hallinnointi, ydintieto

ABSTRACT

Teemu Kontro: Managing Data Quality in Power System Operations
Master's thesis
Tampere University
Master's Programme in Information and Knowledge Management
September 2021

The amount of data has grown steadily due to increasing number of data sources and the development of storage solutions. At the same time, data quality has become increasingly important in business, as erroneous data can incur significant costs, damage an organization's reputation, or hinder strategy implementation. In addition, employees now spend a significant portion of their time correcting various anomalies. Automation of information system processes requires high quality data to function, as people are no longer monitoring all phases of the process.

The purpose of this study is to get acquainted with the quality problems of Fingrid Oyj's power system operations and status data and to offer development proposals based on existing literature. The work consists of a literature review and an empirical section. The literature review covered the definition of data quality through dimensions, data quality assessment methods, data governance, data quality challenges and improvement methods. The empirical part consisted of a qualitative interview study based on the survey included in the AIMQ assessment methodology. The interview material was examined using content analysis. The problems highlighted in the material were divided into five main themes based on the content, within which the interrelationships of the problems were also examined. The observed problems were mirrored to the problems previously identified in the literature, and the maturity of the data quality of the target organization was assessed using models presented in the literature. In the end, five separate proposals for development measures were formulated for the organization.

Five main root cause themes emerged from the interview material: fragmented system architecture, underutilization of data warehousing and reporting, lack of metrics and controls, data governance and common practices, and the grid control ERP system. Fragmented systems make it difficult to access data by splitting it into several different systems and forcing data migration between systems, which increases the risk of data transmission failures. Better utilization of the data warehouse would provide better access to data and facilitate data visualization and reporting, but either the employees are not aware of this possibility, or the refresh cycle of the data warehouse is too slow for their needs. The lack of metrics and controls creates uncertainty among data users, as there are no guarantees about data quality and error detection is often delayed. The target organization has a data governance model in place, but it still in the implementation phase and the designated data stewards do not proactively take responsibility for the quality of the data. The network management information system is perceived by users as confusing and slow, and in addition there are flaws in the device and personnel data it contains.

Data quality development can be approached either proactively or reactively. The proactive strategy seeks to prevent errors from occurring by identifying and eliminating the root causes of the data quality problems. If this is not possible, a reactive strategy can be used, i.e., correcting errors afterwards. In the case of Fingrid's power system operations, business-critical information regenerates in a rapid rate in addition to being largely generated by automatic measurements, hence traditional repair and improvement methods are not always appropriate. The proposed development measures for the target organization include more active data quality control, data centralization, documentation of data flows, practical implementation of their data management model and correction of deficiencies in grid asset data.

Keywords: data quality, data governance, master data

ALKUSANAT

Akateemisia opinnäytetöitä kuvaillaan usein pelottavina mörköinä, jotka seisovat loppuvastuksena opiskelijan ja valmistumisen välissä. Omalla kohdallani diplomityön teko oli lopulta hyvin kaukana tästä mielikuvasta, kiitos mielenkiintoisen aiheen, ymmärtäväisen toimeksiantajan ja laadukkaan ohjauksen. Työtä tehdessä opin valtavasti paitsi datan hallinnasta, myös laadullisen tutkimuksen periaatteista. Erityiskiitos professori Samuli Pekkolalle, joka tarjosi apua ja palautetta ihailtavan nopeasti aina tarvittaessa, sekä Fingridin asiantuntija Mika Laatikaiselle, jonka väsymättömästi sparraili työn suuntaa ja karsi työstä ainakin kolminumeroisen määrän turhia täytesanoja. Kiitokset myös toiselle tarkastajalle Pasi Raatikaiselle sekä Fingridin diplomityön ohjausryhmälle (Mika Laatikainen, Jonne Jäppinen, Markus Virtanen, Mika Latvala) tuesta ja kommenteista.

Tämä diplomityö jää myös tietojohtamisen DI-tutkintoni viimeiseksi opintosuoritteeksi, joten se edustaa samalla ennakoitua huomattavasti pidemmäksi muodostuneen yliopisto-opiskeluni päätöstä tällä erää. Matkalla on tarttunut mukaan huomattava määrä uutta tietoa opintojen kautta, mutta olennaisimmat opit ovat ehkä kuitenkin peräisin luentosalien ulkopuolelta. Erityisesti lukuisat luottamustehtävät opiskelijayhteisössä opettivat paljon erilaisten ihmisten kohtaamisesta ja vastuun kantamisesta. Kiitos näistä kokemuksista TTYY(H18), NMKSV (erityisesti Kivetkin kirjoittaa -vertaisyhteisö), TREY, TEK, Skilta ja kaikki tällä matkalla mukaan tarttuneet ystävät. Erityiskiitos vielä Paulalle, joka oli aina tukena aina silloin, kun sitä eniten tarvitsin.

Tulevaisuus diplomityön jälkeen on vielä tätä kirjoittaessa auki. Ylimääräisen murehtimisen sijaan haluan lainata muutaman sanan eräältä 2000-luvun tunnetuimmalta akateemikolta:

”Taukki! Vollotus! Kummallisuus! Nipistys!”

Tampereella, 28.9.2021

Teemu Kontro

SISÄLLYSLUETTELO

Tampereen yliopisto

KUVALUETTELO

TAULUKKOLUETTELO

1. JOHDANTO

Monet organisaatiot ovat nykyään riippuvaisia datasta, sillä se mahdollistaa sekä operatiivisen toiminnan että liiketoiminnan kehityksen analytiikan kautta (Loshin 2011). Virheellinen data voi aiheuttaa miljoonien kustannukset, heikentää mainetta asiakkaiden keskuudessa tai hankaloittaa strategian jalkauttamista (Redman 1995). Tietotyöläiset käyttävätkin nykyään merkittävästi aikaa datan etsimiseen, virheiden tunnistamiseen ja korjaamiseen sekä epämääräisistä lähteistä saadun datan varmistamiseen (Redman 2013). Jotta ylimääräisiltä kustannuksilta voidaan välttyä, organisaatioiden tulisi määrittää prosesseja datan laadun arviointiin, seurantaan ja kontrollointiin (Loshin 2011). Tämä ei kuitenkaan ole helppoa, sillä samalla datalähteiden kasvava määrä ja monimutkaisuus hankaloittavat datan laadun hallintaa (Batini et al. 2009).

Datan laadun merkitys on suuri myös esimerkiksi automaatiohankkeiden mahdollistajana. Kantaverkkoyhtiö Fingrid Oyj:n yhtenä tehtävänä on pitää sähkön tuotanto ja kulutus tasapainossa. Sääriippuvaisen energiatuotannon kasvun johdosta sähköjärjestelmä on siirtymässä tunnin syklistä 15 minuuttiin, ja toiminnan nopeuttaminen pakottaa automatisoimaan prosesseja. (Määttänen 2020) Muutoksen myötä datan laadun merkitys korostuu, kun ihmiset eivät enää ehdi tarkastamaan ja korjaamaan tietoja. Virheellinen data voi siis vääristää automaattisten laskentojen tuloksia ja johtaa esimerkiksi ylimääräisiin ostoihin reservisähkömarkkinoilta, mikä aiheuttaa ylimääräisiä kustannuksia.

Datan laatua käsittelevä tutkimus on keskittynyt erityisesti verkkosivujen datan laatuun, datan laatuun päätöksenteon tuen näkökulmasta sekä datan laadun arviointiin (Xiao et al. 2014). Datan laadun arviointiin ja kehittämiseen on kehitetty useita teoreettisia menetelmiä, mutta niitä ei ole validoitu laajamittaisesti käytännössä (Batini et al. 2009). Lisäksi tiettyjen organisaatioiden tai toimialojen kohtaamista datan laatuun liittyvistä ongelmista on toteutettu useita tapaustutkimuksia (katso esimerkiksi Haug et al. 2013, Silvola et al. 2011, Umar et al. 1999), jotka tarjoavat tietoa yritysten kohtaamista datan laatu- ja hallintaongelmista. Tämä työ täydentää alan tutkimusta tarjoamalla yhden tapauksen Lee et al. (2002) kehittämän AIMQ-arviointimenetelmän soveltamisesta sekä erityisesti luomalla uutta tietoa kantaverkkoyhtiön käyttötoiminnassa käytännössä havaituista datan laatuongelmista.

1.1 Tutkimusongelma ja -kysymykset

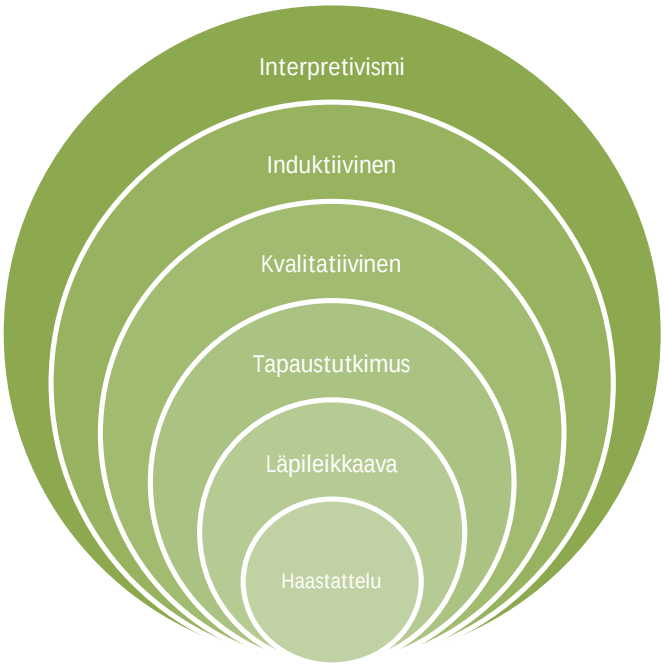
Tämän työn tarkoituksena on arvioida Fingridin kantaverkon käyttötoiminnassa hyödynnettävän datan laatuun liittyviä haasteita ja esittää niiden pohjalta kehitystoimenpiteitä. Tutkimuskysymyksiä on kaksi:

1. Mitä ongelmia käyttötoiminnan datan laadussa on tällä hetkellä?
2. Miten käyttötoiminnan datan laatua voidaan kehittää?

Tutkimuskysymyksiin pyritään vastaamaan sekä empiirisessä osiossa että kirjallisuuskatsauksessa. Teoriaosion luvut 2.4 ja 2.5 esittelevät erilaisia kirjallisuudessa havaittuja datan laadun ongelmia ja keinoja niiden välttämiseen sekä korjaamiseen. Empiirisessä osiossa kartoitetaan kohdeorganisaation ongelmia haastattelemalla datan käyttäjiä ja vastataan näin tutkimuskysymykseen 1. Tarkemmat rajaukset on esitelty luvussa 3. Tutkimuskysymykseen 2 vastataan peilaamalla haastattelusta saatuja tuloksia aiemmassa kirjallisuudessa tunnistettuihin ongelmiin sekä niihin kehitettyihin toimenpide-ehdotuksiin, ja pyritään näin luomaan myös Fingridin käyttötoimintaan sopivia kehitysehdotuksia.

1.2 Tutkimusmetodologia

Tutkimuksen metodologiset valinnat tieteenfilosofisesta koulukunnasta, lähestymistavasta, tutkimusstrategiasta, menetelmän tyypistä, aikahorisontista ja aineiston keräämisestä on esitelty kuvassa 1. Nämä valinnat ohjaavat tutkimuksen etenemistä ja kuvaavat tutkijan omia oletuksia.



Kuva 1. *Metodologiset valinnat (mukaillen Saunders et al. 2019 s. 130)*

Tutkimuksen tieteenfilosofisena pohjana toimii interpretivismi, joka painottaa tutkittavan maailman subjektiivisuutta. Sen voidaan katsoa soveltuvan hyvin liiketoiminnan tutkimukseen, sillä liiketoiminnan tapaukset ovat usein hyvin monimutkaisia ja ainutlaatuisia. Näissä tutkimuksissa kerätään usein laadullista aineistoa pienellä otannalla. (Saunders et al. 2019 s. 148–149) Myös tämä tutkimus tarkastelee hyvin yksilöllistä ja monimutkaista kokonaisuutta laadullisen aineiston avulla. Lisäksi datan laadun tarkastelu on luonnostaan hyvin subjektiivista ja yksilöiden omiin kokemuksiin perustuvaa (katso esimerkiksi Wang & Strong 1996, määritelmät esitelty tarkemmin luvussa 2.1.1), joten interpretivistinen lähestymistapa sopii tutkimukseen hyvin.

Tutkimus on lähestymistavaltaan induktiivinen eli aineistolähtöinen, sillä tutkimuksen tulokset muodostetaan aineiston pohjalta (Juhila 2021a). Tutkimusstrategian muodostamisessa ja erityisesti aineiston keräämisessä hyödynnetään kuitenkin datan laadun substanssiteorioita sekä sen arviointiin kehitettyjä teoreettisia viitekehyksiä. Tutkimuksessa kerättyä aineistoa myös vertaillaan kirjallisuudesta saatuihin havaintoihin. Työn empiirinen osuus on toteutettu tilaajaorganisaatioon kohdistettuna yhden tapauksen tapaustutkimuksena, jossa kerättiin laadullista aineistoa puolistrukturoiduilla haastatteluilla. Tapaustutkimuksen toteutus ja kohdeorganisaatio on kuvattu tarkemmin luvussa 3.

1.3 Tutkimuksen rakenne

Johdannon jälkeisessä toisessa luvussa esitellään työn kannalta olennainen teoria, joka on muodostettu kirjallisuuskatsauksella. Ensimmäisissä alaluvuissa esitellään datan laadun määrittelyä eri ulottuvuuksien kautta sekä datan laadun arviointimenetelmiä. Kolmannessa alaluvussa esitellään datan hallinnoinnin (engl. data governance) sekä ydintiedon hallinnan teoriaa. Lopulta alaluvussa 2.4 käydään läpi aiemmissa tutkimuksissa havaittuja datan laatuongelmia ja alaluvussa 2.5 datan laadun kehittämismenetelmiä. Kirjallisuuskatsaus on muodostettu hakemalla aineistoa Web of Science-, Emerald-, sekä Tampereen yliopiston Andor-tietokannoista. Hakulausekkeita olivat *”data quality assessment”*, *”data quality improvement”* sekä *”master data” AND “data quality”*. Lisäksi aineistoa haettiin niin kutsutulla lumipallomenetelmällä, eli artikkeleita haettiin myös tietokantahaulla löydettyjen tutkimusten lähteistä.

Luvussa 3 esitellään tapaustutkimuksen toteuttaminen eli kuvaillaan kohdeorganisaatio sekä kerrotaan haastatteluaineiston keräämisessä ja analysoinnissa tehdyt metodologiset valinnat perusteluineen. Luvussa 4 esitellään tapaustutkimuksen tulokset eli kohdeorganisaation datan laadun nykytilan ongelmat teemoittain. Tulosten esittelyssä hyödynnetään suoria haastattelusitaatteja sekä kaavioita, joissa esitellään aineistosta nousseі-

den ongelmien syy-seuraussuhteita. Seuraavaksi luvussa 5 pohditaan tarkemmin haastatteluaineistosta nousseita tuloksia ja peilataan niitä kirjallisuudesta löydettyihin havaintoihin muodostaen näin toimenpide-ehdotuksia kohdeorganisaatiolle. Lopulta luvussa 6 vedetään yhteen tutkimuksen tulokset sekä pohditaan tutkimuksen merkitystä ja rajoitteita sekä tarpeita jatkotutkimukselle.

2. DATAN LAADUN ARVIOINTI JA KEHITTÄMINEN

Perinteisesti tietojohtamisen alalla tiedon käsite jaotellaan jalostusasteen mukaan kolmeen osaan: dataan, informaatioon ja tietämykseen. Luokittelun mukaan data on rakenteetonta tietoa, joka voidaan jalostaa analyyseissä hyödynnettäväksi informaatioksi. Tietämys on vielä pidemmälle vietyä inhimillistä tietoa, joka usein perustuu kokemukseen. (Laihonen et al. 2013 s. 18) Datan laatua käsittelevissä tutkimuksissa datasta ja informaatiosta puhutaan kuitenkin usein ristiin, ja niillä tarkoitetaan samaa asiaa (katso esimerkiksi Wang 1998; Strong et al. 1997). Tästä syystä myös tässä työssä ”data” toimii yleiskäsitteenä, joka kattaa sekä datan että jalostetumman informaation.

Dataa voi olla esimerkiksi yksinkertainen asiakastieto, kuten osoite. Heikko laatu voi ilmetä tässä datassa monella tavalla: osoitteesta voi esimerkiksi puuttua kokonaan talon numero, tai kadun nimessä voi olla kirjoitusvirhe. Jälkimmäisessä tapauksessa data voi silti olla käytön näkökulmasta riittävän laadukasta, jos virhe on niin pieni, että osoitteeseen laitettu posti tulee silti perille. Kokonaiskuvassa virheet eivät ole aina ilmiselviä: esimerkiksi postinumero ja osoite voivat olla näennäisesti oikein, vaikka todellisuudessa osoite ei sijaitsisi mainitulla postinumeroalueella.

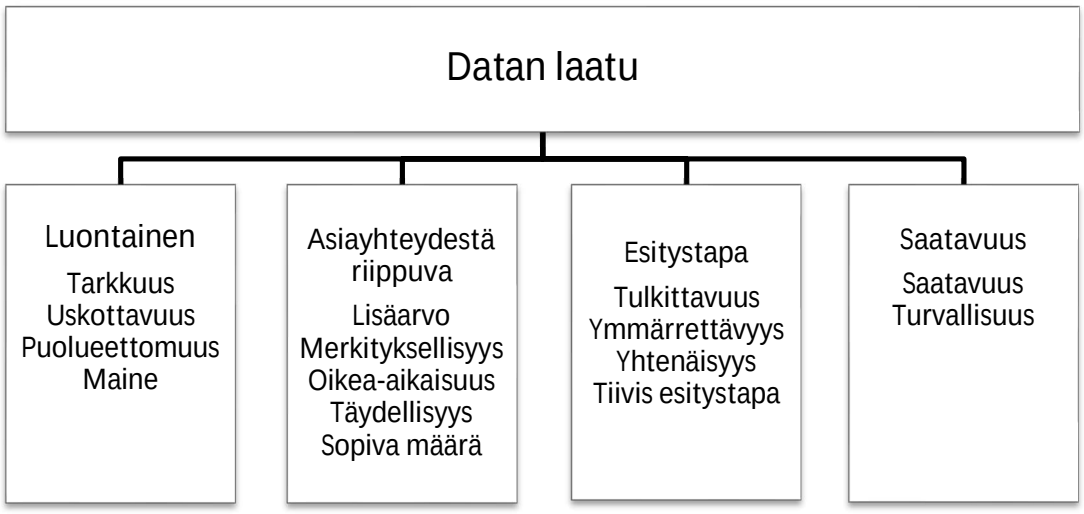
2.1 Datan laadun määritelmä ja ulottuvuudet

Datan laadulle ei ole tarkkaa yleisesti hyväksyttyä määritelmää tutkimuskirjallisuudessa. Vallitsevan käsityksen mukaan data on hyvälaatuista silloin, kun se on käyttöön sopivaa (engl. fitness for use) (Sebastian-Coleman 2013, s. 39–40; Woodall et al. 2013, Strong et al. 1997, Wang & Strong 1996). Näin ollen yhden tehtävän näkökulmasta hyvälaatuinen data voi olla kelvotonta toisen käyttötarkoituksen näkökulmasta (Tayi & Ballou 1998). Tämä voi hankaloittaa laadun arviointia, sillä eri käyttäjäryhmillä voi olla hyvin erilaiset vaatimukset datalle. Toisaalta yksittäisiä datan laadun ulottuvuuksia voidaan myös tarkastella objektiivisemmin laskemalla yksittäisen datajoukon sisältämien poikkeamien määrää. (Ballou & Pazer 1985) Poikkeamia voi laskea esimerkiksi vertaamalla datajoukkoa historiatietoon tai viitteellisiä arvoja sisältävään taulukkoon (Sebastian-Coleman 2013, s. 48). Tällainen menetelmä ei kuitenkaan ota huomioon käyttäjän tarpeita (Wang & Strong 1996). Loshin (2011, s. 130) kuitenkin huomauttaa, että pelkästään käyttäjien anekdootteihin ja esimerkkeihin perustuva arviointi hankaloittaa datan laadun tarkkaa määrittämistä ja mittaamista.

Tarkastelun helpottamiseksi datan laatu voidaan pilkkoa useisiin ulottuvuuksiin. Ulottuvuudella tarkoitetaan joukkoa datan laadun ominaisuuksia, jotka esittävät yhtä näkökulmaa laatuun (Wang & Strong 1996). Kirjallisuus ei tunnista yhtä vakiintunutta listaa datan laadun ulottuvuuksista tai eri ulottuvuuksien määritelmistä (Batini et al. 2009; Lee et al. 2002; Wand & Wang 1996). Sen sijaan eri tutkijat ovat muodostaneet omia näkökulmiaan intuition, aiemman kirjallisuuden ja empiiristen tutkimusten pohjalta (Wand & Wang 1996).

2.1.1 Ulottuvuuksien luokittelu

Ulottuvuuksia voidaan tarkastella erilaisten yläkategorioiden kautta. Wang & Strong (1996) jakavat datan laadun ulottuvuudet neljään luokkaan (kuva 2): luontainen (engl. intrinsic), asiayhteydestä riippuva (engl. contextual), esitystapa (engl. representative) sekä saatavuus (engl. accessibility). Mallissa luontainen datan laatu kuvaa datan itsessään sisältämiä ominaisuuksia eli tarkkuutta, riippumattomuutta, mainetta ja uskottavuutta. Haug et al. (2009) argumentoivat mallia vastaan huomauttamalla, että maine ja uskottavuus eivät kuitenkaan ole datalle luontaisia ominaisuuksia, vaan käyttäjän subjektiivisia kokemuksia. Wand & Wang (1996) tunnistavat ontologiaan pohjautuvassa mallissaan neljä luontaisen datan laadun ulottuvuutta, joiden mukaan datan tulisi olla täydellistä, yksiselitteistä, merkitsevää ja oikein. Maine ja uskottavuus jäävät siis tämän mallin ulkopuolelle.



Kuva 2. *Datan laatu-ulottuvuuksien luokittelu (Wang & Strong 1996)*

Muissa Wang & Strongin (1996) luokissa asiayhteydestä riippuva laatu nostaa esiin datan käyttötarkoituksen ja ympäristön asettamia vaatimuksia: datan täytyy olla merkityksellistä, saatavilla ajoissa, sisältää tarvittavat tiedot ja sitä täytyy olla sopiva määrä, jotta

siitä voi syntyä lisäarvoa. Kaksi viimeistä kategoriaa eli esitystapa ja saatavuus painottavat tietojärjestelmien merkitystä: datan täytyy olla saatavilla käyttäjille selkeässä muodossa. Loshin (2011, s. 131–134) jakaa käytännönläheistä mittaamista painottavassa mallissaan datan laadun vain luontaisiin ja asiayhteydestä riippuvaisiin ulottuvuuksiin.

Myös Yoon et al. (2000) jakavat datan laadun neljään osa-alueeseen: datan esitystavan laatu, datan arvojen laatu, datamallin laatu ja data-arkkitehtuurin laatu. Kuten Wang & Strongin (1996) mallissa, myös tässä esitystavan laatu kuvaa datan esittämistä käyttäjälle. Muut mallin luokat ovat kauempana käyttäjästä: arvojen laatu viittaa siihen, miten dataa on ylläpidetty tietojärjestelmässä, kun taas mallin ja arkkitehtuurin laatu kuvaavat datan rakenteen ja koko organisaation laajuisen datanhallinnan laatua. Haug et al. (2009) puolestaan hyödyntävät toiminnanohjausjärjestelmien datan laadun luokittelussa kolmea kategoriaa: Wand & Wangin (1996) mukaiset luontaiset laatu-ulottuvuudet, saavutettavuuden ulottuvuudet (esimerkiksi käyttöoikeudet ja datan varastointi) sekä hyödyllisyysulottuvuudet (esimerkiksi merkityksellisyys ja lisäarvo). Levitin & Redman (1995) tarjoavat erilaisen näkökulman tarkastelemalla datan arvojen sijaan sen mallin kykyä kuvata todellisuutta kuuden eri kategorian kautta, jotka ovat sisältö, laajuus, yksityiskohtaisuus, koostumus, yhtenäisyys ja muutokseen reagointi. Nämä kategoriat on jaettu edelleen yhteensä 14 eri ulottuvuuteen. Myös tässä mallissa olennaista on datan soveltuminen käyttäjien tarpeisiin.

Laadun ulottuvuuksien määrästä ja niiden keskinäisestä tärkeydestä on esitetty useita poikkeavia näkemyksiä. Ballou & Pazer (1985) kehittämä laskennallinen malli sisältää neljä objektiivisesti tarkasteltavissa olevaa ulottuvuutta: tarkkuus, oikea-aikaisuus, täydellisyys ja johdonmukaisuus. Wang & Strong (1996) ja Lee et al. (2002) täydentävät listaa lukuisilla käyttäjän tarpeita korostavilla ulottuvuuksilla, kuten maineella ja ymmärrettävyydellä. Kirjallisuudesta on kuitenkin erotettavissa neljä tärkeintä ulottuvuutta: tarkkuus, täydellisyys, yhtenäisyys sekä oikea-aikaisuus (Silvola et al. 2016; Batini et al. 2009; Lee et al. 2002). Sebastian-Colemanin (2013, s. 63–64) malli korvaa tarkkuuden oikeellisuudella, jossa datan arvoa verrataan todellisen objektin sijaan ennalta määrättyyn korvikkeeseen mittaamisen mahdollistamiseksi. Lisäksi mallissa on mukana eheys, joka tarkastelee datan pysymistä määrätyn mallin mukaisena huomioiden koko datajoukon sisäisen täydellisyyden ja yhteneväisyyden. Kirjallisuudesta löytyneitä datan laadun ulottuvuuksia on esitelty tarkemmin taulukossa 1.

Suoraan dataa tarkastelevien ulottuvuuksien ohella tarkastelun kohteena voi olla sen määrittelystä kertova dokumentaatio tai datan hallinta. McGilvray (2008) nostaa tarkkuuden, oikea-aikaisuuden, yhtenäisyyden ja muiden vastaavien rinnalle muun muassa da-

tan määrittelytiedot (engl. data specifications), jotka kuvaavat datan malleista, liiketoimintasäännöistä ja muista vastaavista määrittelyistä kertovan dokumentaation laatua. Yoon et al. (2000) puolestaan huomauttavat, että kirjallisuus ei huomio organisaation laajempaa data-arkkitehtuuria laadun ulottuvuuslistauksissa. Tämän pohjalta he ehdottavat datan laadun ulottuvuuksiksi yhdeksää lisäominaisuutta, jotka nostavat esiin muun muassa datan hallinnan, data-arkkitehtuurin hyödyntämisen ja sen kehittämisen näkökulmia. Myöhemmässä kirjallisuudessa tämä näkökulma ei kuitenkaan saa enempää tukea, vaan kokonaisvaltaisempi datan hallinta on pidetty erillään datan laadun ulottuvuuksista. Myös tässä työssä datan hallintaa ja hallinnointia käsitellään erillään luvussa 2.4.

Taulukko 1. *Datan laadun ulottuvuudet kirjallisuudessa*

Ulottuvuus	Ulottuvuuden kuvaus	Ballou & Pazer 1985	Wang & Strong 1996	Sebastian-Coleman 2013	Lee et al. 2002	Batini et al. 2009
Tarkkuus	Datan arvot vastaavat todellisuutta	X	X		X	X
Uskottavuus	Dataa voidaan pitää tarkkana		X		X	
Puolueettomuus	Data on riippumatonta		X		X	
Maine	Dataan luotetaan		X		X	
Lisäarvo	Datan käytöstä saa lisähyötyä		X			
Merkityksellisyys	Data on käyttöön soveltuvaa		X		X	
Oikea-aikaisuus	Data on ajantasaista ja käytettävissä haluttuna hetkenä	X	X	X	X	X
Täydellisyys	Datan kaikki tarpeelliset arvot ovat mukana	X	X	X	X	X
Sopiva määrä	Dataa ei ole liikaa eikä liian vähän		X		X	
Tulkittavuus	Dataa on helppo tulkita		X		X	
Ymmärrettävyys	Dataa on selkeää ja helposti sisäistettävissä		X		X	
Tiivis esitystapa	Data esitetään sopivan tiiviissä muodossa		X		X	
Saavutettavuus	Data on saatavilla aina tarvittaessa		X		X	
Turvallisuus	Dataan on pääsy oikeilla tahoilla		X		X	
Oikeellisuus	Data on ennalta määrätyn standardin mukaista			X		
Yhtenäisyys	Data pysyy muuttumattomana	X	X	X	X	X
Eheys	Data noudattaa datamallin sääntöjä muodostaen eheän kokonaisuuden			X		
Helppokäyttöisyys	Käyttäjän on helppo hyödyntää dataa				X	

2.1.2 Ulottuvuuksien määritelmät

Taulukossa 1 korostuu aiempien kirjallisuuskatsauksien mukaisesti erityisesti tarkkuus, oikea-aikaisuus, täydellisyys ja yhtenäisyys. Samoista nimistä huolimatta niiden sisältö ja painotukset voivat kuitenkin vaihdella (Batini et al. 2009). Esimerkiksi täydellisyyden määritelmä ja arvo voi vaihdella riippuen siitä, halutaanko datan olevan kauttaaltaan täydellistä (Wand & Wang 1996), vai tuleeko sen sisältää vain käyttäjän tarvitsemat arvot (Lee et al. 2002) tai sen prosessointiin tarvittavat arvot (Sebastian-Coleman 2013, s. 62). Näin ollen sen voi myös kategorisoida eri tavoin joko luontaiseksi (täydelliset tiedot) tai asiayhteydestä riippuvaksi (käyttäjän/prosessin tarvitsemat tiedot) laatu-ulottuvuudeksi (Lee et. al 2002).

Tarkkuus eli datan arvojen suhde todellisiin arvoihin on yleinen ja suoraviivainen datan laadun ulottuvuus. Wang & Strong (1996) mukaan tarkka data on oikein, luotettavaa ja vahvistettu virheettömäksi. Wand & Wang (1996) määrittelevät epätarkan arvon edustavan eri reaalimaailman tilaa kuin oli tarkoitus. Ballou & Pazer (1985) määrittelevät datan olevan tarkkaa, jos sen arvot vastaavat todellisia arvoja. Näin määriteltynä tarkkuuden mittaaminen voi kuitenkin olla haastavaa, sillä se vaatii vertailukohteen reaalimaailmasta. Esimerkiksi asiakkaan ilmoittamaa postinumeroa voidaan verrata postinumeroluetteloon ja todeta sen löytyvän luettelosta, mutta tämä ei vielä kerro, asuuko tämä asukas todellisuudessa juuri sillä alueella. (Sebastian-Coleman 2013, s. 63–64) Erilaisia datatyyppejä ei välttämättä pysty vertaamaan suoraan mihinkään ennalta määrättyyn arvoon tai kokonaisuuteen, mikä hankaloittaa tarkkuuden määrittämistä entisestään.

Oikea-aikaisuus esiintyy jossain muodossa kaikissa löydetyissä laadun ulottuvuuksien luokitteluissa, mutta osa tutkimuksista käyttää myös muita aikaan sidottuja ulottuvuuksia (Batini et al. 2009). Myös ajallisten käsitteiden määrittelyissä esiintyy vaihtelua: Ballou & Pazer (1985) määrittelevät oikea-aikaisuuden datan ajantasaisuuden kautta. Samaan tapaan Wang & Strong (1996) mukaan oikea-aikaisuus viittaa datan iän sopivuuteen valitussa tehtävässä, kun taas Sebastian-Coleman (2013, s. 62) nostaa esiin saatavuuteen liittyvän ajan: datan tulee olla käytettävissä käyttäjän tarvitsemalla hetkellä. Loshin (2011, s. 140–142) käyttää kahta eri ajallista ulottuvuutta: ajantasaisuus (engl. currency) tarkastelee datan ikää, kun taas oikea-aikaisuus (engl. timeliness) mittaa aikaa, joka käyttäjältä kuluu tiedon saamiseen sitä tarvittaessa. Lisäksi oikea-aikaisuuden alakäsitteenä esiintyy epävakaisuus (engl. volatility), joka viittaa datan muuttumiseen ajan kuluessa (Sebastian-Coleman 2013, s. 62; Wand & Wang 1996).

Myös yhtenäisyys voidaan määritellä useista eri näkökulmista. Yhtenäisyys voi tarkoittaa datan johdonmukaista esitystapaa eli datan näkymistä käyttäjälle samanlaisena esimerkiksi eri järjestelmien välillä (Loshin 2011, s. 139; Wang & Strong 1996; Ballou & Pazer

1985). Sebastian-Colemanin (2013, s. 63) mallissa yhtenäisyyttä mitataan vertaamalla datajoukkoa toiseen samalla tavalla tuotettuun joukkoon. Yhtenäisyyden voi määritellä viittaavan myös datan sisäisten (esimerkiksi ”Henkilön iän tulee olla vähintään 0”) ja keskinäisten (esimerkiksi ”Elokuvan Oscar-voittovuoden tulee olla sama kuin sen julkaisuvuoden”) sääntöjen noudattamiseen (Batini et al. 2009).

Kaiken kaikkiaan kirjallisuudessa ei määritellä datan laatua tai sen ulottuvuuksia yksimielisesti. Datan laadun ulottuvuuksia voidaan luokitella eri kategorioihin esimerkiksi sen mukaan, ovatko ne sille luontaisia vai asiayhteydestä riippuvaisia ominaisuuksia. Myös itse ulottuvuuksien määrittelyssä on suuria eroja eri tutkimusten välillä, mutta tärkeimmiksi nousevat tarkkuus, täydellisyys, oikea-aikaisuus ja yhteneväisyys. Vaikka nämä neljä ulottuvuutta korostuvat kirjallisuudessa, niillä ei ole yhteisiä, yleisesti hyväksyttyjä määritelmiä.

2.2 Datan laadun arviointi

Datan laadun eri ulottuvuuksia hyödynnetään datan laadun arvioinnissa. Arviointiin on kehitetty useita viitekehyksiä, joissa arvioidaan datan laatua sekä objektiivisilla numeerisilla mittareilla että subjektiivisemmilla arviointimenetelmillä ulottuvuuksien luonteen mukaan. Subjektiivisia ulottuvuuksia, kuten mainetta ja ymmärrettävyyttä, ei voi mitata samaan tapaan kuin esimerkiksi täydellisyyttä ja tarkkuutta, vaan niiden arviointi vaatii esimerkiksi datan käyttäjien haastattelemista (Batini et al. 2009). Olennaista datan laadun arvioinnissa onkin oikeiden ulottuvuuksien ja mittareiden määrittely (Batini et al. 2009, Pipino et al. 2002). Tässä luvussa esitellään datan laadun arviointimenetelmien yleispiirteitä sekä vertaillaan kirjallisuudessa esitettyjä menetelmiä. Batini et al. (2009) huomauttavat vertailussaan, että suuri osa viitekehyksistä on teoreettisia eikä niitä ole sovellettu laajamittaisesti käytännössä. Näin ollen niiden toiminnasta erilaisissa käyttötapauksissa ja organisaatioissa ei ole tietoa, mikä on hyvä huomioida menetelmiä soveltaessa.

Batini et al. (2009) katsauksen mukaan arviointimenetelmät voidaan jakaa edelleen tiettyihin toistuviin elementteihin, jotka ovat:

1. *Data-analyysi*
   Nykytilanteesta muodostetaan kokonaiskuva tutustumalla dataan ja siihen liittyviin sääntöihin
2. *Laatuvaatimusten analysointi*
   Datan käyttäjiltä ja ylläpitäjiltä selvitetään nykytilan ongelmia ja asetetaan uudet laatutavoitteet
3. *Kriittisten alueiden tunnistaminen*
   Valitaan tärkeimmät tietokannat ja datavirrat kvantitatiivista tarkastelua varten

4. *Prosessien mallintaminen*
   Mallinnetaan datan tuotanto- ja päivitysprosessit
5. *Laadun mittaaminen*
   Valitaan havaittuihin ongelmiin liittyvät laatu-ulottuvuudet ja asetetaan niille mittarit

Eri menetelmien käyttämät tekniikat ja niiden tavoitteet vaihtelevat, eivätkä ne välttämättä sisällä kaikkia mainittuja vaiheita. Menetelmien välisten erojen hahmottamisen helpottamiseksi Batini et al. (2009) luokittelevat menetelmät niiden sisällön perusteella neljään eri kategoriaan: operatiivisiin, taloudellisiin, kokonais- ja auditointimenetelmiin. Auditointimenetelmät (katso esimerkiksi Lee et al. 2002) keskittyvät nykytilan arviointiin eivätkä tarjoa tukea toiminnan kehittämiseen, kun taas operatiiviset menetelmät (katso esimerkiksi Wang 1998) tarkastelevat sekä arviointi- että kehitystoimia teknisestä näkökulmasta. Taloudelliset menetelmät puolestaan keskittyvät datan laatuun liittyvien kustannusten arviointiin. Kokonaismenetelmät kattavat sekä teknisen että taloudellisen tarkastelun sekä arvioinnin että kehitystoimenpiteiden osalta. Tässä esiteltävät menetelmät (taulukko 2) on jaettu kahteen kategoriaan: vakiomuotoiset menetelmät on lähtökohtaisesti tarkoitettu käytettäväksi sellaisenaan, kun taas modulaarisissa menetelmissä valitaan kuhunkin datan laadun arviointiprojektiin sopivat osat.

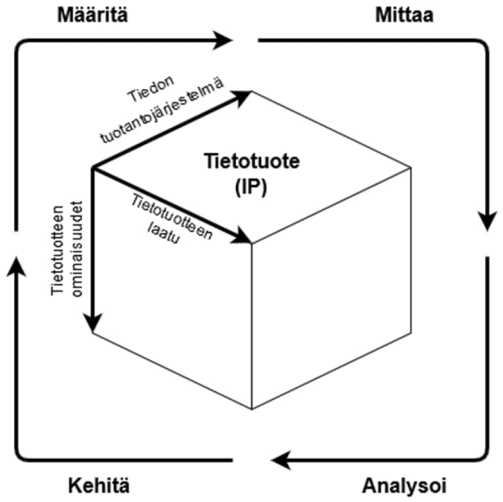
Taulukko 2. *Työssä huomioidut datan laadun arviointimenetelmät*

Menetelmä	**Lähde**	**Kategoria**
Data processing quality control model	Ballou & Pazer (1985)	Vakiomuotoinen
Total Data Quality Management (TDQM)	Wang (1998)	Vakiomuotoinen
A Methodology for Information Quality Assessment (AIMQ)	Lee et al. (2002)	Vakiomuotoinen
Data Quality Assessment Framework (DQAF)	Sebastian-Coleman (2013)	Vakiomuotoinen
Data Quality Assessment (DQA)	Pipino et al. (2002)	Vakiomuotoinen
Ten Steps Process	McGilvray (2008)	Modulaarinen
Hybrid Approach	Woodall et al. (2013)	Modulaarinen

2.2.1 Vakiomuotoiset menetelmät

Datan laadun arviointiin on kehitetty menetelmiä kymmenien vuosien ajan. Varhaisimpana menetelmänä Ballou & Pazer (1985) esittivät neljää ulottuvuutta (tarkkuus, täydellisyys, oikea-aikaisuus ja yhtenäisyys) mittaavaa mallia, jota voidaan soveltaa ainoastaan numeerisia arvoja sisältävään dataan. Tämän mallin tarkoituksena on tuottaa tietoa poikkeamien suuruusluokasta ja seurata virheiden syntymistä datavirran eri kohdissa.

Yksi varhaisimmista datan laadun viitekehyksistä on Wangin (1998) kehittämä Total Data Quality Management (TDQM), joka pohjautuu aiemmassa luvussa esiteltyihin Wang & Strongin (1996) laadun ulottuvuuksiin. Mallin periaate on esitetty kuvassa 3. TDQM-mallissa organisaation tulee ajatella informaatiota valmistusprosessin läpi kulkevana tuotteena samaan tapaan kuin perinteisessä valmistavassa teollisuudessa – fyysinen tuote valmistetaan raakamateriaalista tuotantolinjalla, ja samaan tapaan tietotuote valmistetaan raakadatasta tietojärjestelmässä. Mallin tarkoitus on toimittaa tiedon kuluttajille laadukkaita tietotuotteita.



Kuva 3. *TDQM-metodologian osat (mukaillen Wang 1998)*

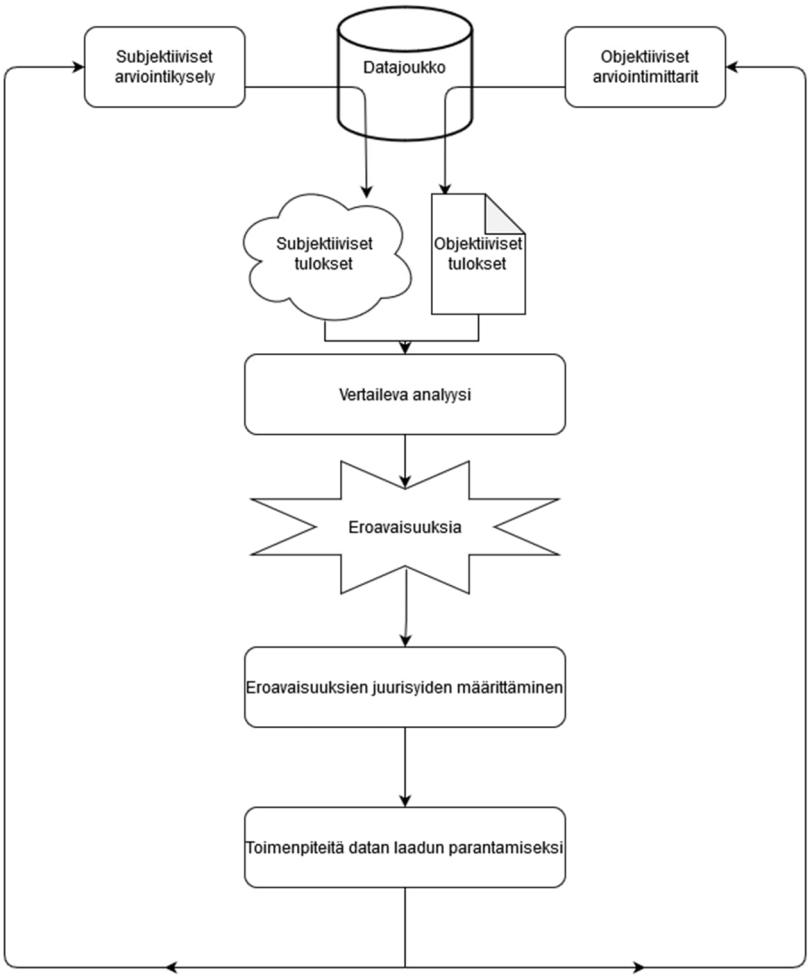
Malli koostuu iteratiivisesta prosessista, jonka vaiheet ovat määrittely, mittaaminen, analysointi ja kehittäminen. Määrittelyvaiheessa tunnistetaan tietotuotteen ominaisuudet, sen tärkeimmän laadun ulottuvuudet laatuvaatimuksineen sekä tietojärjestelmä, jossa tietotuote tuotetaan. Nämä kolme osaa muodostavat tietotuotekuution. Edelleen mittausvaiheessa tuotetaan sopivat laatumittarit, analysointivaiheessa tunnistetaan laatuun liittyvien ongelmien juurisyyt sekä lasketaan niistä aiheutuvat kustannukset ja lopulta kehittämisvaiheessa tuotetaan menetelmiä laadun kehittämiseksi soveltuvien ulottuvuuksien kautta. (Wang 1998) Näin ollen menetelmä sisältää laatuvaatimuksien analysointia lukuun ottamatta kaikki Batinin (2009) listaamat arviointivaiheen osat, vaikka menetelmä sisältää määrittelyvaiheessaan myös vaatimusten arviointia (Wang 1998). Myös Woodall et al. (2013) toteavat TDQM-menetelmän sisältävän laatuvaatimusten määrittelyn.

Lee et al. (2002) kehittämässä AIMQ-menetelmässä datan laatua arvioidaan subjektiivisesti. Menetelmä koostuu kolmesta osasta, joita voidaan hyödyntää myös itsenäisesti. Ensimmäinen osa on 2x2 -matriisi, joka kuvaa datan laadun merkitystä sen käyttäjille ja hallinnoijille. Matriisin kentät jaottelevat datan laadun ulottuvuudet neljään kategoriaan vakaaseen, luotettavaan, hyödylliseen ja käyttökelpoiseen informaatioon. Toinen osa on

kyselylomake, jonka avulla voidaan arvioida organisaation datan laatua pisteyttämällä väitteitä, jotka ovat muodostettu datan laadun ulottuvuuksien pohjalta. Kolmas osa koostuu kahdesta vaihtoehtoisesta analyysimenetelmästä, joissa verrataan kuiluanalyysillä kyselystä saatuja tuloksia joko saman organisaation eri yksiköiden tai erikseen valitun hyväksi todetun verrokkikohteen tuloksiin. Batini et al. (2009) huomauttavat, että kirjallisuudesta ei löydy tietokantaa, joka mahdollistaisi vertailun toiseen organisaatioon. Menetelmä myös erottuu muista sen subjektiivisuudella, mutta toisaalta tämä on linjassa datan laadun *fitness for use* -määritelmän kanssa. Menetelmä myös painottuu puhtaasti datan laadun arviointiin eikä tarjoa työkaluja laadun kehittämiseen.

Osa viitekehyksistä painottaa enemmän objektiivisia, numeerista sisältöä tuottavia arviointimenetelmiä. Sebastian-Colemanin (2013) Data Quality Assessment Framework (DQAF) sisältää ainoastaan objektiivisia mittareita, joilla datan laatua valvotaan automaattisesti ja jatkuvasti. Objektiviisiin mittareihin on päädytty, koska datan tulisi silti täyttää tietyt perusvaatimukset ollakseen käyttökelpoista, vaikka datan laatu määritelläänkin sen käyttäjien tarpeiden kautta. DQAF-viitekehys tarjoaa yleisen mallin jatkuvaan datan oikea-aikaisuuden, täydellisyyden, oikeellisuuden, yhtenäisyyden ja eheyden mittaamiseen. Malli sisältää kaikkiaan 48 erilaista mittaria näille ulottuvuuksille. Cappiello et al. (2004) huomauttavat, että algoritminen datan laadun mittaaminen voi sivuuttaa määritelmällisesti olennaiset käyttäjien erilaiset dataan kohdistetut vaatimukset ja esittelevät mallin, jossa automaattinen mittausprosessi voidaan räätälöidä käyttäjien vaatimusten perusteella.

Erilaisia subjektiivisia ja objektiivisia arviointimenetelmiä voidaan myös yhdistellä samassa viitekehyksessä. Pipino et al. (2002) kehittämässä Data Quality Assessment -viitekehyksessä arvioidaan datan laadun nykytilaa sekä subjektiivisilla että objektiivisilla menetelmillä, jonka jälkeen niiden tuloksia verrataan. Jos joko subjektiivisessa tai objektiivisessa tarkastelussa todetaan puutteita tai tulosten välillä on poikkeamia, prosessissa edetään ongelmien juurisyiden tutkimiseen. Juurisyiden analyysin pohjalta muodostetaan edelleen tapauskohtaisia kehitysehdotuksia. DQA-prosessi on visualisoitu kuvassa 4. Malli kehottaa organisaatioita muotoilemaan tarkoituksiinsa sopivat mittarit tapauskohtaisesti, mutta tarjoaa niiden pohjaksi kolmea eri luokkaa: haluttujen arvojen määrän osuus kaikista arvoista (engl. simple ratio), minimin tai maksimin laskeminen sekä painotettu keskiarvo.



Kuva 4. *DQA-menetelmän vaiheet (mukaillen Pipino et al. 2002)*

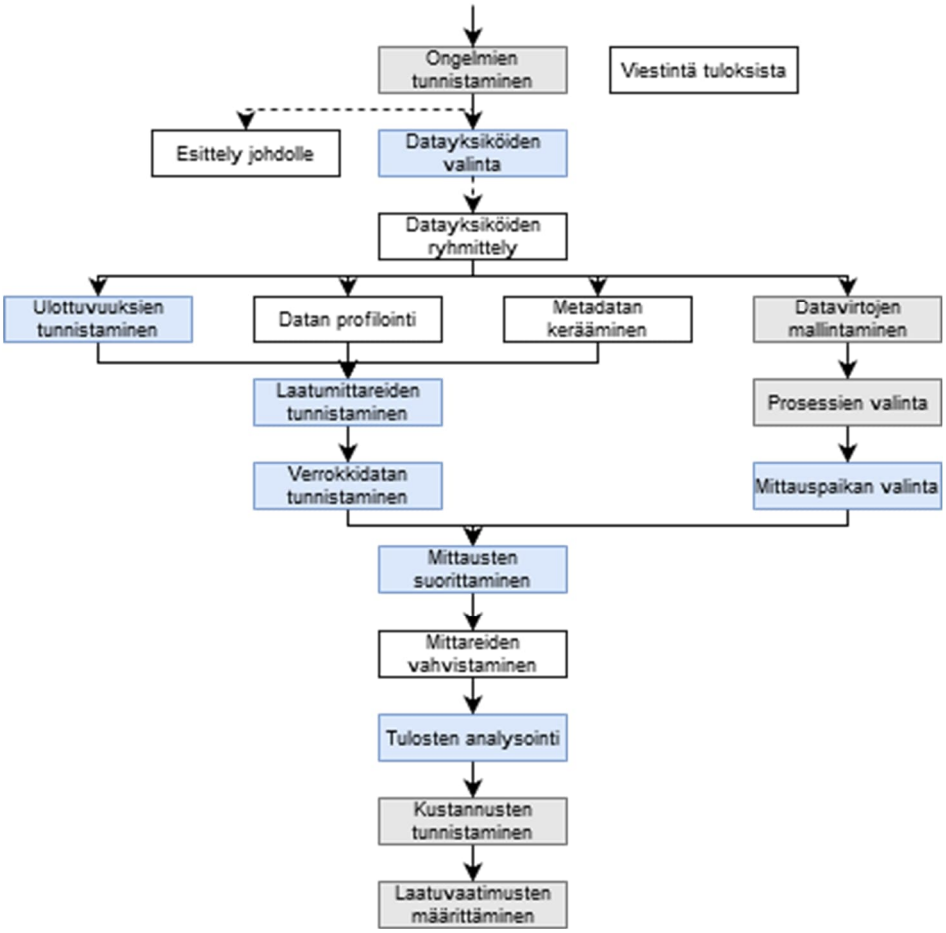
DQA-malli on vakiomuotoisista malleista vapaamuotoisin, eikä se tarjoa suoraan yhtä tiukkoja ohjeita ja konkreettisia työkaluja datan laadun arviointiin tai varsinkaan sen kehittämiseen. Tämä kuitenkin mahdollistaa mallin soveltamisen erilaisiin käyttötapauksiin, kun mittarit ja mahdolliset kehitystoimenpiteet on joka tapauksessa tarkoitus määrittää erikseen kohteen ominaispiirteitä silmällä pitäen.

2.2.2 Modulaariset menetelmät

Osassa menetelmiä myös niiden sisältämät toimenpiteet valitaan vastaamaan kunkin tapauksen yksilöllisiä tarpeita. McGilvroy (2008) esittää kymmenvaiheista iteratiivista mallia, jossa käytettävät vaiheet valitaan projektin vaatimusten mukaan. Malli pohjautuu PDCA-sykliin (Plan-Do-Check-Act), ja siinä on kolme osiota: arviointi, ymmärrys ja toiminta. Arviointiosiossa käydään läpi neljä ensimmäistä vaihetta, jotka ovat liiketoiminnan tarpeiden määrittäminen, tietoympäristön analysointi, datan laadun arviointi sekä liiketoimintavaikutusten arviointi. Tämän jälkeen ymmärrysosiossa tunnistetaan juurisyyt havaittujen ongelmien taustalla ja kehitetään niiden pohjalta suunnitelma kehitystoimenpiteistä. Lopulta toimintavaiheessa ehkäistään tulevia virheitä datassa, korjataan nykyiset

virheet ja otetaan käyttöön valvontamenetelmät. Kymmenes vaihe on viestintä toimenpiteistä ja tuloksista, ja se läpileikkaa kaikkia osioita jatkuvana toimintana.

Myös Woodall et al. (2013) painottavat aikaisempia malleja yhdistelevässä hybridilähestymistavassaan arviointimenetelmien muotoilua kunkin organisaation tarpeiden mukaisesti. Hybridilähestymistapa ei ole suoraan valmis toimintamalli, vaan se tarjoaa neljä vaihetta organisaatiokohtaisen datan laadun arviointimenetelmän laatimiseen. Ensimmäisessä vaiheessa määritellään arvioinnin tarkoitus, joka voi olla esimerkiksi aiemmin havaitun datan laatuongelman mittaaminen tai organisaation datan laadun nykytilan arviointia ja havaittujen ongelmien priorisointia. Toisessa vaiheessa tunnistetaan organisaation vaatimukset, joiden tulee olla linjassa ensimmäisen vaiheen tavoitteen kanssa. Organisaation asettamia vaatimuksia voivat olla esimerkiksi heikosta datan laadusta aiheutuvien kustannusten laskenta tai datavirtojen mallintaminen. Kolmannessa vaiheessa valitaan organisaation vaatimuksiin sopivat arviointimenetelmien toiminnot. Toiminnot on tunnistettu aiemmasta datan laadun arviointimenetelmiä käsittelevästä kirjallisuudesta ja ne on kuvattu ja luokiteltu tarkemmin kuvassa 5. Toinen ja kolmas vaihe myös tukevat toisiaan ja niitä voidaan suorittaa iteratiivisesti, sillä vaatimuksia voi olla vaikea hahmottaa ilman tietoa arviointimenetelmistä. Lopulta vaiheessa 4 toiminnot asetetaan toimivaan järjestykseen niiden erilaiset riippuvuussuhteet huomioiden.



Kuva 5. *Hybridilähestymistavassa käytettävät toiminnot (mukaillen Woodall et al. 2013)*

Kuvassa on esitetty kaikki hybridilähestymistavassa huomioidut arviointimenetelmien toiminnot. Siniset toiminnot ovat suositeltuja toimintoja, jotka löytyivät kaikista tutkimuksen arvioimista menetelmistä. Nämä vaiheet ovat datayksiköiden valinta, laatu-ulottuvuuksien tunnistaminen, laatumittareiden tunnistaminen, verrokkidatan tunnistaminen, mittauspaikan valinta, mittausten suorittaminen ja tulosten analysointi. Harmaat toiminnot ovat säädettäviä toimintoja, jotka voidaan suorittaa useammassa eri kohdassa muiden valittujen toimintojen mukaan. Valkoiset toiminnot ovat hajanaisemmin kirjallisuudesta tunnistettuja toimintoja, joita voidaan hyödyntää sopivissa tilanteissa. Katkoviiva kuvaa toiminnon riippuvuutta seuraavasta: esimerkiksi organisaation johdolle ei voida pitää esitystä ilman ongelmien tunnistamista.

2.3 Datan hallinnointi

Datan laadun onnistunut hallinta vaatii ymmärrystä erilaisista datan laatuun vaikuttavista organisatorisista toimenpiteistä, jotka ovat osa datan hallinnointia (engl. data gover-

nance). Tarkemmin sanottuna datan hallinnointi on datan kannalta relevanttien prosessien, vastuiden, ohjeistuksien ja menettelytapojen määrittämistä (Dreibelbis et al. 2008). Hallinnoinnin tavoitteena on varmistaa datan ja liiketoiminnan yhteensopivuus, mikä sisältää myös datan laatuvaatimusten täyttymisen (Brous et al. 2016). Datan laatua ei välttämättä voida pitää hyvänä, jos sitä ei tueta datan hallinnoinnin menetelmillä, kuten ohjeistuksilla ja selkeällä vastuunjaolla (Mahanti 2019 s. 401).

Kirjallisuudessa esiintyy aiheeseen liittyen useita osittain rinnakkaisia käsitteitä: datan laadun hallinnan (engl. data quality management, DQM) voidaan katsoa liittyvän olennaisesti datan hallinnoinnin ja ydintiedon hallinnan (engl. master data management) kokonaisuuksiin, joskin käsitteille ei ole tunnistettu yksiselitteisiä määritelmiä. Ydintiedon hallinnan tavoitteena on varmistaa datan laadukkuus kehittämällä organisaation prosesseja, toimintatapoja ja teknologioita (Vilminko-Heikkinen & Pekkola 2019). Laihonen et al. (2013 s. 20) mukaan suuri datamäärä voi pakottaa organisaation keskittymään vain olennaisimman datan eli ydintiedon laatuun. Joskus datan laadun hallintaa pidetään yhtenä ydintiedon hallinnan osana, mutta erityisesti datan hallinnointi ja muut ennaltaehkäisevät laadunhallintamenetelmät voidaan nähdä erillisenä ydintiedon hallintaa tukevana käsitteenä (Otto et al. 2012). Datan hallinnoinnin voidaan siis katsoa olevan datan laadun hallintaa ja ydintiedon hallintaa tukeva kattotermi, jonka yhtenä tavoitteena on varmistaa hyvälaatuinen data liiketoiminnan käyttöön. Seuraavissa aliluvuissa käydään läpi ydintiedon hallinnan teoriaa datan laadun näkökulmasta sekä esitellään kirjallisuudessa yleisesti tunnettuja datan hallinnoinnin rooleja ja vastuita.

2.3.1 Ydintiedon hallinta ja datan laatu

Organisaation ydintieto (engl. master data) kuvaa sen toiminnan olennaisimpia sisältöjä, kuten esimerkiksi asiakkaita, tuotteita, palveluita, tavarantoimittajia ja henkilöstöä (Silvola et al. 2011; Smith & McKeen 2008; Dreibelbis et al. 2008). Ydintiedolle on ominaista sen läpileikkaavuus organisaatiossa: ydintieto on oltava yhtenäistä ja käytettävissä eri yksiköiden välillä esimerkiksi asiakkaita laskuttaessa (Dreibelbis et al. 2008). Ydintiedon tietyt osat ovat usein muuttumattomia ajan suhteen: esimerkiksi tietyn materiaalin ominaisuudet pysyvät aina samana (Otto & Hüner 2009). Ideaalitilanteessa kaikki organisaation ydintieto olisi tallennettuna samaan paikkaan, jossa sitä voitaisiin hallita (Silvola et al. 2011; Dreibelbis et al. 2008). Tällöin organisaation prosesseja ja tietojärjestelmiä kehitettäisiin tätä silmällä pitäen. Käytännössä yksi ydintietojärjestelmä ei kuitenkaan usein ole realistinen vaihtoehto, sillä vaadittavat integraatiot voivat olla kalliita. (Silvola et al. 2011) Ydintiedon ympäristöineen tulisi olla mukautuvaa, sillä niiden pitäisi pystyä kehittymään liiketoiminnan muuttuessa ajan myötä (Dreibelbis et al. 2008).

Ydintiedon hallinnan tavoitteena on mahdollistaa tällainen ideaalitilanne hyödyntämällä arkkitehtuuria, teknologiaa ja liiketoimintaprosesseja (Dreibelbis et al. 2008). Smith & McKeen (2008) määrittelevät ydintiedon hallinnan sovelluksista riippumattomaksi prosessiksi, joka kuvaa, omistaa ja hallinnoi ydinliiketoiminnan datakokonaisuuksia. Saman määritelmän mukaan datan ei kuitenkaan tarvitse sijaita yhdessä yhteisessä datalähteessä. Ydintiedon hallinta ja siinä käytettävät tietojärjestelmät voidaan jakaa analyyttiseen ja toiminnalliseen kategoriaan käyttötarkoituksen mukaan (Loshin 2011). Ydintiedon hallinnan projekteja johtaa usein organisaation IT-osasto, mutta niiden haasteet kohdistuvat yleensä ihmisiin ja prosesseihin (Silvola et al. 2011). Vilminko-Heikkinen & Pekkola (2013) jakavat ydintiedon hallinnan toteuttamisprosessin kymmeneen vaiheeseen:

1. tarpeen tunnistaminen
2. organisaation olennaisimman datan ja sitä käyttävien prosessien tunnistaminen
3. hallinnoinnin määrittely
4. ylläpitoprosessien määrittely
5. datastandardien määrittely
6. mittarien määrittely
7. arkkitehtuurimallin suunnittelu
8. koulutuksen ja viestinnän suunnittelu
9. ydintiedon hallinnan kehityssuunnitelman teko
10. ydintiedon hallintasovelluksen ominaisuuksien määrittely.

Onnistunut ydintiedon hallinta edellyttää siis ydintietojen tunnistamista ja huolellista määrittelyä. Ydintietoja voidaan tunnistaa joko analysoimalla liiketoimintaprosessissa hyödynnettäviä tietoja tai tarkastelemalla käytössä olevia datajoukkoja erikseen (Loshin 2009). Epäselvästi määritellyt ydintiedot voivat aiheuttaa ongelmia viestinnässä sekä datan laadussa, ja datan laatuongelmat ovat yksi suurimmista haasteista ydintiedon hallinnassa (Silvola et al. 2011).

2.3.2 Roolit ja vastuut

Oikeanlainen roolitus ja vastuunjako ovat olennainen osa datan laadun hallintaa. Erilaisia rooleja on tunnistettu kirjallisuudessa laajalti: Strong et al. (1997) tunnistavat kolme eri roolia datan tuotantojärjestelmissä: datan tuottajat (engl. data producer) ovat ihmisiä tai muita lähteitä, jotka luovat dataa; datan valvojat (engl. data custodian) järjestävät ja hallinnoivat tietoteknisiä resursseja datan prosessointiin sekä varastointiin ja datan kuluttajat (engl. data consumers) lopulta käyttävät dataa. Wang (1998) tunnistaa TDQM-mallissaan näiden kolmen luokan lisäksi tietotuotepäälliköt (engl. IP manager), joiden

vastuulla on koko tietotuotteen tuotantoprosessin hallinta tuotteen elinkaaren ajan. Sebastian-Coleman (2013 s. 19–20) tunnistaa datan tuottajien ja kuluttajien lisäksi datan välittäjät, jotka eivät suoraan tuota dataa, mutta mahdollistavat sen kuluttamisen muille käyttäjille.

Erityisen suuri merkitys datan hallinnoinnissa ja ydintiedon hallinnassa on erilaisilla määrätyillä vastuurooleilla (Haug et al. 2013, Smith & McKeen 2008), jotka eivät välttämättä ota kantaa datan tuotantoon tai hyödyntämiseen. Tällaisia rooleja ovat muun muassa tietovastaavat (engl. data steward) sekä tiedon omistajat (engl. data owner). Tietovastaavan rooli voidaan luokitella joko tekniseksi tai liiketoimintalähtöiseksi tehtävän sisältämien vastuiden perusteella (Vilminko-Heikkinen & Pekkola 2019, Weber et al. 2009). Loshinin (2011, s. 122–124) mukaan rooli ei välttämättä ole tietotekninen, eikä sen välttämättä tule olla kokoaikainen rooli. Sebastian-Coleman (2013) toteaa, että liiketoiminnan tietovastaavan ja tietoteknisemmän datan valvojan roolien erottelu ei ole hyödyllistä, sillä IT-osastolla on joka tapauksessa velvollisuus ymmärtää datan merkitystä liiketoiminnalle edes jossain määrin. Joka tapauksessa liiketoiminnan tietovastaavan rooliin kuuluu hänen oman liiketoiminta-alueensa tiedoista vastaaminen ohjeistusten mukaisesti (Vilminko-Heikkinen & Pekkola 2019). Tietovastaavalla on tietämystä datan merkityksestä liiketoiminnalle sekä siihen liittyvistä säännöistä (Smith & McKeen 2008). Vastuualueeseen kuuluu myös datan laadun standardien kehittäminen sekä laadunvalvonta (Loshin 2011, s. 123–124).

Tietovastaavan yläpuolella hierarkiaan sijoittuu datan omistaja, jonka vastuulla on tietyn tietoalueen (engl. data domain) ylläpito ja kehittäminen (Vilminko-Heikkinen & Pekkola 2019). Datan omistajan määrittäminen voi olla hankalaa sen abstraktin luonteen takia (Sebastian-Coleman 2013 s. 21), ja yritysten datan hallinnan ongelmat voivat olla osittain peräisin huonosti määritellystä omistajuudesta (Silvola et al. 2011). Datajoukkojen omistajuuksien määrittely on välttämätöntä liiketoimintayksiköiden osallistumisen varmistamiseksi (Vilminko-Heikkinen & Pekkola 2013). Datan omistajuus on siis tärkeää halki organisaation, jotta ydintiedon hallintaa voidaan tehdä menestyksekkäästi.

Omistajuuden määrittämisessä IT:n ja liiketoiminnan välillä on omat haasteensa. Liiketoiminta voi haluta omistaa datan, sillä he hyötyvät sen hallintavallasta. Toisaalta jos vastuu datan käsittelyjärjestelmistä on IT-osastolla, liiketoiminta ei välttämättä koe hallitsevansa dataa käytännössä. Samaan aikaan IT ei halua olla vastuussa datasta, jonka sisältöä he eivät hallitse, vaikka heillä on joka tapauksessa suuri rooli sen käsittelyjärjestelmien hallinnassa. (Sebastian-Coleman 2013, s. 23) Järjestelmien ja prosessien omistajuuden kautta myös datan omistajuus voidaan helposti liittää juuri IT:n vastuulle. (Vilminko-Heikkinen & Pekkola 2017). Vaikka organisaation IT tukee ydintiedon hallintaa

tietoteknisellä osaamisellaan, juuri liiketoiminnalle annettu omistajuus olisi tärkeää, sillä he myös käyttävät dataa omassa päätöksenteossaan (Smith & McKeen 2008).

Omistajuus-termiä on myös kritisoitu. Redman (2008) huomauttaa, että datan omistajuudesta puhuminen voi heikentää datan jakamista organisaation sisällä, sillä ”omistaminen” sisältää sanana tiettyjä oikeuksia, mitkä tuovat mukanaan myös valtaa. Tällöin datan hallinnointi voi johtaa sisäiseen valtakilvoitteluun ja ristiriitatilanteisiin. Sebastian-Colemanin (2013) mukaan datan omistajien nimeäminen voi kummuta halusta ratkaista monimutkaiset ongelmat yksinkertaisella tavalla, mutta käytetystä terminologiasta riippumatta selkeä organisaation sisäinen vastuunjako on yksi tehokas lähestymistapa.

Liiketoiminta-aluekohtaisten vastaavien ja omistajien lisäksi datan hallinnoinnissa voidaan hyödyntää ylempiä vastuurooleja sekä erilaisia ohjausryhmiä tai vastaavia toimielimiä. Esimerkiksi ydintiedon hallintahankkeissa voidaan nimittää johtoryhmätason konseptin omistaja, jonka vastuulla on ydintiedon hallinnan kehittäminen, sekä operatiivinen omistaja, joka on vastuussa teknisestä toteutuksesta (Vilminko-Heikkinen & Pekkola 2013). Yleisemmin datan laadun hallinnasta konseptin omistajasta voidaan käyttää myös termiä sponsori (engl. executive sponsor) (Weber et al. 2009).

Yksittäisille henkilöille nimettyjen vastuu- ja omistusroolien lisäksi datan laadun hallinnassa ja ydintiedon hallintahankkeissa voidaan hyödyntää jonkinlaista ohjausryhmää, jonka vastuulla on datan hallinnointijärjestelmän kehittäminen ja käyttöönotto. Ryhmä voi koostua esimerkiksi liiketoimintayksiköiden ja IT:n johtajista sekä tietovastaavista. (Weber et al. 2009) Hallinnointimalli voi myös sisältää useita eri ryhmiä eri tasoilla: Loshinin (2009) esittämässä ratkaisussa hallintomallin toimintaa valvoo ylimpänä datan hallinnoinnin valvontakomitea, ja lähempänä operatiivista tasoa ennen tietovastaavia toimii datan koordinaationeuvosto, joka huolehtii esimerkiksi laatumittareista sekä tietovastaavien toimien priorisoinnista. Vastuuta voi siis jakaa usealla eri tavalla, ja jokaisen organisaation tulisi suunnitella oma datan hallinnointiratkaisunsa (Weber et al. 2009).

2.4 Heikkolaatuinen data ja syyt sen taustalla

Kuten datan laadun ulottuvuudet kertovat, data voi olla puutteellista monin eri tavoin. Wand & Wang (1996) toteavat ontologisessa mallissaan datan puutteiden syntyvän, kun käyttäjän havainto reaalimaailmasta on ristiriidassa tietojärjestelmästä saadun reaalimaailmaa kuvaavan datan kanssa. Näin määriteltynä data voi muuttua virheelliseksi ajan myötä reaalimaailman tilan muuttuessa, vaikka data pysyisi muuttumattomana (Maydanchik 2007). Strong et al. (1997) määrittelevät datan laatuongelman miksi tahansa jostain laatu-ulottuvuudesta ilmaantuneeksi vaikeudeksi, joka tekee datasta osittain tai täysin

käyttökelvotonta. Tämä määritelmä sitoo ongelmat aiemmin käsiteltyihin laatu-ulottuvuuksiin ja laajalti hyväksyttyyn datan laadun määritelmään sen käyttöön sopivuudesta. Ongelmia voi olla hankala hahmottaa, sillä datan käyttö jakaantuu läpileikkaavasti organisaatiossa. Puutteet datan laadussa voivat ilmetä muissa liiketoiminnan prosesseissa, kuten esimerkiksi asiakkaiden palautteessa tai erilaisten korjaavien toimenpiteiden määrän kasvuna. (McGilvray 2008)

Ongelmia voidaan tarkastella ja luokitella monin tavoin. Strong et al. (1997) jakavat ongelmat laatu-ulottuvuuksien ja niiden yläkategorioiden (kuva 1) mukaisesti luontaisiin, asiayhteydestä riippuviin, sekä yhdistettyihin saatavuus- ja esitystapaongelmiin. Hieman samaan tapaan Redman (1996) luokittelee ongelmat reaalimaailmaa kuvaavien mallien (esimerkiksi merkityksellisyys, yksityiskohtaisuus), datan arvojen (tarkkuus, täydellisyys), datan esitystavan (tulkittavuus, esitystavan sopivuus tehtävään) tai muihin ongelmiin (luottamuksellisuus, omistajuus) (katso Redman 1998). Kaikki datan laatuongelmat eivät ole teknisiä (Umar et al. 1999), vaan laadukkaan datan esteenä ovat usein pehmeämmät organisatoriset, poliittiset ja sosiaaliset ongelmat (Redman 2004). Näin ollen myös mahdolliset hallinnolliset tekijät on hyvä huomioida datan laatua arvioidessa ja kehittäessä.

2.4.1 Ongelmien ilmeneminen datassa

Kirjallisuudessa on listattu jonkin verran erilaisissa käyttötapauksissa ilmenneitä puutteita datan laadussa. Redman (2008, s. 41–45) esittelee seitsemän tyypillistä datan laatuongelmaa: dataa ei löydetä, virheellinen data, heikko datan määrittely, datan yksityisyys/turvallisuus, epäyhteneväisyys datalähteiden välillä, liian suuri määrä dataa sekä organisatorinen epäjärjestys, kuten tietämättömyys oman datan käyttökohteista ja tärkeydestä. Valtaosa näistä ongelmista vastaa lähes suoraan joitain aiemmin esiteltyjä datan laadun ulottuvuuksia, kuten saatavuutta, tarkkuutta, turvallisuutta, yhtenäisyyttä ja sopivaa määrää.

Ge & Helfert (2007) kokoavat kirjallisuudessa havaittuja datan laatuongelmia 2x2 -matriisimalliin: sarakkeet jaottelevat ongelmat data- tai käyttäjänäkökulmaan ja rivit jaottelevat ongelmat joko asiayhteydestä riippumattomaksi tai riippuvaiseksi. Malli on esitelty taulukossa 3. Vasemman yläneliön ongelmat viittaavat tietokannassa olevaan dataan, ja niitä voi ilmaantua missä tahansa datajoukossa. Vasen alaneliö kuvaa liiketoiminnan asettamia sääntöjä rikkovia ongelmia, jotka voidaan havaita asettamalla yhteyteen sopivia sääntöjä. Oikean yläneliön ongelmat voivat syntyä dataa prosessoidessa, ja oikean alaneliön ongelmat syntyvät, kun data ei täytä käyttäjien asettamia vaatimuksia.

Taulukko 3. *Laatuongelmien luokittelu (Ge & Helfert 2007)*

	Datan näkökulma	**Käyttäjän näkökulma**
Asiayhteydestä riippumaton	Kirjoitusvirhe Puuttuvaa data Kaksoiskappale Virheellinen arvo Epäjohdonmukainen muoto Vanhentunut data Vajavainen muoto Syntaksivirhe Ainutlaatuisen arvon rikkomus Eheysrajoitteiden rikkominen Tekstin muotoilu	Tieto on saavuttamattomissa Tieto ei ole turvattua Tieto on hädin tuskin saatavilla Tietoa on hankala koota Virheet tiedon muuntamisprosessissa
Asiayhteydestä riippuvainen	Alueen rajoitteiden rikkominen Liiketoiminnan asettamien sääntöjen rikkominen Yhtiön ja julkishallinnon sääntelyn rikkominen Tietokannan ylläpitäjän asettamien rajoitteiden rikkominen	Tieto ei pohjaudu faktoihin Tieto ei ole luotettavaa Tieto on puolueellista Tieto ei ole merkityksellistä työn kannalta Tieto koostuu epäjohdonmukaisista merkityksistä Tieto on esitetty tiiviisti Tietoa on hankala käsitellä Tietoa on hankala ymmärtää

Taulukosta voidaan nähdä yleisimpien ongelmien olevan melko yksinkertaisia, kuten kirjoitusvirheistä tai muusta vastaavasta virheestä syntynyt poikkeama datassa, tai joko tietojärjestelmän tai liiketoimintaympäristön asettamien rajoitteiden rikkominen. Käyttäjälle nämä ongelmat voivat näkyä lukuisien eri laatu-ulottuvuuksien heikentymisenä.

Datan laadun ongelmia on tutkittu myös tarkemmin tämän tutkimuksen kannalta relevanteissa tapaustutkimuksissa. Karkouch et al. (2016) listaavat kuusi erilaista esineiden internetin sensoridatan laatuongelmien esiintymismuotoa: toimittamatta jääneet lukemat, epäluotettavat lukemat, ristiriidat eri datalähteiden välillä, datan kaksoiskappaleet, datan vuotaminen sekä aikapoikkeamat eri datalähteissä. Suuri osa näistä ongelmista johtuu sensorien määrästä ja monimuotoisuudesta sekä niiden mittausten yleisestä epäluotettavuudesta. Liu et al. (2020) erottelevat katsauksessaan ongelmat hieman tarkemmin mittauksen virheisiin (esim. sensori sijoitettu väärin), kohinaan, artefaktivirheisiin, datan vääristymiseen, likaiseen dataan, poikkeamiin, puuttuvaan dataan, puuttuviin päivityksiin, datan häviämiseen ja datan lähetysviiveeseen. Toisin sanoen sensoreiden ja tietoliikenteen epävarmuus korostuvat esineiden internetin tuottaman datan laatuongelmissa.

Umar et al. (1999) tutkivat tietoliikenneyhtiöissä esiintyviä datan laatupuutteita ja löysivät 80 erilaista ongelmaa, jotka yhdistettiin viideksitoista laajemmaksi ongelmaksi. Nämä

ongelmat on eritelty taulukossa 4. Alkuperäisessä tutkimuksessa ongelmat jaoteltiin kategorioittain data-, ohjelmisto- ja prosessiongelmiin, mikä kertoo datan laatuun liittyvien ongelmien monimuotoisuudesta.

Taulukko 4. *Datan laatuongelmat tietoliikenneyhtiöissä (Umar et al. 1999)*

Kategoria	**Lyhyt kuvaus**
Järjestelmien välinen epäyhtenäisyys	Dataa ei löydy kaikista tarvittavista järjestelmistä oikeassa muodossa
Prosessien kehittäminen	Prosesseja täytyy kehittää ja automatisoida
Mittareiden tarve	Datan laatuongelmien vaikutusta ja kehitystä täytyy mitata
Järjestelmien välinen virtaus	Tietovirrat aiheuttavat epäyhteneväisyyksiä datassa
Juurisyyt ja datan tarpeettomuus	Dataa ei päivitetä vaatimuksien mukaiseksi ohjelmistojen kehittyessä
Järjestelmärkkitehtuuri ja -evoluutio	Data on siiloutunutta ja sitä kehitetään paikallisesti
Standardisointi	Datalla ei ole yhtenäistä formaattia tai yhtä syöttöpistettä
Ristiriita todellisuuden kanssa	Tiedot eivät vastaa todellisuutta, esim. varusteiden käyttäjiä ei voida tietää
Datan syöttö/validointi	Jatkoa standardisoinnille: dataa joudutaan siivoamaan
Dataan pääsy ja turvallisuus	Tietoja haetaan vanhoista järjestelmistä, jotka eivät ole käyttäjille tuttuja
Yksi päädatalähde	Tarvitaan yksi varasto datalle metadatan kokoamista varten
Viestintä/hallinnon monimutkaisuus	Ohjelmistopäivityksistä tiedottamiseen tarvitaan yhteinen alusta
Omistajuus ja vastuu	Datayksiköillä pitäisi olla vastuuhenkilö
Metodologia	Jatkuvaa laadun kehittämistä varten tarvitaan metodologia
Datan ristiriidoista palautuminen	Ohjelmistovirheiden aiheuttamia virheitä datassa ei aina korjata

Chen et al. (2017) ovat tutkineet datan laatuongelmia älykkäissä sähköverkoissa. He jakavat sähkön kulutusdatan laatuongelmat kolmeen kategoriaan: kohinadata, epätäydellinen data sekä poikkeamadata. Kohinadata tarkoittaa järjestelmille vaikeasti ymmärrettävää dataa, joka rikkoo joko tietomallin tai liiketoimintalogiikan sääntöjä. Epätäydellinen data ei ole välttämättä ongelma sähkönkulutusdatassa, sillä puuttuvat kohdat voivat sisältää hyödyllistä tietoa. Poikkeamadata on näistä kategorioista merkittävin, sillä kaikki poikkeamat datassa eivät ole virheellisiä, vaan ne voivat johtua esimerkiksi hajonneesta laitteesta, käyttökatkosta tai muusta vastaavasta reaalimaailman tilanteesta, jota data edelleen kuvaa tarkasti. Poikkeamista on siis tärkeää erotella todelliset häiriötilanteet ja

virheellinen data. (Chen et al. 2017) Poikkeamien luonnolliset esiintymät voivat tehdä sähköverkon datan laatuongelmien tunnistamisesta hankalaa, sillä tarkkuuden tai oikeellisuuden arviointi vaatii alan asiantuntemusta ja tietoa mahdollisista reaalimaailman poikkeustilanteista.

2.4.2 Laatuongelmien juurisyyt

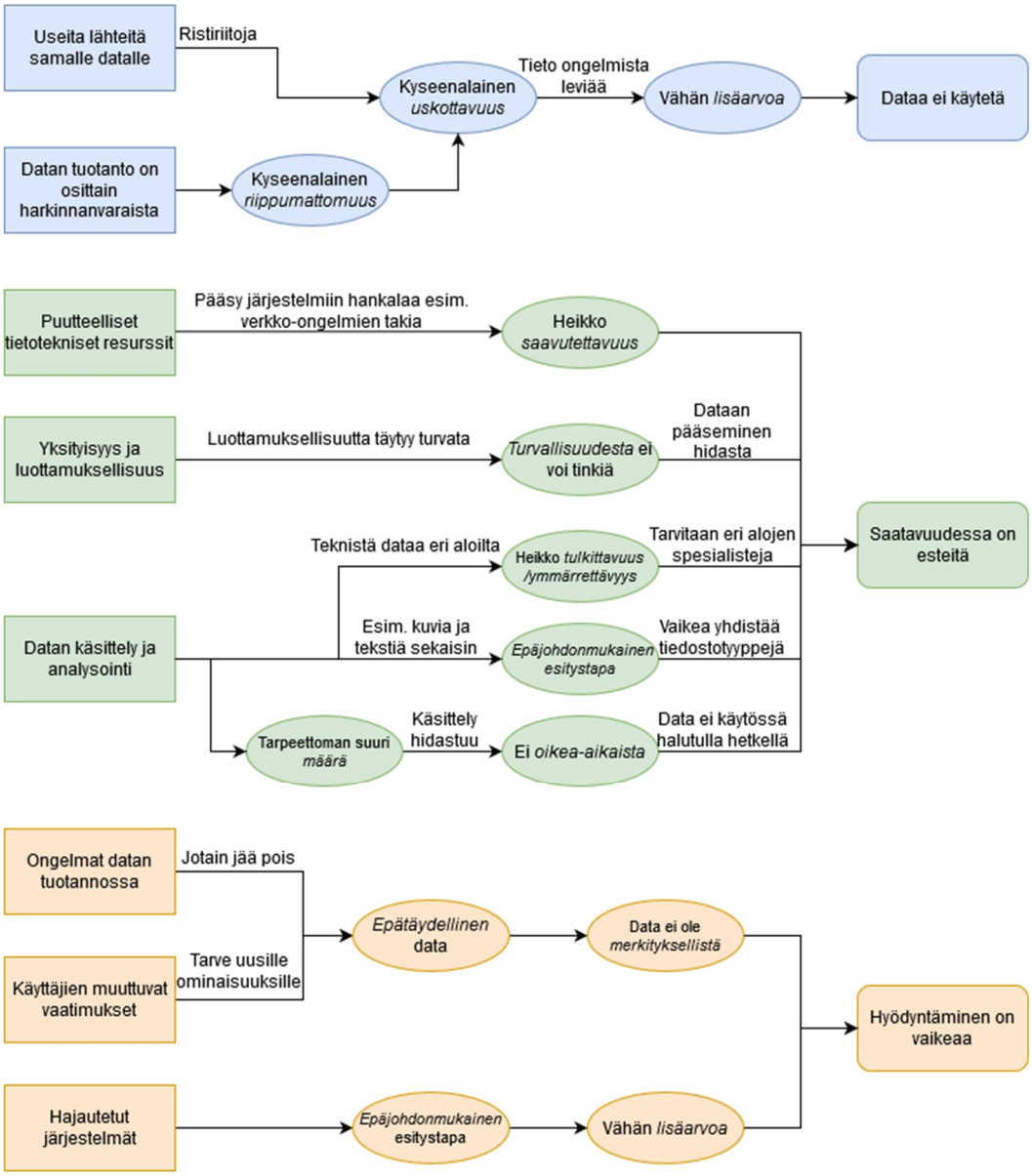
Ongelmien taustalla piilevät tekijät voivat olla hyvin erilaisia. Yoon et al. (2000) jakavat epätäydellisen datan taustatekijät kahteen luokkaan: käytäntöpainotteiset tekijät johtuvat epätäydellisen datan keräämisestä tai käsittelystä tietojärjestelmässä, kun taas rakennepainotteiset tekijät ovat seurausta käyttäjän vaatimusten ja varsinaisen datajärjestelmän toiminnallisuuden ristiriidoista. Käytäntöpainotteisia tekijöitä voidaan korjata perusteellisilla datanhallintamenetelmillä, kun taas rakennepainotteisten ongelmien korjaaminen vaatii perustavanlaatuisia muutoksia data-arkkitehtuuriin. Maydanchik (2007) puolestaan tunnistaa kolme kategoriaa datan laatua heikentäville prosesseille: dataa ulkopuolelta tuovat prosessit (esimerkiksi manuaalinen syöttö ja reaaliaikaiset rajapinnat), sisältä dataa muuttavat prosessit (esimerkiksi datan prosessointi) sekä datan rappeutumista aiheuttavat prosessit (esimerkiksi järjestelmäuudistukset).

Käytännössä virheellisten arvojen takana voi olla monia syitä. Merkittävä osa virheellisestä datasta syntyy ihmisten virheistä tietojen syöttövaiheessa (Mahanti 2019; Umar et al. 1999). Muita syitä datan laadun heikkenemiseen voivat olla datan epäselvä määrittely tai epäyhtenäinen datamalli, joka johtaa virheisiin datassa, kun käytössä on useita tietojärjestelmiä. Tietojärjestelmien tasolla myös järjestelmien väliset integraatiot voivat heikentää datan laatua, kun integraatioiden yhteydessä osa datasta jää siirtämättä tai arvot ovat väärässä paikassa. (Silvola et al. 2011) Integraatioiden yhteydessä havaitut laatuongelmat voivat johtua teknisestä virheestä tiedonsiirrossa tai lähdejärjestelmässä olleista korjaamattomista puutteista (Mahanti 2019, s. 353).

Ongelmien syyt ja seuraukset voivat muodostaa monimutkaisen ketjun. Strong et al. (1997) tutkivat kolmea yritystä ja tunnistivat niissä kolme yleistettävää ongelmaa: dataa ei käytetä, sen saatavuudessa on esteitä tai sen hyödyntäminen on hankalaa. Näiden ongelmien taustalta tunnistettiin edelleen rakenteita, jotka kertovat syistä niiden takana. Käyttämättömyys johtuu datan heikosta lisäarvosta, joka on peräisin monien datalähteiden tai subjektiiviseksi koetun datan tuotantoprosessin huonosta uskottavuudesta ja riippumattomuudesta. Saatavuuden esteet voivat olla peräisin muun muassa tietoturvavaatimuksista, datan vaikeasta ymmärrettävyydestä tai suurten datamassojen prosessoinnin hitaudesta. Hyödyntäminen voi olla hankalaa heikon lisäarvon tai merkityksellisyyden

takia, mitkä voivat aiheutua datan epätäydellisyydestä tai epäjohdonmukaisesta esitystavasta eri järjestelmien välillä.

Myös Lee et al. (2006, s. 80–92) löytävät laatuongelmille samat juurisyyt ja rakenteet hieman eri sanoilla kuvattuna. Tarkemmin ongelmien syy-seuraussuhteita on esitelty kuvassa 6. Ongelmiin liittyvät laatu-ulottuvuudet on esitetty ellipsien sisällä ja ulottuvuuksien nimet on kursivoitu. Luontaiseen laatuun liittyvä rakenne on esitetty sinisellä, saatavuuslaatuun liittyvä rakenne vihreällä ja asiayhteydestä riippuvaan laatuun liittyvä rakenne oranssilla taustavärillä. Tarvittaessa rakenteen eri osien välistä suhdetta on avattu lyhyesti tekstimuodossa.



Kuva 6. *Laatuongelmien rakenteet (mukaillen Lee et al. 2006, s. 92; Strong et al. 1997)*

2.4.3 Esteet laadukkaalle datalle

Luvun 2.4.2 katsauksesta voidaan todeta, että organisaatioissa tunnistetut datan laatuongelmat eivät välttämättä ole ainoastaan virheellistä dataa, vaan esimerkiksi puutteelliset vastuut ja heikosti käyttöön soveltuvat tietojärjestelmät voivat olla esteenä datan korkean laadun saavuttamiselle. Tällaisia esteitä on tutkittu kirjallisuudessa laajalti, ja ne saattavat olla paremmin yleistettävissä kuin datalähtöiset ongelmat. Viiden tutkimuksen (Haug et al. 2013; Silvola et al. 2011; Haug & Arlbjørn 2011; Redman 2004; Umar et al. 1999) nostamat erityishuomiota vaativat tekijät on esitelty laajemmin taulukossa 5.

Umar et al. (1999) analysoivat datan laadun ongelmia hallinnollisesta näkökulmasta ja listaavat mallissaan kuusi potentiaalista ongelmaa, joihin tulisi kiinnittää huomiota. Näitä ovat muun muassa roolien ja vastuiden jakaminen, henkilöstön palkitseminen sekä muut hallinnolliset yksityiskohdat, kuten henkilöstön koulutus ja viestintä. On huomattava, että tässä tutkimuksessa ei varsinaisesti tutkittu esteitä hyvälaatuisen datan saavuttamiseen, vaan listattuihin tekijöihin on tärkeää kiinnittää huomiota datan laadun kehittämiseksi. Sen sijaan Redman (2004) vie ajatusta pidemmälle ja listaa 12 estettä onnistuneelle datan laadun hallinnalle, joista erityisen tärkeiksi nostetaan kaksi: heikko ymmärrys datan laadun ja liiketoiminnan tuloksen välisestä yhteydestä sekä vastuiden määrittäminen väärille tahoille. Myös Silvola et al. (2011) korostavat vastuun suurta roolia nostamalla epäselvät datan omistajuusmäärittelyt yhdeksi prosessilähtöisistä ongelmista epäselvien datanhallintakäytäntöjen ja jatkuvien datan laatukäytäntöjen puutteen ohella. Omistajuuden määrittely voi olla paikoitellen puutteellista, tai se voi puuttua organisaatiolta kokonaan.

Osa tutkimuksista painottaa erityisesti ydintiedon laadun hallintaa ja sen esteitä. Haug & Arlbjørn (2011) tunnistavat aiemmasta kirjallisuudesta viisi ylätason estettä, jotka edelleen korostavat selkeitä vastuita, palkitsemiskäytäntöjä ja henkilöstön osaamista. Haug et al. (2012) löytävät 12 estettä, jotka täydentävät aiempia listauksia muun muassa korostamalla johdon roolia datan laadun hallinnan tärkeyden esiintuomisessa. Tutkimuksessa myös havaittiin tietojärjestelmäesteiden roolin olevan pienempi suurilla yrityksillä. Vastaavasti organisatoristen ongelmien painoarvo oli niillä suurempi.

Taulukko 5. *Potentiaaliset ongelmat tai esteet datan laadun hallinnassa*

Potentiaalinen ongelma tai este	**Ilmeneminen kirjallisuudessa**
Vastuunjako	Rooleja ja vastuita ei määritelty, datan laadulla ei ole omistajaa (Umar et al. 1999); Vastuiden määrittäminen väärille tahoille (Redman 2004); Puutteelliset ylläpitovastuut (Haug & Albjørn 2011); Epäselvät datan omistajuuksien määrittelyt (Silvola et al. 2011); Tiettyjen ydintietojen vastuutusten puuttuminen, datan luonti-, käyttö- ja ylläpitovastuisiin liittyvät epäselvyydet (Haug et al. 2013);
Palkitseminen	Ei palkitsemis-/nuhtelujärjestelmää (Umar et al. 1999); Puutteellinen palkitseminen oikeellisuudesta (Haug & Albjørn 2011); Puutteellinen palkitseminen tai nuhtelu datan laadusta (Haug et al. 2013)
Henkilöstön osaaminen	Viestintää ja kouluttamista ei ole hoidettu (Umar et al. 1999); Puutteellinen henkilöstön osaaminen (Haug & Albjørn 2011); Datan käyttäjien puutteellinen kouluttaminen (Haug et al. 2013)
Tietojärjestelmät	Väärinymmärrykset tietotekniikan roolista (Redman 2004); Tietojärjestelmien heikko käyttäjäystävällisyys (Haug & Albjørn 2011, Haug et al. 2013); Datan hallintaan ei ole sopivia tietojärjestelmiä, nykyisissä tietojärjestelmissä ei ole sopivia tiedonsyöttömahdollisuuksia (Haug et al. 2013)
Puutteelliset menettelytavat datan hallinnassa	Organisatoristen toimintamalleja ei kehitetä (Umar et al. 1999); Ymmärtämättömyys datan laadun hallinnan parhaista käytännöistä, hallinnon ja datavirtojen heikko yhteensopivuus (Redman 2004); Puutteelliset ydintiedon hallintarutiinit (Haug & Albjørn 2011); Epäselvät datan hallintamenetelmät, ei jatkuvia datan laatukäytäntöjä (Silvola et al. 2011); Tehottomat menettelytavat organisaatisossa (Haug et al. 2013)
Mittaaminen	Nykytilan mittaamisen välttely totuuden pelossa (Redman 2004); Datan laatua ei mitata (Haug et al. 2013)
Muut ongelmat	Aikataulutusskenaarioita ei määritelty (Umar et al. 1999); Heikko ymmärrys datan laadun ja liiketoiminnan yhteydestä, vallankäyttö datan jakamisessa, haluttomuus organisaatiorajojen ylittämiseen, datan standardoimisen vaikeus, epärealistiset tavoitteet, "laatu"-termin negatiiviset mielikuvat, puutteellinen yksityisyyteen liittyvä sääntely (Redman 2004); Johto ei keskity datan laatuun riittävästi, kirjallisia laatuohjeistuksia ei ole, johto ei korosta datan laadun tärkeyttä riittävästi (Haug et al. 2013)

Taulukossa voi havaita vastuunjakoon liittyvien haasteiden korostuvan – kaikki viisi tutkimusta nostavat ne esiin yhdellä tai useammalla tavalla. Toisin sanoen datan laadun täytyy olla jonkun vastuulla, jotta se voi olla hyvällä tasolla. Tämä on myös linjassa aiem-

min esiteltyjen ydintiedon hallinnan ja datan hallinnoinnin periaatteiden kanssa. Hallinnointiin viittaavat myös maininnat puutteellisista menettelytavoista sekä mittaamista. Kolme tutkimusta mainitsee datan laadun palkitsemisjärjestelmän puutteen, joka omalta osaltaan myös vastuuttaisi organisaation toimijoita huolehtimaan laadusta omalta osaltaan. Samaa teemaa täydentää aineistosta esiin nouseva henkilöstön osaamisteema. Hallinnollisissa ongelmissa nousi esiin myös läheisesti tietojärjestelmiin liittyviä ongelmia – on huomionarvoista, että niistäkin vain osa (käyttäjäystävällisyys, tiedon syöttö) on suoranaisesti tekniikkaan liittyviä. Muissa ongelmissa näkyy vahvasti johdon rooli datan hallinnan onnistumisessa: johdon täytyy panostaa datan hallintaan ja organisaation tulee ymmärtää datan merkitys liiketoiminnalle.

2.5 Datan laadun kehittäminen

Datan laatua voidaan parantaa useilla eri menetelmillä. Redmanin (2008, s. 55) mukaan ne voidaan jakaa kahteen luokkaan: ongelmat voidaan etsiä ja korjata, tai ne voidaan estää niiden alkulähteillä. Mahanti (2019, s. 319) käyttää samaa luokittelua puhuen reaktiivisesta ja proaktiivisesta lähestymistavasta. Molemmat lähteet pitävät proaktiivista eli ennaltaehkäisevää lähestymistapaa parempana, sillä yksittäiset virheet datassa voivat kertautua nopeasti, jolloin niiden korjaaminen voi osoittautua kalliiksi. Batini et al. (2009) luokittelevat kehitysmenetelmät hieman samaan tapaan data- ja prosessijohtoisiin strategioihin. Datajohtoiset menetelmät muokkaavat suoraan datan arvoja esimerkiksi päivittämällä tietokannan arvoja, kun taas prosessijohtoiset menetelmät pohjautuvat datan käsittelyprosessien uudelleensuunnitteluun esimerkiksi lisäämällä prosessiin valvontatoiminnon ennen datan tallentamista. Lee et al. (2006, s. 106) käyttävät vastaavaa jaottelua ja pitävät datan tuotantoon kohdistettuja prosessiorientoituneita menetelmiä välttämättöminä datan laadun kehittämiseksi. Myös Umar et al. (1999) jaottelevat menetelmät datan siivoamiseen ja prosessien siivoamiseen.

Silvola et al. (2011) jakavat lähestymistavat tarkemmin neljään kategoriaan: passiiviseen, reaktiiviseen, aktiiviseen ja proaktiiviseen. Passiivisella tasolla datan laatua ei valvota lainkaan, ja ongelmien ilmaantuessa organisaatio siirtyy reaktiiviselle tai aktiiviselle tasolle, jossa ongelma pyritään korjaamaan. Tämän jälkeen organisaatio voi siirtyä jälleen passiiviseen tilaan. Aktiivisella tasolla datan laatua valvotaan reaaliaikaisesti, ja proaktiivisella tasolla ongelmat ehkäistään ennen niiden syntymistä.

Osa tällaisista keinoista on sidottu yhteen datan arviointimenetelmien kanssa, ja nykytilan tunteminen on välttämätöntä ennen kehitystoimenpiteiden toteuttamista (Woodall et al. 2013). Batini et al. (2009) tunnistavat menetelmien kehitysosioista seuraavat vaiheet:

1. *Kustannusten arviointi*
   Arvioidaan datan laadusta aiheutuvat suorat ja epäsuorat kustannukset
2. *Prosessivastuiden asettaminen*
   Tunnistetaan prosessiomistajat ja määritellään heidän vastuunsa datan laadun suhteen
3. *Datavastuiden asettaminen*
   Tunnistetaan dataomistajat ja määritellään heidän vastuunsa datan laadun suhteen
4. *Ongelmien syiden tunnistaminen*
   Tunnistetaan laatuongelmien taustalla olevat syyt
5. *Strategioiden ja tekniikoiden valinta*
   Valitaan datan laadun kehittämiseen tehokkain strategia ja siihen sopivat tekniikat ja työkalut
6. *Prosessien hallinta*

   Määritellään tarkastuspisteet datan tuotantoprosesseissa datan laadun valvomiseksi
7. *Prosessien uudelleensuunnittelu*

   Määritellään datan laadun kehittymiseen tähtäävät prosessien kehitystoimenpiteet
8. *Kehityksen hallinta*

   Määritellään organisaation laajuiset säännöt datan laadulle
9. *Kehityksen seuranta*

   Asetetaan säännölliset seurantatoimenpiteet, jotka tarjoavat tietoa kehitysprosessin vaikutuksista

Luettelon ensimmäiset vaiheet ovat hallinnollisia ja valmistelevia, ja vasta ongelmien tunnistamisen jälkeen voidaan tietää, tuleeko ongelmia lähestyä data- vai prosessilähtöisellä strategialla. Seuraavissa aliluvuissa eritellään tarkemmin eri lähestymistavoissa hyödynnettäviä kehitystoimenpiteitä sekä tavoitteiden asettamisessa hyödynnettäviä organisaation datan laadun kypsyysmalleja.

2.5.1 Proaktiiviset menetelmät

Proaktiivisessa lähestymistavassa puutteet datan laadussa pyritään ehkäisemään jo niiden alkulähteillä suurempien ongelmien välttämiseksi. Tämä kuitenkin vaatii resursseja dataongelmien syiden tunnistamiseen, mikä näkyy myös datan laadun kehittämismenetelmissä: Batini et al. (2009) vertailussa virheiden aiheuttajien tunnistaminen on menetelmien yleisin vaihe. Yksittäisen ongelman taustalla voi olla monta tekijää, joten juurisyiden selvittäminen voi olla haastavaa. Juurisyiden tunnistamiseen ei ole yhtä kaikkiin tapauksiin sopivaa menetelmää, vaan oleellista on toteuttaa perusteellinen selvitys. (Lee et al. 2006, s. 109) Juurisyiden tunnistaminen vaatii yleensä teknologia- ja liiketoiminta-asiantuntijoiden yhteistyötä (Mahanti 2019 s. 331; Loshin 2011 s. 215–216)

Yksi konkreettinen apukeino ongelmien juurisyiden löytämiseen on tietovirran mallintaminen datan tuotannosta sen eri käyttökohteisiin (Loshin 2011 s. 212, Silvola et al. 2011). Mahdollisen virheen havaitsemisen jälkeen tietovirtaa seurataan taaksepäin, kunnes virheen syntykohta löytyy esimerkiksi datan tuotannosta, muokkaamisesta tai siirrosta järjestelmien välillä. (Loshin 2011 s. 215) Mallintamisessa voidaan hyödyntää IP-MAP-työkalua (Information Production Map), jolla tietovirta voidaan mallintaa kahdeksan elementin avulla: datalähteet, (dataa hyödyntävät) prosessoinnit, datavarastot, päätöskohdat, laatutarkastukset, tietojärjestelmien rajat, organisaatio- tai liiketoimintaprosessien rajat sekä tietotuotteet (Loshin 2011 s. 214, Shankaranarayan et al. 2003). IP-MAP pohjautuu samaan näkökulmaan tiedosta tuotteena kuin luvussa 2.2 esitelty TDQM-viitekehys (Shankaranarayan et al. 2003). Juurisyiden tunnistamisen lisäksi IP-MAP-visualisointi auttaa ymmärtämään tietovirtojen kokonaisuuksia yleisellä tasolla (Silvola et al. 2011). Batini et al. (2009) kuitenkin huomauttavat, että IP-MAPin vaatima prosessien mallinnus voi olla hyvin kallista sekä joissain tapauksissa myös mahdoton toteuttaa käytännössä.

Aktiivinen kehitystyö vaatii myös jatkuvaa valvontaa, ja kirjallisuudessa jatkuva datan laadun seuranta mainitaan usean datan laadun kehittämismallin osana. Loshinin (2011, s. 17–18) esittämässä viisiosaisessa mallissa tunnistetaan dataongelmien vaikutukset, määritellään datan laatutavoitteet, suunnitellaan ja toteutetaan laatua parantavat toimenpiteet sekä lopulta valvotaan datan laatua vertaamalla nykytilaa määriteltyihin tavoitteisiin. Jos valvonnassa paljastuu ongelmia, sykli alkaa jälleen alusta. McGilvrayn (2008) mallissa korjaavien toimenpiteiden jälkeen suunnitellaan ja otetaan käyttöön jatkuva valvonta ja mittarit, jotta toimenpiteiden vaikutusta voidaan seurata eikä organisaatio palaa vanhaan malliin ongelmineen. Sebastian-Colemanin (2013, s. 117–119) DQAF-mallin kantavana ajatuksena on mittaamiseen perustuva jatkuva kehittäminen, mikä on peräisin valmistavan teollisuuden laadunvalvontafilosofiasta. Toinen mallin jatkuvan mittauksen etu on nopeampi reagointi datan muutoksiin, jotka voivat syntyä teknisten tai liiketoimintaprosessien muuttuessa. Myös Silvola et al. (2016) toteavat valvonnan mahdollistavan jatkuvan kehittämisen.

2.5.2 Reaktiiviset menetelmät

Mikäli ongelmien lähdettä ei voida poistaa, virheellistä dataa voidaan korjata suoraan muokkaamalla sitä tai korvaamalla se kokonaan. Ennen korjaustoimia ongelmat tulisi priorisoida, jotta resursseja käytetään tehokkaasti (Loshin 2011 s. 208–212). Tavoitteena datan laatua kehittäessä ei tulisi olla kaikkien ongelmien ratkaisu, vaan riittävän hyvä tilanne (Silvola et al. 2011). Datalähtöisiä kehitysmenetelmiä ovat muun muassa datan korvaaminen uudella laadukkaammalla datalla, standardointi (poikkeavien arvojen

korvaaminen standardilla, esimerkiksi lempinimen korvaaminen oikealla nimellä) ja tietueiden linkitys (samaa asiaa eri tietokannoissa kuvaavien tietojen yhdistäminen) (Batini et al. 2009).

Myös reaktiivinen datan laadun kehittäminen vaatii ongelmien havaitsemista vähintään kertaluontoisesti. Yksi tätä helpottava menetelmä on datan profilointi, eli metatietotiivistelmän teko eri datajoukoista erilaisia analyysimenetelmiä hyödyntäen (Abedjan et al. 2015). Käytännössä profiloinnissa hyödynnetään erilaisia algoritmeja, jotka tuottavat tietoa datajoukon mahdollisista laatupuutteista (Loshin 2011 s. 241), kuten esimerkiksi epäyhteneväisistä formaateista, puuttuvista arvoista tai selvistä poikkeamista (Abedjan et al. 2015). Profiloinnin lopputuloksena saadaan tarkempi kuva datan rakenteesta, sisällöstä, sisäisistä säännöistä sekä suhteista (Sebastian-Coleman 2013 s. 49). Datan profilointi voidaan kohdistaa yksittäiseen sarakkeeseen, sarakkeiden vertailuun tai kokonaisten tietokantataulujen vertailuun (Loshin 2011 s. 245–247). Datan profilointiin on tarjolla lukuisia valmiita tietojärjestelmäratkaisuja, joskin ne eivät kykene jatkuvaan datan laadun valvontaan (Ehrlinger et al. 2019).

2.5.3 Organisaation kypsyysmallit

Kehittämismallien tarjoamien tekniikoiden ohella datan laadun kehittämisessä voidaan hyödyntää organisaation kyvykkyyksiä tarkastelevia kypsyysmalleja. Kypsyysmallit tarjoavat apua organisaation kehityskohteiden tunnistamiseen visualisoimalla datan laatutoimintojen nykytilaa ja mahdollista tavoitetasoa (Loshin 2011). Organisaatio voi siis tunnistaa heikkouksia toiminnassaan ja pohtia mahdollisia kehitystoimenpiteitä seuraavien tasojen pohjalta. Osa kirjallisuudesta keskittyvistä kypsyysmalleista keskittyy suoraan datan laadun hallinnan kypsyystasoon (katso Mahanti 2019; Loshin 2011), kun taas osassa datan laatu on upotettu osaksi datan hallinnoinnin tai ydintiedon hallinnan kypsyyden arviointia (katso Spruit & Pietzka 2015). Myös datan laadun kypsyysmallit voidaan pilkkoa edelleen osiin: esimerkiksi Loshin (2011) tarkastelee mallissaan erikseen odotuksia datan laadulle, datan laadun ulottuvuuksia, toimintaperiaatteita, menettelytapoja, hallinnointia, standardeja, teknologiaa sekä suorituskyvyn johtamista. Näille osa-alueille on erikseen kuvatut sisällöt jokaiselle tasolle, mikä helpottaa mallin hyödyntämistä. Vastaavasti Mahanti (2019, s. 295) jakaa kypsyyden tarkastelun tekniikkaan, asenteeseen, lähestymistapaan, ihmisiin ja hyötyihin.

Taulukossa 6 esitellään tiivistettynä Loshinin (2011) ja Spruitin & Pietzkan (2015) mallit, jotka sisältävät Mahantin (2019) mallin tapaan viisi eri tasoa organisaation kypsyyden

arviointiin: alkeellinen, toistettava, määritelty, hallittu sekä tehokas. Valtaosa organisaatioista sijoittuu kahdelle ensimmäiselle tasolle (Mahanti 2019, s. 294), joten keskimäärin datan laatun hallinta on hyvin alkeellista.

Taulukko 6. *Datan laadun kypsyystasot*

Taso	**Loshin (2011)**	**Spruit & Pietzka (2015) data quality**
Alkeellinen	Toiminta on laatuongelmiin reagointia ja prosessit muodostetaan tarpeen mukaan. Korjaustoimenpiteet ovat yksittäisiä, eivätkä todennäköisesti paranna laatua pitkällä aikavälillä. Tietoa ja kokemuksia ei jaeta.	Organisaatiossa on tunne siitä, että data on hyvä- tai heikkolaatuista. Tiettyjen laatuongelmien mainehaitat tiedostetaan. Tiedetään, että heikkolaatuisen datan taustalla on useita syitä. Tunnistetaan alueita, jossa laatu on heikkoa.
Toistettava	Alustava hallinnointipohja on olemassa rajallisen dokumentaation muodossa. Tietoa jaetaan prosessien mukaisesti ja hyviä käytäntöjä tunnistetaan. Käytäntöjen käyttöönottotahti vaihtelee eri yksiköiden välillä. Teknologiafokus ajaa liiketoimintatarpeiden yli.	Datan laadun mitattavat osa-alueet ovat tiedossa. Tiettyjen laatuongelmien suorat kustannukset tiedetään. Tiedetään, mitkä tekijät aiheuttavat puutteita datan laadussa. Korkealaatuisen datan merkityksestä ollaan tietoisia.
Määritelty	Datan hallinnointiohjeistukset, datan laatuodotusten määrittelyprosessit, teknologiakomponentit ja datan laadun valvontaprosessit ovat dokumentoitu ja saatavilla koko organisaatiolle. Vastuuroolit datan laadulle on määritelty ja niitä valvotaan hallinnointiryhmässä.	Datan laatu on määritelty sidosryhmien tarpeiden kautta. Huonolaatuisen ydintiedon taloudelliset vaikutukset tiedostetaan. Heikkolaatuisen datan taustarakenteita on tutkittu. Organisaatiossa on käytössä datan laadun vertailujärjestelmä.
Hallittu	Datan laadun seurannassa huomioidaan liiketoimintavaikutukset. Laatuodotusten toteutumista seurataan painotettujen mittareiden avulla. Datan laatua kehitetään proaktiivisesti ja puutteet havaitaan tietovirran alkuvaiheissa. Korjaavia toimenpiteitä hallitaan dokumentoitujen prosessien avulla.	Datan laatua mitataan objektiivisesti ja jokaisen ydintietoyksikön laatu on tiedossa. Huonolaatuisen ydintiedon ei-taloudelliset vaikutukset (maine, asiakaspito jne.) tiedostetaan. Työntekijät ovat tietoisia laatuongelmien syistä ja niiden vaikutuksista päivittäisessä työssään. Kehitystoimenpiteitä on tehty.
Tehokas	Datan laadun kehitysmahdollisuudet havaitaan koko yrityksen laajuisen suorituskykymittariston avulla. Strategista kehitystyötä ja jatkuvaa prosessien valvontaa tehdään visuaalisten raporttien avulla.	Jokaiselle dataryhmälle toteutetaan säännöllisesti laatuarviointi. Huonolaatuisen ydintiedon vaikutukset liiketoimintaan osataan esittää rahallisin perustein. Syyt heikon laadun taustalla ja niiden ilmentymät tiedetään. Laatua valvotaan säännöllisesti vertailujärjestelmän avulla.

Taulukosta voidaan nähdä, että jo alkeellinen taso vaatii organisaatiolta jonkinlaista ymmärrystä datan laadusta ja sen vaikutuksista. Spruit & Pietzka (2015) toteavat tämän olevan tarkoituksellista: jos organisaatiossa ei ole mitään tietoa nykytilan ongelmista, ei sillä ole myöskään kypsyyttä. Kypsyystasoja ei siis voida hyödyntää, jos organisaatiossa ei ole lainkaan ymmärrystä datan laadun nykytilasta. Alkeellisella tasolla toimintaa leimaa lyhytjänteisyys ja reaktiivisuus, mutta organisaatiolla on kuitenkin jonkinlainen käsitys nykytilan haasteista. Toistettavalla tasolla ongelmien vaikutuksista ja syistä niiden taustalla voi olla alkeellinen ymmärrys, ja dokumentaatiota ja standardeja on otettu jollain tasolla käyttöön. Määritellyllä tasolla datan laadun hallinta on jo varsin organisoitua ja

lähestymistavaltaan ennaltaehkäisevää: laatuodotukset ja liiketoimintasäännöt on dokumentoitu, ja laatua myös valvotaan. Hallitulla ja tehokkaalla tasolla korostuu liiketoimintavaikutukset ja niiden huomiointi: dataa voidaan pitää yhtenä keinona saavuttaa kilpailuetua, joten sen laatua valvotaan raportointityökaluilla läpi koko organisaation.

Mahantin (2019) osiin pilkottu malli (taulukko 7) mahdollistaa nopean kypsyysarvioinnin yksinkertaisten tasojensa kautta. Näin esitettynä mallissa korostuvat organisatoriset tekijät, kuten asenne, lähestymistapa ja henkilöstön roolitus, mikä on linjassa luvun 2.3 ja 2.4 päätelmien kanssa – datan laatuun liittyvät ongelmat ja esteet ovat harvoin puhtaasti teknisiä.

Taulukko 7. *Datan laadun kypsyysmalli osineen (mukaillen Mahanti 2019, s. 295)*

	Alkeellinen	**Toistettava**	**Määritelty**	**Hallittu**	**Tehokas**
Tekniikka	Yleissovellus (esim. Excel), manuaalisia prosesseja, tarpeen mukaan toteutettuja rutiineja	Taktisen tason työkaluja sovellustasolla tai yksiköissä siiloutuneesti	Laatutyökaluja profilointiin ja siivoamiseen, tietovarasto, BI-sovelluksia	Tietovarastoa ja BI-sovelluksia pidemmälle vietyjä laatutyökaluja, metatiedon hallintatyökaluja	Työkalut on standardoitu organisaation läpi. Alustaratkaisu datan profilointiin, valvontaan ja visualisointiin.
Asenne	Datan laatu nähdään kustannuksena	Alustava tietoisuus datan hallinnan merkityksestä	Dataa käsitellään organisaatiotasolla tuloksen kannalta kriittisenä	Dataa käsitellään mahdollisena kilpailuedun lähteenä	Data nähdään kriittisenä ja datan laatu mahdollistajana
Lähestyminen	Tulipalojen sammuttamista tarpeen vaatiessa, ei juurisyiden analysointia	Datan laadun dokumentointi mahdollistaa toistettavuuden	Isommat ongelmat dokumentoitu, mutta ei kokonaan ratkaistu	Proaktiivinen ehkäisevä työ	Strategista optimointia
Ihmiset	Ei datarooleja, ei tietoisuutta datan hallinnan käytännöistä	Tietokannan ylläpitäjä, datan laatu IT:n vastuulla	Datan ylläpitäjä, tietovastaava ja -omistajaroolit nousemassa	Useamman tason tietovastaavaroolit käytössä	Keskeinen datarooli
Hyödyt	Ei lainkaan tai rajatusti	Vähän taktisen tason hyötyjä	Olennaisia taktisen tason hyötyjä	Taktisen ja strategisen tason hyötyjä	Strategisen tason hyötyjä

Kypsyysmallit tarjoavat yksinkertaisen visualisaation nykytilan ja seuraavien tasojen vaatimien toimenpiteiden analysointiin. Toisaalta ne eivät tarjoa kovin konkreettisia kehitystoimenpiteitä tai apua datalähtöisten ongelmien havaitsemiseen ja ratkaisemiseen, vaan toimivat ylemmän tason työkaluna.

3. TAPAUSTUTKIMUKSEN TOTEUTUS

Tutkimuksen empiirinen osio koostuu yhden tapauksen tapaustutkimuksesta, eli siinä analysoidaan ajallisesti ja tilallisesti rajattua tapausta, joka on jollain tavalla esimerkki tutkittavasta ilmiöstä (Vuori 2021). Tässä luvussa kuvataan tapaustutkimuksen kohteena oleva organisaatio taustayhtiöineen sekä käydään läpi aineiston keräämisen ja sen analysoinnin toteuttaminen valintoineen.

Kohdeorganisaationa toimii kantaverkkoyhtiö Fingrid Oyj:n Voimajärjestelmän käyttö -toiminto, jonka tehtävänä on pitää Suomen sähköjärjestelmä jatkuvasti toimintakykyisenä. Kohdeorganisaation hyödyntämä ydintieto ei ole datanhallinnan kontekstissa tyypillistä asiakas- tai tuotetietoa, vaan pääsääntöisesti sähköverkon tilasta ja käytöstä kertovaa reaaliajassa kerättävää ja hyödynnettävää aikasarjadataa. Näin ollen sitä voidaan pitää epätyypillisenä tapauksena, jolloin on perusteltua toteuttaa tapaustutkimus yhden tapauksen tarkasteluna (Yin 2018, s. 47–51).

Tutkimuksen aikahorisontti on läpileikkaava, eli tutkimuksessa kuvataan organisaation nykytilaa. Tämä on tarkoituksenmukaista, sillä tavoitteena oli tunnistaa nimenomaan nykyhetken ongelmia ja valmistaa organisaatiota tulevien kehityshankkeiden vaatimusten mukaiseksi muotoilemalla toimenpide-ehdotuksia havaittujen ongelmien korjaamiseksi. Nykytilaa haluttiin tutkia keräämällä laadullista aineistoa kohdeorganisaation dataa työssään hyödyntäviltä asiantuntijoilta luvussa 2.1 esitellyn datan laadun *fitness for use* -määritelmän mukaisesti. Arviointi toteutettiin luvussa 2.2.2 esiteltyyn Lee et al. (2002) AIMQ-menetelmän kyselylomakkeeseen pohjautuen subjektiivisiin kokemuksiin keskittyen. Menetelmän kehittäjien mukaan se soveltuu hyvin datan laadun ongelmien tunnistamiseen (Lee et al. 2002). Subjektiiviseen arviointiin päädyttiin, sillä kohdeorganisaatiolla ei ollut vielä tietoa mahdollisten ongelmien juurisyistä, vaikutuksista tai luonteesta ja laaja-alaisten objektiivisten mittausten toteuttaminen ilman pohjatietoja olisi ollut hyvin työlästä. Esimerkiksi Lee et al. (2006, s.27), Batini et al. (2009) ja Woodall et al. (2013) määrittävät nykytilan ongelmien selvityksen datan laadun arviointi- ja kehitysprosessin ensimmäiseksi vaiheeksi, jonka pohjalta työtä voidaan jatkaa.

3.1 Kohdeorganisaatio

Voimajärjestelmän käyttö -toiminnon ydintehtäviin kuuluu muun muassa häiriötilanteiden selvittämistä, sähkön tuotannon ja kulutuksen tasapainon ylläpitoa sekä sähkönsiirtoka-

pasiteetin tarjoamista maan sisällä ja maiden välillä. Toiminnon kriittisin osa-alue on kantaverkkokeskus, joka valvoo reaaliaikaisesti voimajärjestelmää. Näin ollen datalla ja reaaliaikaisella tiedonsiirrolla on hyvin suuri rooli toiminnon tehtävissä, sillä keskuksen täytyy saada tietoa vuorokauden ympäri muun muassa sähkön tuotannosta ja kulutuksesta, verkon tilasta sekä sähkön laadusta. Mahdolliset häiriöt vaativat välittömästi toimenpiteitä operaattoreilta, joten tiedon täytyy kulkea mahdollisimman nopeasti ja datan täytyy olla laadukasta. Teknologian kehitys ja vaikeammin ennustettavan uusiutuvan energiatuotannon kasvu vaatii entistä nopeampaa reagointia muutoksiin, joten kantaverkkokeskuksen toimintoja pyritään automatisoimaan mahdollisimman laajasti. Tämä asettaa myös uusia vaatimuksia datalle.

Tutkimuksen toteuttamishetkellä Fingridissä on loppumaisillaan datanhallintahanke, jossa on määritelty tietoalueet, niiden ydintiedot sekä näistä vastaavat henkilöt läpi organisaation. Ydintietojen hallinnan rinnalla kulkee tietovaraston jatkuva kehittäminen sekä analytiikan jalkauttaminen organisaation toimintaan raportointityökalujen käyttöä lisäämällä. Jokaiselle tietoalueelle on määritelty myös tietovarastoinnista sekä analytiikasta vastaavat henkilöt, joiden tehtävänä on edistää niiden hyödyntämistä omassa liiketoimintayksikössään. Lisäksi yrityksessä ollaan ottamassa käyttöön erillistä datakatalogia, jonka on tarkoitus sisältää esimerkiksi tietoalueiden ja ydintietojen kuvaukset metatietoineen.

Toiminnon kannalta relevantit ydintiedot kuuluvat erilliseen Verkon tila- ja käyttötieto -tietoalueeseen, joka jakautuu edelleen neljään tietoryhmään: verkonhallintaan, tasehallintaan, siirtojen hallintaan sekä tasehallintaan. Yrityksen datanhallintamallin mukaisesti tietoalueelle on määritelty toiminnon sisältä tietoalueen omistaja sekä jokaiselle ydintiedolle oma tietovastaava. Ydintietoja on tunnistettu yhteensä 31, joista 12 täytyy olla käytettävissä jokaisena vuorokauden hetkenä. Kaikki ydintiedot liittyvät olennaisesti kantaverkon hallintaan, sähkön tuotantoon ja kulutukseen tai sähkön siirtämiseen verkossa. Tarkemmin tietoalueen ydintiedot on esitelty liitteessä A.

Valtaosa ydintiedoista on tyypiltään aikasarjoja, mutta kaikkiaan tietoalueen ydintiedot ovat hyvin monimuotoisia: esimerkiksi verkonhallinnan tiedoissa on paljon sisäisten asiantuntijoiden tuottamia suunnitelmia, kun taas tasehallinnan tiedoissa korostuvat automaattisesti ja reaaliaikaisesti kerättävät voimajärjestelmän mittaustiedot. Myös ydintietojen rakenne vaihtelee: esimerkiksi erilaiset ennusteet, kuten kulutus- ja tuotantoennuste, muodostetaan erillisessä ennustejärjestelmässä sekä sisäistä (esim. historialliset mittaustiedot) sekä ulkoista dataa (esim. sääennusteet) hyödyntäen. Toisaalta esimerkiksi sähkön siirtotieto perustuu suoraan kantaverkosta kerättävään mittausdataan.

3.2 Aineiston kerääminen

Aineistoa kerättiin puolistrukturoiduilla haastatteluilla. Haastattelukysymykset muodostettiin Lee et al. (2002) AIMQ-menetelmän IQA-kyselylomakkeen pohjalta, joka on esitelty tarkemmin luvussa 2.2.2. IQA-lomakkeen ulottuvuuskohtaiset kysymykset mahdollistavat kattavan nykytilan analyysin sisältäen luvussa 2.1 tunnistetut merkittävimmät datan laadun ulottuvuudet. Aineistonkeruu toteutettiin haastatteluna kyselyn sijaan, sillä näin haastattelija pystyi esimerkiksi tarkentamaan erilaisten laatu-ulottuvuuksien sisältöä haastateltaville tai kysymään esimerkkejä tilanteista, joissa laatuongelma haittaa työskentelyä. Alkuperäisen lomakkeen laatu-ulottuvuuksista tietoturvallisuus (engl. *security*) jätettiin pois, sillä saatavuuden kysymysten koettiin sisältävän pääsynhallinnan mahdolliset ongelmat. Lisäksi ymmärrettävyyden ja tulkittavuuden kysymykset yhdistettiin saman otsikon alle, sillä ne koettiin osittain päällekkäisiksi. Alkuperäisen mallin väitteet on muutettu kysymysmuotoon luontevamman haastattelun aikaansaamiseksi. Myös alkuperäiseen IQA-kyselyyn kuuluvista numeroarvosanoista luovuttiin, sillä tutkimuksen tarkoituksena oli kartoittaa käyttäjien kokemia ongelmia, jolloin numeeristen arvojen vertailulle ei nähty tarvetta.

Haastattelut aloitettiin lämmittelykysymyksellä haastateltavan työnkuvasta, työskentelyajasta kohdeorganisaatiossa ja hänen työssään käyttämistä ydintiedoista, jotta haastateltava osaisi ajatella työn rajauksen mukaista dataa pohtiessaan eri ulottuvuuksien ongelmia. Tämän jälkeen esitettiin kysymyksiä laatu-ulottuvuuksista aloittaen helpommin käsitettävistä ulottuvuuksista, kuten saatavuus ja tarkkuus edeten kohti abstraktimpia ulottuvuuksia. Lopussa ulottuvuuksien läpikäynnin jälkeen haastateltavalta kysyttiin vielä, onko hänellä muuta lisättävää mahdollisista datan laatuun liittyvistä ongelmista. Kokonainen kysymysrunko on esitelty liitteessä B. Kysymysrunkoa testattiin koehaastattelussa, jonka jälkeen runkoon lisättiin kysymys haastateltavan roolista suhteessa ydintietoihin. Haastattelurunko on laaja, sillä kohdeorganisaation mahdollisia datan laadun ongelmia haluttiin tarkastella kattavasti. IQA-lomakkeen tarjoama pohja mahdollistaa tämän, ja samalla vakiintuneen menetelmän käyttäminen lisää tutkimuksen luotettavuutta pienentämällä tutkijan omien ennakkoasenteiden vaikutusta (Saunders et al. 2019 s. 447, Eskola et al. 2018). Toisaalta laaja kysymyspatteristo voi heikentää haastattelun vuorovaikutuksellisuutta (Eskola et al. 2018), joten haastatteluissa keskityttiin erityisesti reagoimaan haastateltavan vastauksiin sekä tarkentavien kysymyksien esittämiseen. Lisäksi osa kysymyksistä jätettiin pois, jos ne eivät olleet haastateltavalle relevantteja: esimerkiksi osa datan käyttäjistä ei käsitellyt tai yhdistellyt dataa työssään, joten helppokäyttöisyysteeman kysymykset rajattiin pois näistä haastatteluista.

Haastateltavat valittiin harkinnanvaraisella otannalla kohdeorganisaation sisältä niin, että haastatteluilla saataisiin luotua mahdollisimman kattava läpileikkaus tietoalueen datan laadusta. Haastateltavat muodostivat neljä eri ryhmää, joista kaikista haastateltiin datan parissa työskenteleviä asiantuntijoita tietovirran eri vaiheissa. Näin datan laadun ongelmista saatiin kerättyä tietoa kattavasti niin, että ryhmien sisällä vastaukset ovat vertailukelpoisia. Haastateltavat on esitelty ryhmiteltynä taulukossa 8. Haastateltavilta kysyttiin myös heidän rooliaan datan käsittelyssä: ovatko he datan tuottajia vai käyttäjiä tai onko heillä määritelty vastuu jostain ydintiedosta (tietovastaava) tai tietojärjestelmästä (sovellusvastaava). Osa sovellus- ja tietovastaavista kokivat olevansa sekä datan käyttäjiä että tuottajia, sillä he hyödynsivät dataa itse sekä olivat vastuussa sen tuotantoprosessista tai -järjestelmästä. Tällöin voitaisiin puhua myös Sebastian-Colemanin (2013) käyttämästä datan välittäjän roolista.

Taulukko 8. *Haastateltavat henkilöt ryhmittäin ja heidän roolinsa datan käsittelyssä*

Haastateltava	**Ryhmä**	**Rooli(t) datan näkökulmasta**
T1	Tasehallinta	Käyttäjä, tuottaja, sovellusvastaava, tietovastaava
T2		Käyttäjä, tuottaja, sovellusvastaava
T3		Käyttäjä
R1	Reservienhallinta	Käyttäjä, tuottaja, tietovastaava
R2		Tuottaja, tietovastaava
R3		Käyttäjä
V1	Verkonhallinta	Tuottaja, sovellusvastaava, tietovastaava
V2		Käyttäjä, sovellusvastaava
V3		Käyttäjä, tietovastaava
S1	Siirtojen hallinta	Käyttäjä
S2		Käyttäjä, tuottaja, tietovastaava
S3		Käyttäjä

Ennen haastatteluita haastateltaville toimitettiin sähköpostitse lyhyt kuvaus tutkimuksen tarkoituksesta ja toteutustavasta sekä haastattelussa käsiteltävistä teemoista. Viestin liitteenä oli kuva tarkasteltavasta tietoalueesta ydintietoineen (liite A), johon haastateltavia pyydettiin vielä perehtymään ennen haastattelua. Tämän tarkoituksena oli kertoa haastatteluteemoista jo ennakkoon uskottavuuden lisäämiseksi (Saunders et al. 2019, s. 452) sekä terävöittää tutkimuksen rajausta haastateltaville, jotta vastaukset käsittelisivät oikeaa dataa. Haastattelut toteutettiin koronapandemian aiheuttamien tapaamisrajoitusten

vuoksi videopuheluina Microsoft Teams-alustalla. Lyhyin haastattelu kesti noin 35 minuuttia, ja pisin haastattelu 65 minuuttia. Keskimääräinen haastattelun kesto oli noin 47 minuuttia. Haastattelut tallennettiin ja ne litteroitiin tallenteen pohjalta jälkikäteen, jotta haastattelun aikana haastattelija pystyi keskittymään kuuntelemiseen ja mahdollisten tarkennusten kysymiseen.

3.3 Aineiston analysointi

Haastatteluaineistoa analysointiin sisällönanalyysin menetelmillä, eli aineistosta poimittiin tutkimuskysymysten näkökulmasta olennaisia kohtia, jotka yhtenäistettiin koodaamalla ja teemoiteltiin, eli jaoteltiin tutkimuskysymyksen kannalta olennaisiin kokonaisuuksiin (Juhila 2021b). Analysointivaihe aloitettiin litteroimalla tallennetut haastattelut tekstidokumenttiin. Haastateltavien kommentit kirjoitettiin lähes sanasta sanaan jättäen pois ainoastaan pohdiskelevia täytesanoja. Haastattelut pyrittiin litteroimaan mahdollisimman pian niiden jälkeen, jotta mahdolliset haastattelurungon kehityskohteet tai lisätietoa vaativat kohdat saatiin huomioitua myöhemmissä haastatteluissa. Litteroitu aineisto luettiin muutaman kerran, ja tässä yhteydessä aineistosta lihavoitiin erikseen tutkimuskysymyksen kannalta merkittävät kohdat, kuten esimerkiksi maininnat työssä kohdatuista puutteista datan laadussa.

Tämän jälkeen merkityt kohdat siirrettiin taulukkoon, jonka vaakariveinä toimivat haastattelurungon laadun ulottuvuudet ja pystysarakkeina yksittäiset haastattelut. Tätä matriisia käytiin läpi riveittäin sarake kerrallaan, ja aineistosta nostetut havainnot koodattiin lyhyiksi lausekkeiksi. Jos havaintoa vastaava koodi oli jo olemassa, kasvatettiin sen havaintomäärää yhdellä. Jos havaintoa merkitykseltään vastaavaa koodilauseketta ei vielä ollut, luotiin sitä varten uusi koodi. Esimerkiksi erään haastateltavan kommentti integraatio-ongelmien aiheuttamasta katkoksesta arvojen päivittymisessä sekä toisen haastateltavan lausunto laskennoista puuttumaan jääneistä lähtötiedoista koodattiin molemmat tiedonsiirto-ongelmiksi.

Koodilausekkeita ja niiden mainintamääriä tarkastellessa niiden taustalta tunnistettiin kuusi erillistä teemaa, joihin koodatut havainnot ryhmiteltiin. Teemat muodostettiin aineistolähtöisesti niin, että haastateltavien mainitsemien ongelmien taustalta pyrittiin tunnistamaan yhteisiä taustatekijöitä ja juurisyitä. Esimerkiksi jotkut haastateltavat kertoivat datan saatavuuden heikentyneen, kun tietoa pitää etsiä useasta eri järjestelmästä. Osa taas kertoi datan maineen kärsivän alkuperäisen lähdejärjestelmän hämärtyessä tiedon kulkiessa usean järjestelmän läpi. Eri laatu-ulottuvuuksiin kohdistuvista vaikutuksista huolimatta molemmat näistä ongelmista ovat seurausta datan ja järjestelmien hajanaisuudesta, joten ne sijoitettiin osaksi kyseistä teemaa.

Teemojen pohjalta kirjoitettiin yhteenveto tuloksista, eli tämän työn luku 4. Teemapohjainen esittelytapa valittiin toiston välttämiseksi sekä olennaisten havaintojen korostamiseksi. Esimerkiksi eri käyttäjäryhmien tai tietoalueryhmien tulosten välillä ei havaittu tutkimuskysymyksen näkökulmasta merkittäviä eroja, joten erittelyä niiden perusteella ei pidetty mielekkäänä. Itse haastattelurunko oli muodostettu teorian esittelemien laatuulottuvuuksien pohjalta, mutta tulosten esittely niiden perusteella olisi tuonut raporttiin toistoa sekä jättänyt olennaisimmat havainnot vähemmälle huomiolle kuin teemoittain ryhmitelty esitystapa. Teemojen sisältöä pyrittiin avaamaan yhteenvedossa taulukoilla, kaavioilla sekä suorilla sitaateilla haastatteluista.

4. KÄYTTÖTOIMINNAN DATAN LAADUN NYKY-TILA

Haastatteluiden tulokset on jaoteltu tässä luvussa esiteltäviin teemoihin, jotka kuvaavat juurisyitä haastatteluissa ilmenneiden datan laatuun liittyvien haasteiden taustalla. Teemoilla ja luvun sisällöllä kokonaisuudessaan pyritään vastaamaan työn ensimmäiseen tutkimuskysymykseen *”Mitä ongelmia käyttötoiminnan datan laadussa on tällä hetkellä?”*. Teemat ja niiden sisältämät koodatut maininnat aineistossa on tiivistetty taulukkoon 9.

Taulukko 9. *Aineistosta nousseet teemat sisältöineen*

Teema	**Teeman sisältämät ongelmat**
Hajanainen järjestelmä-arkkitehtuuri	Data hajallaan eri järjestelmissä; Tiedonsiirto-ongelmat; Epäjohdonmukaisuus järjestelmien välillä; Datan löytäminen on vaikeaa; Tiedon lähde hämärtynyt; Järjestelmiä on liikaa; Sama tieto useassa paikassa; Käsittely on työlästä; Ei tietoa, missä dataa käytetään
Tietovaraston vajaa-käyttö	Tietovarastossa ei riittävästi tietoa; Tietovaraston olemassaolosta ei olla tietoisia; Tarvitaan jalostetumpaa tietoa; Tietovaraston hidas sykli; Visualisoinnissa haasteita
Valvonnan puute	Virheitä on vaikea havaita; Datan laatua ei valvota; Tarkkuudesta ei voi olla varma; Laatukriteerien puute; Datalähteiden laatua ei voi arvioida; Tiedonsiirtoa ei valvota;
Ennusteiden ongelmat	Ennusteiden maine on kärsinyt; Ennusteet epätarkkoja/-luotettavia; Ennusteet ei tuoreita; Ennusteet epäuskottavia; Ennusteiden tulee olla tulevaisuudessa parempia
Datanhallintamalli ja yhteiset käytännöt	Tietovastaavat ei tietoisia tarkoista vastuistaan; Tietoalueen ydintiedot ei ajan tasalla; Datajoukkojen nimeäminen epäyhteneväistä; Etumerkit epäyhteneväisiä
Toiminnanohjausjärjestelmä	Järjestelmän puutteet; Tietojen puutteet;

Jokaisen teeman aliluvussa esitellään taulukkomuodossa sen alle kuuluvat koodit, niiden mainintojen määrä sekä laatu-ulottuvuudet, joissa koodin alle kuuluvia asioita on mainittu haastatteluissa. Mainintoja-sarake kertoo asiasta maininneiden haastateltavien määrän, eli vaikka haastateltava olisi tuonut aiheen esiin useasti, tähän se lasketaan vain yhtenä. Maksimimäärä mainintoja on näin ollen 12. On myös huomattava, että mainintojen määrä ei välttämättä korreloi suoraan ongelman laajuuden tai merkityksellisyyden

kanssa – yksikin maininta voi kertoa suuresta vaikutuksesta yksittäiseen liiketoiminta-prosessiin, johon muut haastateltavat eivät osallistu, eivätkä siten huomaa myöskään puutetta datassa. Ulottuvuuksien kohdalla ”Muu” tarkoittaa, että haastateltava mainitsi asiasta varsinaisen ulottuvuuksien mukaan rakennetun kysymysrungon ulkopuolella eli esimerkiksi viimeiseen ”Tuleeko vielä mieleen jotain muuta mitä haluaisit sanoa?” -kysymykseen vastatessaan.

4.1 Hajautettu järjestelmäarkkitehtuuri

Jokainen haastateltava toi esiin haasteita, joiden taustalla voidaan katsoa olevan kohdeorganisaation tietojärjestelmien hajanaisuus. Haastateltavien näkökulmasta tämä näkyy esimerkiksi järjestelmien paljoutena, datan hajanaisuutena eri järjestelmissä, sekä tiedonsiirtokatkoksina eri järjestelmien välillä. Tarkemmin kaikki tämän teeman ongelmat oheistietoineen on esitelty taulukossa 10.

Taulukko 10. *Hajautetun järjestelmäarkkitehtuurin ilmeneminen aineistossa*

Ongelma	**Mainintoja**	**Ulottuvuudet**
Data hajallaan eri järjestelmissä	11	Saatavuus, Tiivis esitystapa, Helppokäyttöisyys
Tiedonsiirto-ongelmat	11	Saatavuus, Täydellisyys, Tarkkuus, Oikea-aikaisuus, Helppokäyttöisyys, Maine, Muu
Epäjohdonmukaisuus järjestelmien välillä	9	Johdonmukainen esitystapa, Helppokäyttöisyys
Datan löytäminen on vaikeaa	7	Saatavuus, Sopiva määrä, Muu
Tiedon lähde hämärtynyt	6	Maine
Järjestelmiä on liikaa	3	Sopiva määrä
Sama tieto useassa paikassa	3	Tarkkuus, Sopiva määrä
Käsittely on työlästä	3	Saatavuus
Ei tietoa, missä dataa käytetään	1	Ymmärrettävyys

Osa haastateltavista puhui ongelmasta ylätasolla: joko itse tietojärjestelmiä on liikaa tai datalähtöisemmin sanottuna data on hajautettuna liian moneen erilaiseen järjestelmään. Järjestelmien suuren määrän koetaan hidastavan työtä ja hankaloittavan tiedon käyttämistä. Yksi käytännön haaste on myös järjestelmien tietoturvavaatimuksista aiheutuvat hidasteet, sillä käyttäjät tarvitsevat oikeudet useaan eri järjestelmään ja oikeudet eivät pysy aina ajantasaisena.

> *”Ohjelmia on niin monia mitä pitää käyttää, niin siinä hypit sit ohjelmien välissä ihmettelemässä niitä juttuja.”* (S3)

> *”Meillä on liian monta järjestelmää. Meillä on varmaan 30 eri sovellusta mitä me käytetään.”* (T3)

Yhtä vaille kaikki haastateltavat totesivat datan olevan liian hajallaan eri järjestelmissä. Tämän koettiin hidastavan työskentelyä, kun jo pelkästään datan saaminen käyttöön vaatii aikaa ja työtä. Tällöin nopeat tarkastukset eivät ole mahdollisia. Myös datan käsittely ja yhdistely vaikeutuu, kun tietoja joudutaan hakemaan useasta erilaisesta järjestelmästä, joiden toimintaperiaatteet voivat olla erilaisia. Osa haastateltavista kuitenkin huomautti, että yrityksessä jokin aika sitten käyttöönotettu tietovarasto on helpottanut tilannetta jossain määrin.

> *”Se [data] on vähän hajallaan siellä ja täällä eri järjestelmissä ja muussa niin se on aika lailla sit järjestelmäkohtasta miten se on saatavilla”* (T1)

> *”Niitä lähdejärjestelmiä on tosi paljon, ja monta, niin niitten välillä se yhdistely ei kyllä oo helppoa ilman näitä data-alustoja”* (V1)

Datan ja järjestelmien hajanaisuus ilmenee myös muunlaisina ongelmina haastateltavien työssä. Ilman keskitettyä alustaa tietoa joudutaan siirtämään paljon eri järjestelmien välillä, joka voi johtaa alkuperäisen tietolähteen hämärtymiseen sekä saman tiedon päätymisen useaan eri järjestelmään, mikä voi hankaloittaa datan hyödyntämistä. Epäselvyydet tietojen alkuperässä voivat heikentää datan mainetta ja siten vähentää sen käyttöä. Saman tiedon löytyminen useasta eri paikasta voi aiheuttaa myös virheitä, sillä osa haastateltavista kertoi löytäneensä ristiriitaisia arvoja samalle suureelle eri järjestelmistä. Tiedoille ei ole ilmoitettu pääjärjestelmää tai muuta ohjeistusta siitä, mistä saatava tieto on luotettavin.

> *”Välillä tietysti lähdejärjestelmä rupee vähän hämärtyyn jos ne tulee kauheen monen mutkan kautta. … ehkä se vaatii kokemusta, että tietää mistä mikäkin data tulee. Sitten kun on muutaman vuoden näiden asioiden kanssa pyöriny, niin tietää että tää tulee tuolta ja tuo tuolta, mutta ei sitä järjestelmästä itsestään nää mistä se on tänne ponnahtanu”* (S1)

> *”Ainakin aiemmin oli paljonkin semmosia tilanteita, että vähän eri järjestelmässä saattaa olla eri arvot, niin joudut miettimään, että niin, missähän tää on oikein vai onko missään. Semmosen penkominen on sitten aika työllistävää.”* (R3)

Erityisen suuri ongelma kohdeorganisaatiossa on tiedonsiirrossa ilmenevät katkokset, joiden seurauksena tietoja ei löydy halutusta järjestelmästä lainkaan tai ne voivat olla

puutteellisia. Tiedonsiirto- tai integraatio-ongelmat nousivat esiin yhtä vaille kaikissa haastatteluissa, vaikka haastateltavat käyttivät kattavasti eri järjestelmiä ja ydintietoja. Tiedonsiirron katkeaminen näkyy käyttäjille useina käytännön ongelmina: mittaustiedot jäävät saamatta, tiedot täytyy aina tarkistaa päivittämättä jääneiden tai puuttuvien arvojen varalta ja laskennat sekä ennusteet tuottavat virheellistä tietoa lähtötietojen puuttuessa. Vaikka virheet huomattaisiin ajoissa, niiden korjaaminen vaatii paljon aikaa. Huomaamatta jääneet puutteet voivat puolestaan aiheuttaa ongelmia useassa eri prosessissa ja järjestelmässä virheiden kertautuessa. Tämä huolettaa osaa käyttäjistä ja heikentää datan mainetta.

> *”Melkein mistä tahansa datasta kun puhuu, niin jos siitä jonkun pitkän pätkän ottaa, niin kyl se pitää aina jollain kammalla käydä läpi, et onks siellä joku arvo jääny jumiin tai puuttuuks sieltä jotain pätkiä”* (T1)

> *”Enemmän ne datan laatuongelmat liittyy siihen, että se ei siirry sieltä lähteestä, en tiedä onko se lähteen vai tiedonsiirron hyvyydestä kiinni. Lähde kykenee tuottaan arvon, mutta sitä ei saada siirrettyä sinne paikkaan missä sitä käytetään”* (R1)

Järjestelmien hajanaisuus myös lisää työtä järjestelmiä päivittäessä tai uusia kehittäessä. Automaattisen tiedonsiirron toteuttaminen erilaisten järjestelmien välillä koetaan työlääksi, mutta se on myös välttämätöntä. Työmäärää kasvattaa erilaisen datan epäyhteneväinen muoto, jota voidaan joutua muokkaamaan paljonkin eri järjestelmien välillä. Lisäksi integraatioiden toteuttamisprosessi koetaan työlääksi erityisesti osallistujien suuren määrän takia: samassa toteutuksessa voi olla mukana esimerkiksi sovellusvastaava, jokin järjestelmän käyttäjätaho, sisäinen IT sekä ulkopuolinen toteuttajataho. Ongelma ei ole ainoastaan kohdeorganisaation sisäinen, sillä samoja tietoja ja järjestelmiä hyödynnetään myös yhteispohjoismaisissa toiminnoissa, joissa omat vaikutusmahdollisuudet esimerkiksi käytettäviin tiedostotyyppeihin ovat rajalliset.

Datan hajanaisuudella on myös suorempia seurauksia. Pirstaloituneisuudesta johtuen datan käyttäjien on vaikea löytää tarvitsemaansa tietoa, erityisesti jos kyseessä on harvemmin tarvittu aineisto. Dataa ylipäätään on paljon, ja sille ei ole olemassa tarkkaa dokumentaatiota, josta selviäisi tiettyjen tietojen sijainti. Näin ollen datan löytäminen vaatii kokemusta. Toisaalta datan suurta määrää pidettiin positiivisena asiana, ja kokemuksen todettiin auttavan datan löytämisessä.

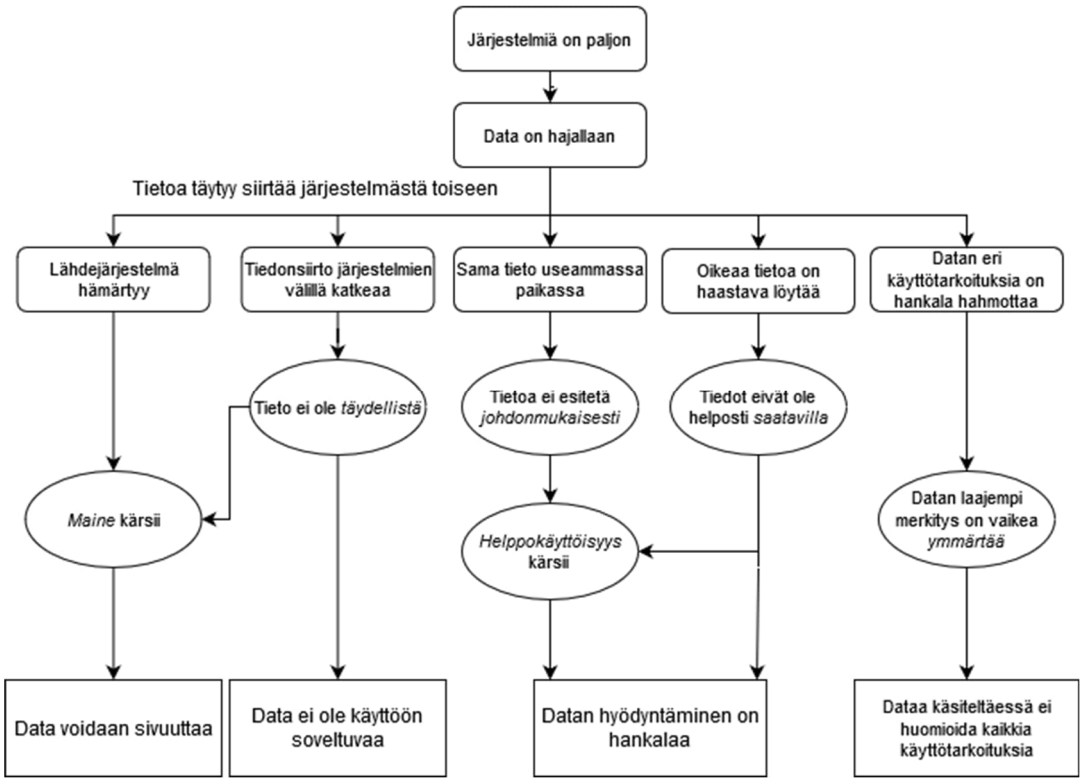
> *”Hyvin pitkälti pitää tietää, mitä on missäkin järjestelmässä, ei oo semmosia katalogeja mistä vois miettiä.”* (R2)

Tiedon löytämisen ohella myös sen yhdistäminen ja käsittely on hankalaa, kun tietoaineistojen etsinnän vaivalloisuuden lisäksi järjestelmät myös esittävät ja välittävät eteenpäin dataa eri muodoissa. Käytännössä eroja voi olla esimerkiksi tyhjien arvojen käsittelyssä sekä kellojen siirtojen vaikutuksessa aikaleimoihin. Tietoja käsittelevät haastateltavat hyödyntävät dataa yhdistellessään esimerkiksi Exceliä, sillä varsinaisten käyttötoiminnan omien tietojärjestelmien sisällä dataa on hankala siivota tarkoitukseen sopivaksi.

> *”Joutuu aina tarkisteleen miten vaikka kesäaikaan siirtyminen on käsitelty siinä, lisäileekö se sinne lisätunteja vai onko ne missä UTC-aikaleimoissa ne datat”* (R2)

Hajanaisuus voi näkyä myös epäsuoremmin datan käsittelyssä, sillä datan tuottajat ja käyttäjät eivät aina tiedä kaikkia sen käyttötarkoituksia. Osaa tiedoista julkaistaan muun muassa Fingridin avoin data -palvelussa, jolloin asiakkaat ja muut sidosryhmät voivat käyttää dataa omiin tarkoituksiinsa. Nämä käyttötavat eivät kuitenkaan ole välttämättä dataa hallinnoivien ihmisten tiedossa, joten mahdollisia niiden asettamia vaatimuksia datan laadulle ei voida huomioida.

Järjestelmien paljous ja datan hajanaisuus aiheuttavat epätietoisuutta, hankaloittavat datan hyödyntämistä sekä pakottavat siirtämään dataa useiden eri järjestelmien välillä, mikä aiheuttaa ongelmia tiedonsiirtokatkosten ollessa yleisiä. Haastatteluissa mainitut ongelmat, niiden keskinäiset seuraussuhteet sekä vaikutukset datan laadun eri ulottuvuuksiin sekä organisaation toimintaan on esitelty kaaviomuodossa kuvassa 7.



Kuva 7. *Hajanaisen arkkitehtuurin ongelmat ja vaikutukset*

Ongelmat siis ilmenevät datan parissa työskenteleville ihmisille eri tavoin, mutta juurisyy niiden taustalla on tietojärjestelmien suuri määrä, joka aiheuttaa datan siiloutumista. Tämä taas johtaa edelleen tietolähteiden hämärtymiseen, tietojen siirtelyyn, ylimääräisiin kopioihin sekä hankaloittaa kokonaisuuden hahmottamista.

4.2 Tietovaraston ja raportoinnin vajaakäyttö

Osa haastateltavista mainitsi potentiaalisena datan saatavuuden edistäjänä organisaation tietovaraston. Aineiston perusteella tietovarastoon liittyy kuitenkin haasteita, joiden takia sen käyttö on jäänyt vähäiseksi. Osaltaan ongelma näkyy myös kohdeorganisaation raportoinnissa ja tiedon visualisoinnissa, sillä nykyisen mallin mukaan raportteja ja visualisointeja tehdään tietovarastossa olevista tietoaineistoista. Tietovaraston ja raportoinnin vajaakäytön haasteet on eritelty taulukossa 11. Teeman kokonaismainintojen määrä on pienehkö, mutta kuten taulukosta huomataan, kaikki haastateltavat eivät olleet edes tietoisia tietovaraston kaltaisen keskitetyn ratkaisun olemassaolosta. Näin ollen osa haastateltavista ei ole myöskään osannut ottaa aihetta esiin haastatteluissa.

Taulukko 11. *Tietovaraston ja raportoinnin vajaakäytön ilmeneminen aineistossa*

Ongelma	**Mainintoja**	**Ulottuvuudet**
Tietovarastossa ei riittävästi tietoa	3	Helppokäyttöisyys, Muu
Tarvitaan jalostetumpaa tietoa	3	Oikea-aikaisuus, Sopiva määrä, Merkityksellisyys
Tietovaraston hidas sykli	2	Saatavuus, Oikea-aikaisuus
Visualisoinnissa haasteita	2	Tiivis esitystapa
Tietovaraston olemassaolosta ei olla tietoisia	1	Muu

Tietovarasto nousi aineistossa esiin potentiaalisena tulevaisuuden ratkaisuna datan hajanaisuuden aiheuttamiin ongelmiin. Nykyisessä tilanteessa se ei kuitenkaan ole ratkaisu, sillä tietovarastoon on siirretty hyvin vähän kohdeorganisaation kannalta relevantteja tietoja.

> *”Tietty meillä on sitä dataa paljon eri lähdejärjestelmissä, ja vaikka me nyt ollaan otettu tätä tietovarastoa käyttöön, niin sieltä ei varmaan vielä saa kuin pienen osuuden siitä datasta mitä meillä syntyy.”* (R3)

Syitä tietovaraston pienelle datamäärälle on useita. Tietovaraston kehitystiimi on nykyiselläänkin työllistetty, vaikka tietoja ei ole siirretty vielä suuria määriä. Kaksi haastatelta-

vaa toivat myös esiin, että nykyinen tietovaraston päivitystahti ei ole riittävä. Kuten aiemmin todettiin, kohdeorganisaation tietoalueen tiedot ovat suurelta osin reaaliajassa tai ainakin vuorokauden sisällä päivittyviä. Nykyinen tietovarasto sen sijaan päivittyy kerran vuorokaudessa, jolloin tiedot ovat jo vanhentuneita moneen käyttötarkoitukseen.

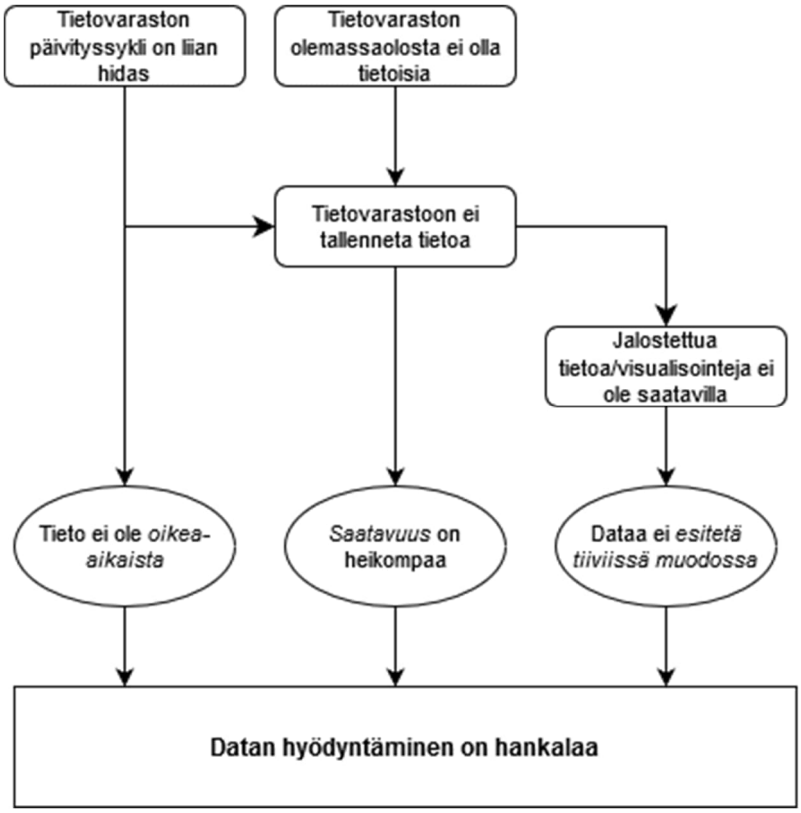
> *”Me ollaan vähän tuskailtu sen tietovaraston kanssa siinä, että sen oletuspäivitysväli on kerran vuorokaudessa, ja kun käynnissä oleviin keskeytyssuunnitelmiin tai kytkentäohjelmiin tehdään muutoksia, niin ois erittäin tärkeetä, että kaikilla ois se viimeisin versio käytössä, jotta ei tehdä päätöksiä väärän tiedon perusteella. Siinä ei riitä se vuorokauden päivitysväli oikein.”* (V1)

Oikea-aikaisuuteen liittyvien teknisten rajoitteiden lisäksi myös työntekijöiden tietämys tietovarastosta ja sen käytöstä saattaa olla vähäistä. Yksi haastateltava toi esiin kehitysehdotuksen keskitetystä data-alustasta, johon voitaisiin koota dataa yli yksikkörajojen. Toisen haastateltavan mukaan juuri tämä on tietovaraston tarkoitus. Haastateltavilta ei kuitenkaan suoraan kysytty heidän tietovaraston tuntemuksestaan, joten tarkkaa kuvaa tilanteesta ei voida näiden tietojen perusteella muodostaa. Joka tapauksessa tämä voi olla yksi syy tietovaraston vähäisen käytön taustalla.

Tietovaraston vajaakäyttö vaikuttaa paitsi datan yleiseen saatavuuteen, niin myös tiedon jalostamiseen visuaalisiksi raporteiksi. Fingridin datanhallintamallin mukaisesti jokaisella työntekijällä on pääsy Power BI -työkaluun, jota käytetään raporttien rakentamiseen tietovarastosta löytyvän datan pohjalta. Haastatteluissa nousi useasti esiin tarve jalostetummalle tiedolle päätöksenteon nopeuttamiseksi. Osa haastateltavista myös mainitsi erikseen Power BI -raportit mahdollisena ratkaisuna, mutta niiden tekemistä rajoittaa osaltaan myös tietovarastosta löytyvän datan vähäinen määrä.

> *”Se vois olla vähän jotenki jalostetumpaa se data sen puoleen, tai sitä vois enemmän jalostaa dataa jollain tavalla, esimerkiks jotain valmiita… esimerkiks Power BI:hin vois tehdä jotain valmiita kuvaajia tai piirtoja siitä datasta.”* (S2)

Kaiken kaikkiaan tietovarasto ei siis vielä toimi kunnolla keskitettynä data-alustana muun muassa henkilöstön heikon tietämyksen sekä päivitystahtiin liittyvien teknisten rajoitteiden takia. Tämä näkyy osaltaan myös tiedon jalostamisessa. Tietovaraston vajaakäyttöön liittyvät tekijät ja niiden vaikutukset on esitelty kuvassa 8.



Kuva 8. *Tietovaraston ja raportoinnin vajaakäytön vaikutukset*

Kuvasta voidaan havaita, että nykymuotoisena tietovarastosta saatava data on hankalasti hyödynnettävissä, sillä se on käyttöhetkellä jo vanhentunutta. Toisaalta ihmiset tuntevat tämän rajoitteen, eikä tietovarastoon juuri tallenneta nopeaa käsittelyä vaativaa dataa, jolloin se ei ole myöskään keskitetysti saatavilla. Lopulta myös datan esitystapa kärsii, kun raportteja ei pystytä valmistamaan tietovaraston vähäisen datamäärän takia.

4.3 Mittarien ja valvonnan puute

Valtaosa haastateltavista toi jossain muodossa esiin huolen siitä, että datassa olevia puutteita on vaikea havaita, ja ilmoitukset niistä tulevat pahimmassa tapauksessa arvoketjun loppupäästä asiakkailta tai muilta sidosryhmiltä. Kohdeorganisaatiossa ei ole käytössä laajamittaista valvontaa tai mittaristoa datan laadun seurantaan, kuten taulukko 12 kertoo.

Taulukko 12. *Mittarien ja valvonnan puutteiden ilmeneminen aineistossa*

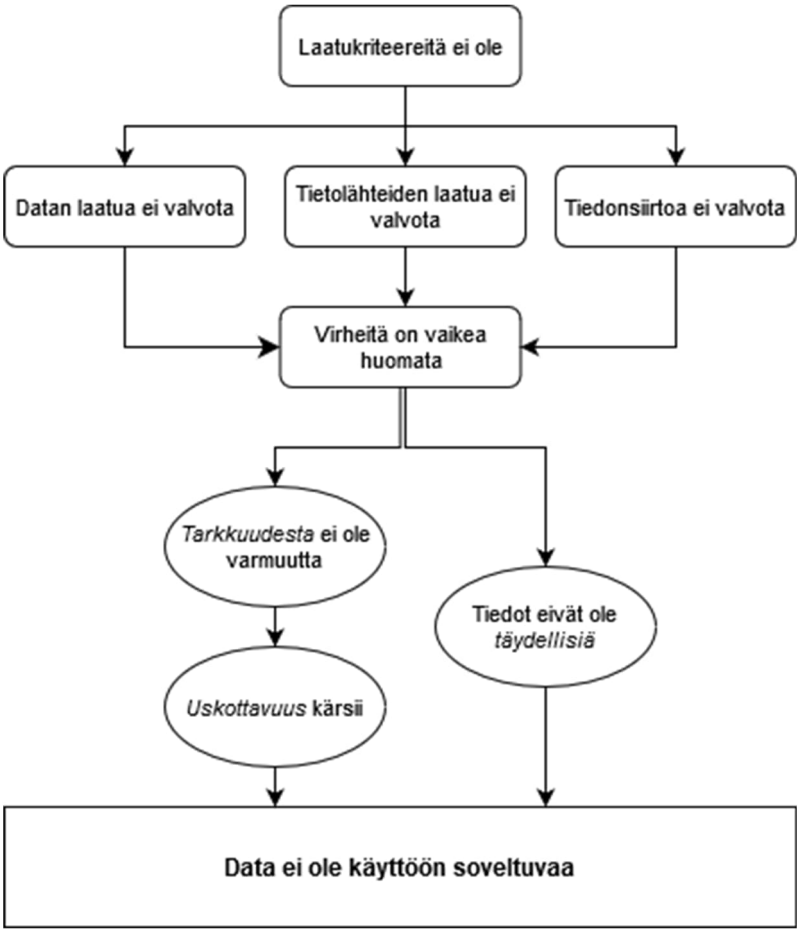
Ongelma	**Mainintoja**	**Ulottuvuudet**
Virheitä on vaikea havaita	5	Täydellisyys, Tarkkuus, Oikea-aikaisuus, Uskottavuus
Datan laatua ei valvota	3	Maine, Muu
Tarkkuudesta ei voi olla varma	3	Tarkkuus, Täydellisyys, Uskottavuus
Laatukriteerien puute	1	Tarkkuus
Datalähteiden laatua ei voi arvioida	1	Maine
Tiedonsiirtoa ei valvota	1	Täydellisyys

Yleisimmin ongelmat tulivat esiin haastateltavien puheissa virheiden havaitsemisen vaikeutena. Esimerkiksi laskennoissa käytettävien pohjatietojen virheet havaitaan, jos laskennan tulos vaikeuttaa poikkeukselliselta, ja tämän arvion joutuu tekemään asiantuntija itse. Näin ollen virheen havaitseminen on riippuvainen ihmisten kokemuksen ja tietämyksen määrästä. Pienet virheet tarkkuudessa voivat jäädä huomaamatta jopa vuosiksi, sillä niiden olemassaoloa ei voi päätellä suoraan tulosten oikeellisuudesta. Tämä voi johtaa ongelmiin, sillä organisaation päätöksenteko voi nojata vahvasti näihin laskelmiin ja virheelliset tulokset voivat kertautua, kun niitä käytetään jossain toisessa prosessissa lähtöarvoina.

> *”Esimerkkinä tossa vähän aikaa sitten kävi ilmi, että yhdessä laskennassa pari lähtöarvoa onkin laskettu siellä kahteen kertaan, vaikka pitäis olla vaan yhteen kertaan, ja sitten me ollaan sitä dataa käytetty mallien virittämiseen. Ja sitten huomataankin, että okei, täällä on tää data ollut muutaman vuoden ajan 40 % pielessä, niin semmonen on sitten vähän ikävä tilanne. Toki inhimillistä että näitä virheitä tulee, mutta se on niin kun hankala, että niitä on tosi vaikee sieltä huomata ja ne saatetaan huomata vasta tosi pitkän ajan päästä.”* (R3)

Yhtenä poikkeuksena nousivat esiin sähköasemilta tulevat mittaustiedot, joita hyödyntävä järjestelmä tarjoaa huomioita datan laadusta varoittamalla esimerkiksi raja-arvot ylittävistä tai pitkään päivittämättä olleista luvuista. Näiden mittaustietojen tarkkuudesta haastateltavat nostivat esiin valtaosin positiivisia huomioita. Joskus mittaukset voivat kuitenkin jäädä jumiin, eikä järjestelmä aina merkitse näitä erillisellä hälytyksellä. Tällöin päivittymättömän mittaustiedon havaitseminen jää operaattorin vastuulle, joten virheen havaitsemisessa voi kestää pitkään. Ajoittain järjestelmän antamat kriittiset varoitukset ovat myös virheellisiä, jolloin niiden selvittelyyn kuluu aikaa turhaan.

Muissa järjestelmissä virheiden heikko havaitseminen johtuu todennäköisesti automaattisen valvonnan puutteesta, kuten osa haastateltavista toi esiin. Itse datan laatua ei valvota suurimmassa osassa järjestelmiä, ja jos valvotaan, niin mahdollisten virheilmoitusten havaitseminen jää usein käyttäjän omalle vastuulle. Virheiden käsittelyyn ja laadun jatkuvaan kehittämiseen ei ole myöskään olemassa tiettyä prosessia, jonka mukaan toimia. Haastateltavat tunnistivat useita kohtia tietovirroissa, joissa valvonta voisi olla hyödyllistä: itse datan sisältämiä arvoja voitaisiin tarkastella järjestelmissä, tiedonsiirron onnistumista järjestelmien välillä voitaisiin valvoa sekä eri tietolähteiden luotettavuudesta voisi olla saatavilla tietoa niitä hyödynnettäessä. Tällä hetkellä datalle, tiedonsiirrolle tai lähdejärjestelmille ei kuitenkaan ole määritelty erillisiä laatukriteereitä, jotka olisivat yleisessä tiedossa ja mahdollistaisivat niiden toteutumisen seurannan. Datan luotettavuuden ja oikeellisuuden todentaminen jää siis asiantuntijoiden oman tietämyksen varaan.



Kuva 9. *Valvonnan ja mittarien puutteen vaikutukset*

Kuvassa 9 on esitelty valvonnan ja mittarien puutteen keskinäiset vaikutukset haastatteluaineistosta nousseisiin ongelmiin sekä niiden vaikutukset datan laatuun ja niiden kautta datan hyödyntämiseen. Keskeisin ongelma on virheiden havaitsemisen vaikeus, joka voi johtaa käyttökelvottomaan dataan. Syitä ongelman taustalla ovat valvonnan puute, mutta

valvontaa ei voida myöskään toteuttaa ilman yhdessä määriteltyjä laatukriteereitä tai -mittareita.

4.4 Ennustetietojen ongelmat

Aiemmista teemoista poiketen ennustetiedot erottuivat aineistosta datalle luontaisten laatu-ulottuvuuksien ongelmien muodossa. Ennusteet ovat myös luonteeltaan poikkeuksellisia tietoja, sillä niille ei ole suoranaista luontihetkellä tiedossa olevaa reaalimaailman vastinetta, vaan niiden tarkkuutta voidaan arvioida vasta jälkikäteen vertaamalla niitä todellisiin mitattuihin arvoihin. Ongelmat ennustetietojen tarkkuudessa eivät välttämättä siis ole suoranaisesti datan laadun ongelmia, vaan poikkeamat ennusteen ja myöhemmin mitatun arvon välillä voivat johtua ennusteen laskennasta. Aineistosta löytyi kuitenkin myös datalähtöisiä selityksiä ennusteiden poikkeamille. Ennusteita myös käytetään lähtötietoina muissa prosesseissa, joten niiden tarkasteleminen datan laadun näkökulmasta on perusteltua. Aineistossa ilmenneet ennustetietojen ongelmat on esitelty taulukossa 13.

Taulukko 13. *Ennustetietojen ongelmien ilmeneminen aineistossa*

Ongelma	**Mainintoja**	**Ulottuvuudet**
Ennusteiden maine on kärsinyt	5	Maine
Ennusteet epätarkkoja/-luotettavia	4	Täydellisyys, Tarkkuus
Ennusteet ei tuoreita	2	Oikea-aikaisuus
Ennusteet epäuskottavia	2	Uskottavuus
Ennusteiden tulee olla tulevaisuudessa parempia	2	Täydellisyys

Ennusteiden – erityisesti säästä riippuvaisen tuulivoimaennusteen - suurin ongelma on niiden hyödyntäjien heikko luottamus niihin. Tämä johtuu niiden historiallisesta epätarkkuudesta, joka on voinut johtua useasta syystä. Yksi taustatekijä on aliluvussa 4.1 mainitut tiedonsiirtokatkokset, joiden takia ennusteprosessin lähtötiedot jäävät puuttumaan ja varsinaista ennustetta ei saada tuotettua käyttäjien näkyviin lainkaan. Toisaalta ennusteita säädetään tuotantomittauksiin perustuvien laskelmien avulla: jos mittauksissa on jokin virheellinen tai päivittämättä jäänyt arvo, virhe siirtyy myös ennusteeseen. Tällaista poikkeamaa voi olla myös hyvin hankala havaita, kun yksittäisistä tuotantolaitoksista lasketaan koko maan summaa.

> *”Jos meillä tuotantomittauksissa on joku arvo jumissa, puuttuu joku laitos, niin se tavallaan näkyy heti siinä kulutusmittauksessa. Ja jos me kulutusmittauksen perusteella säädellään… tai kaikki ennustemallit ja muut säätelee itseään sen perusteella ja pitää sitä referenssiarvona… Jos se on vaikka 200 megaa pielessä, niin me ollaan sitten heti sen verran pielessä.”* (T1)

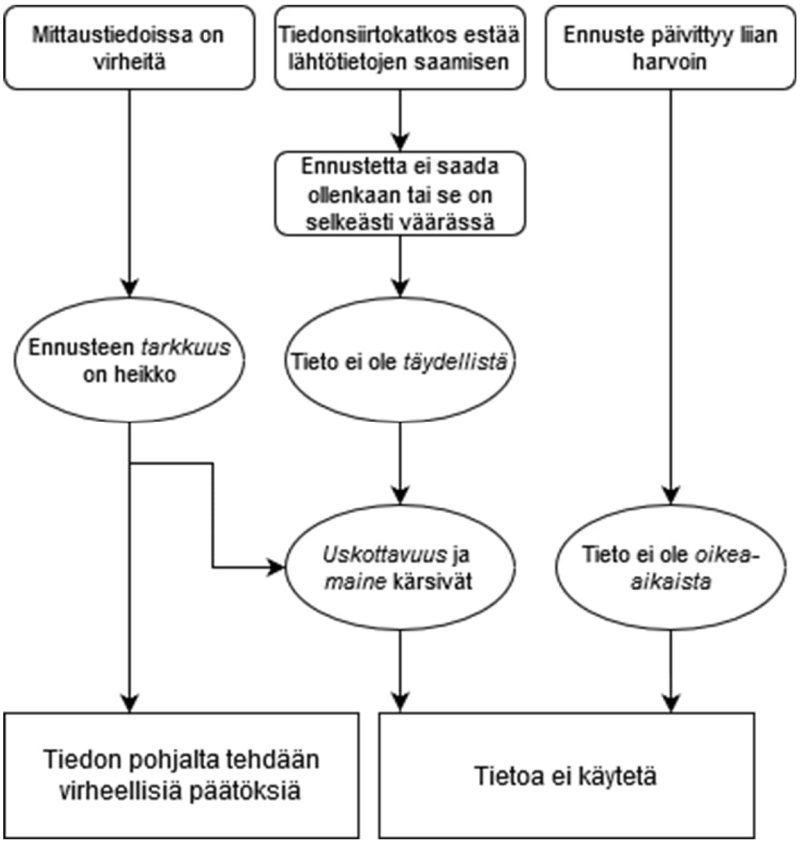
Tuulivoimaennusteen osalta on myös havaittu etumerkkivirheitä lähtötiedoissa, eli yksittäisen tuulipuiston tuotanto saattaa olla negatiivista, vaikka sen pitäisi olla positiivista. Mittauksissa saattaa myös esiintyä virheitä tuotannon ollessa nollassa niin, että mittaustieto jää jumiin johonkin nollaa lähellä olevaan arvoon, mikä sotkee ennustelaskelmat. Lisäksi tuulivoimaennusteen osalta epäiltiin, että mallissa ei olisi aina mukana tuoreimpia mittaustietoja.

Ennustetta tuottava järjestelmä tarjoaa myös varoituksia lähtötietojen laadusta, mutta tällä hetkellä niitä ei aktiivisesti seurata, vaan mahdolliset virheet havaitaan vasta laskentojen tulosten vääristyessä merkittävästi. Tällöin virheiden havaitseminen voi tapahtua hyvin suurella viiveellä, ja tiedot voivat olla jo merkittävästi virheellisiä.

> *”Meidän ennustejärjestelmässä on kyllä varotuksia, mutta niitä ei kyllä kukaan seuraa aktiivisesti tai oikeastaan ees puoliaktiivisesti, ei oo vaan mahdollisuuksia käyttää aikaa siihen että seurais niitä. Seurataan vaan sitä koko Suomen summaa ja kuinka realistinen se on, ja sit hommat menee jo tosi pieleen kun siellä rupee oleen huomattavaa virhettä. Et yksittäiset puistot, jos niissä on jotain virhettä, niin voi olla että ei huomata kuukausiin.”* (V2)

Suurin osa loppukäyttäjistä eivät osanneet eritellä syitä ennusteiden epätarkkuuden taustalla, joten on mahdollista, että osa ongelmista johtuu epävarmuustekijöistä ennusteen laskennassa puutteellisten lähtötietojen sijaan. Mainintoja ennusteiden ongelmista tuli ensisijaisesti käyttäjiltä, mutta myös tuulivoimaennusteen vastuuhenkilö oli tietoinen käyttäjien luottamuspulasta. Luotettavuusongelmien lisäksi osa loppukäyttäjistä toivoi, että ennusteet olisivat nykyistä ajantasaisempia. Varsinaista ongelmaa nykyisessä mallissa ei nähty, mutta ennustejärjestelmiltä toivottiin tiiviimpää päivittymissykliä.

Tarkemmin ennusteiden ongelmat datan laadun näkökulmasta on esitelty kuvassa 10. Nykytilan ongelmat voivat pahimmillaan johtaa ennusteiden täydelliseen sivuuttamiseen niiden käyttäjien päivittäisessä työssä, ja käyttäjät myös kertoivat tehneensä näin. Toisaalta ennusteita hyödynnetään lähtötietoina myös muiden ydintietojen tuottamisessa.



Kuva 10. *Ennusteiden ongelmat ja niiden vaikutukset*

Aineistosta nousi esiin myös, että ennusteiden tarkkuuteen kohdistuu kehityspainetta tulevaisuudessa alan kehityksen myötä. Esimerkiksi tuulivoiman määrän kasvaessa sama prosentuaalinen virhe tuulivoimaennusteessa aiheuttaa entistä suuremmat kustannukset. Lisäksi taseselvitysjakson lyheneminen tunnista 15 minuuttiin voi pakottaa muuttamaan toimintatapoja sekä tuottamaan ennusteita tiiviimmällä tahdilla. Näin ollen ennustetiedot kaipaavat kokonaisvaltaista kehitystyötä.

4.5 Datan hallinnointi ja yhteiset käytännöt

Vaikka haastateltavilta ei kysytty suoraan näkemyksiä datan hallinnasta tai hallinnoinnista, osa esiin tulleista ongelmista ja kehitysehdotuksista voidaan katsoa liittyvän nimenomaan datan hallinnointiin. Haastatteluissa hallinnolliset maininnat liittyivät yrityksen datanhallintamallissa määriteltyihin vastuisiin ja ydintietoihin sekä yhtenäisten käytäntöjen puutteeseen. Tarkemmin nämä neljä osa-aluetta on esitelty taulukossa 14.

Taulukko 14. *Datan hallinnointiin liittyvien ongelmien ilmeneminen aineistossa*

Ongelma	**Mainintoja**	**Ulottuvuudet**
Tietovastaavat eivät ole tietoisia tarkoista vastuistaan	4	-
Datajoukkojen nimeäminen epäyhteneväistä	3	Tarkkuus, Johdonmukainen esitystapa, Ymmärrettävyys, Muu
Etumerkit epäyhteneväisiä	2	Johdonmukainen esitystapa, Ymmärrettävyys
Tietoalueen ydintiedot eivät ole ajan tasalla	1	Muu

Haastateltavina oli useita datanhallintamallissa liiketoiminnan tietovastaaviksi määriteltyjä henkilöitä. Heillä oli kuitenkin hyvin vaihtelevat tiedot roolistaan, ja vain yksi osasi nimetä roolinsa ja vastuunsa tarkalleen. Valtaosa oli tietoinen nimellisestä roolistaan, mutta he eivät aina osanneet nimetä mistä ydintiedoista he ovat vastuussa, erityisesti jos niitä oli useita. Yksi haastatelluista tietovastaavista myös kommentoi, ettei oikeastaan muista mitä roolin sisältöön kuuluu, sillä hän ei ole tehnyt mitään asian eteen ydintietojen kuvausten kirjoittamisen jälkeen. Tilannetta on toisaalta ymmärrettävä, sillä datanhallintamalli on otettu käyttöön kohdeorganisaatiossa vasta viime vuonna.

Myös datanhallintamallin osana vuotta aikaisemmin määritellyn tietoalueen sisältöä pohdittiin vanhentuneeksi. Kehityshankkeet ovat edenneet, ja niissä hyödynnetään uusia tietoja, joille ei nyt ole määritelty vastuuhenkilöitä. Näiden ennustaminen ei ole ollut mahdollista, mutta tulevaisuutta varten voisi olla järkevää määritellä uusia ydintietoja hankkeiden edetessä.

Varsinaisen datanhallintamallin ohella havaittiin, että tietoaineistojen nimeämisessä ei ole kattavasti käytössä yhtenäisiä käytäntöjä. Tämä hankaloittaa aineistojen löytämistä, hyödyntämistä ja yhdistelyä, kun vähän tietoja käyttävä henkilö ei tunnista tietoja epäyhteneväisten nimien takia. Erityisen ongelmallisina koettiin tuulivoimatuotantoon liittyvät tiedot.

> *”Ehkä toi tuulivoima tulee vastaan, kun se muuttuu jatkuvasti, niin ei oo mitään yhtenäistä nimeämiskäytäntöä, vaikka millä tuotantojärjestelmät olis eri järjestelmissä samalla nimellä, vaan joutuu ristiintsekkaamaan ja ettimään, että mikähän tää nyt mahtaa olla verkkomallissa, kun ennustejärjestelmässä se oli tällanen.”* (S1)

> *”Aikasarjoilla on hyvin erilaisia nimeämistapoja yhden järjestelmän sisällä ja sitten kun se viedään toiseen järjestelmään niin se saattaa olla siellä sillä alkuperäsen kaltasella nimellä tai eri nimellä”* (R2)

Samanlainen ongelma on myös joidenkin mittaustietojen etumerkeissä – samantyyppiset mittaustiedot voivat olla positiivisia tai negatiivisia ilman selkeää logiikkaa. Joissain tapauksissa taas etumerkin määräytymiseen on sääntö, mutta sitä ei ole selkeästi dokumentoitu ja tietoa harvemmin käyttävät henkilöt joutuvat tarkistamaan asian erikseen.

> *”Jos ottaa järjestelmästä Olkiluodon pätötehon niin se on negatiivinen, mutta jos ottaa Vuosaaren kaasarin pätötehon niin se on positiivinen, eli eri etumerkkisääntö.”* (R3)

> *”Esimerkiks joku rajajohtosiirto Suomen ja Ruotsin välillä, niin se on joko plus- tai miinusmerkkistä se data, niin sä et oikein tiedä, että jos on miinusmerkkistä siirtoa, niin onko se tuontia Suomeen vai vientiä Ruotsiin jos ei sitä oo jossain kerrottu.”* (S2)

4.6 Verkonhallinnan toiminnanohjausjärjestelmä

Muut teemat ovat koskettaneet kaikkia tietoalueen tietoryhmiä, mutta verkonhallinnan tietoryhmän haastatteluista löytyi myös lähinnä kyseisen ryhmän tietoja ja niiden käyttäjiä koskettavia ongelmia. Verkonhallinnassa hyödynnettävä toiminnanohjausjärjestelmä koettiin hankalaksi, sekä sen sisältämät verkko-omaisuuteen liittyvät tiedot paikoitellen puutteellisiksi. Nämä tiedot eivät itsessään kuulu käyttö- ja tilatiedon tietoalueeseen, mutta niitä käytetään lähtötietoina useassa liiketoimintaprosessissa. Tästä syystä ongelmat saivat hajanaisia mainintoja myös muiden tietoryhmien haastateltavilta. Mainittujen ongelmien mainintojen määrä ja vaikutuksen kohteena olevat laatu-ulottuvuudet on eritelty taulukossa 15.

Taulukko 15. *Toiminnanohjausjärjestelmän maininnat aineistossa*

Ongelma	**Mainintoja**	**Laatu-ulottuvuudet**
Järjestelmän heikko käytettävyys	4	Saatavuus, Täydellisyys, Tarkkuus, Tiivis esitystapa, Johdonmukainen esitystapa, Helppokäyttöisyys
Puute omaisuustiedoissa	3	Täydellisyys, Tarkkuus, Oikea-aikaisuus, Uskottavuus, Muu

Toiminnanohjausjärjestelmän heikko käytettävyys vaikuttaa myös datan saatavuuteen. Järjestelmää pidetään sekavana erityisesti vähemmän sitä käyttävien haastateltavien

toimesta. Järjestelmässä on hyvin laajasti dataa, mutta yksittäinen käyttäjä tarvitsee tehtävissään vain pientä osaa tästä massasta, jolloin oikeiden tietojen etsiminen voi olla työlästä. Myös tietojen esitystapaa pidetään epäloogisena, jolloin tietojen tehokas hakeminen vaatii kokemusta järjestelmän käytöstä. Yksi käytännön esimerkki epäloogisuudesta on erilaisten työmääräysten (engl. *work order*) yhteinen numeroavaruus, johon sisältyy lukuisia erilaisia tietotyyppejä, kuten kytkentäohjelmat, siirtokeskeytykset ja häiriöt. Käyttäjä ei siis pysty tunnistamaan tiedon tyyppiä nopealla vilkaisulla, ja järjestelmä esittää erilaisia tyyppejä sekaisin samassa paikassa. Toisaalta valtaosa järjestelmän käyttäjistä työskentelevät sen parissa jatkuvasti, jolloin nämä ongelmat eivät välttämättä vaikuta toimintaan merkittävästi. Kokeneemmat käyttäjät myös pitivät suurta datamäärää hyvänä asiana. Yleisen sekavuuden ohella järjestelmää kritisoitiin myös sen hitaudesta, mikä luonnollisesti hidastaa myös datan hyödyntämistä.

> *”Järjestelmä nyt ei oo kaikista loogisin ja vikkelin ja nopein, mutta siellä on paljon tietoa ja kyllä ne sieltä löytyy kunhan osaa vaan etsiä oikeesta paikasta.”* (V3)

Käytettävyysongelmien ohella järjestelmässä on puutteita, jotka voivat johtaa virheisiin sen sisältämissä verkonhallinnan tiedoissa. Ongelmia aiheuttaa esimerkiksi kenttien pakollisuuden määrittäminen tietoja syötettäessä: eri prosesseissa kaikkia niiden kannalta olennaisia kenttiä ei ole määritelty pakollisiksi, jolloin tietoja jää syöttämättä. Jotkut kentät ovat myös sisällöltään tulkinnanvaraisia, mikä aiheuttaa myös ongelmia tiedon syöttäjien tulkitessa niitä eri tavoin. Vajaaksi jääneet kentät voivat aiheuttaa ongelmia, sillä kenttien jäädessä puutteellisiksi jokin tietoa hyödyntävä prosessi voi tuottaa virheellistä tietoa. Yhtenä konkreettisena esimerkkinä toimii siirtokeskeytyssuunnitelma-ydintieto, johon käyttäjät merkitsevät keskeytyksen alku- ja loppumisajankohdan. Tätä varten järjestelmässä on erillinen rivityyppi, jossa ajankohdat ovat pakollisia. Käyttäjä voi kuitenkin valita myös yleismaailmallisemman rivityypin, jossa näitä kenttiä ei ole pakko täyttää. Tällöin ne voivat jäädä tyhjäksi, ja tiedot keskeytyksestä eivät siirry automaattisesti päivittyvään verkkomalliin.

Järjestelmän haasteiden lisäksi sen sisältämä tieto on osin puutteellista erityisesti verkko-omaisuustietojen osalta. Toiminnanohjausjärjestelmässä on paljon vanhentuneita henkilötietoja, ja niiden päivittäminen on paikoitellen hankalaa, sillä tunnisteena käytetään puhelinnumeroa. Tällöin yhteyshenkilön vaihtuessa kumppaniyrityksessä uutta henkilötietoa ei välttämättä pystytä luomaan, sillä puhelinnumero on edellisen henkilön käytössä, mikäli sitä ei ole erikseen huomattu poistaa. Lisäksi varsinaista verkko-omaisuutta koskevat tiedot voivat olla paikoitellen puutteellisia esimerkiksi sähköasemien, kytkinlaitteiden ja voimajohtojen osalta. Erityisesti uusien kytkinlaitteiden kohdalla tiedot voivat olla aluksi puutteelliset, mutta tämä on harvinaista, ja tilanne on yleensä

korjattu nopeasti. Osa voimajohtojen tiedoista sen sijaan on tälläkin hetkellä puutteellisia tai arvoiltaan virheellisiä ainakin kuormitettavuuksien osalta. Tämä on ongelmallista, sillä puuttuvat tai virheelliset kuormitettavuudet voivat sekoittaa esimerkiksi verkkomalleja tai muuta käyttövarmuuslaskentaa.

5. KÄYTTÖTOIMINNAN DATAN LAADUN KEHITTÄMINEN

Työn toiseen tutkimuskysymykseen ”*Miten käyttötoiminnan datan laatua voidaan kehittää?*” vastataan tässä luvussa peilaamalla haastatteluaineistosta saatuja havaintoja luvussa 2 esiteltyyn kirjallisuuteen. Ensin tuloksissa tunnistettuja ongelmia analysoidaan ja vertaillaan niitä aiemmissa tutkimuksissa havaittuihin ongelmiin ja kehitystoimenpiteisiin. Tämän jälkeen pohditaan kohdeorganisaation datan laadun kypsyyttä hyödyntäen kirjallisuudessa esiteltyjä kypsyystasoja. Kypsyystasojen avulla priorisoidaan kohdeorganisaation datan laadun kehityksen kohteita. Lopuksi kohdeorganisaatiolle esitetään näiden pohdintojen pohjalta toimenpide-ehdotuksia datan laatuun liittyvien ongelmien korjaamiseksi.

5.1 Havaittujen ongelmien analyysi

Kokonaisuutena kohdeorganisaation datan laadun nykytila on kelvollisella tasolla, ja esimerkiksi toimenpiteitä vaativia tarkkuusongelmia ei juurikaan havaittu haastatteluissa. Haastateltavat myös tiesivät melko hyvin mistä heidän havaitsemansa datan laadun ongelmat johtuivat – moni osasi esimerkiksi kertoa, että puuttuvat tai selkeästi virheelliset arvot syntyvät usein tiedonsiirtokatkoksen seurauksena – minkä ansiosta jo tulosten teemoitteluvaiheessa pystyttiin pohtimaan laatuongelmien juurisyitä. Kaikkiaan ongelmat ovat myös melko ainutlaatuisesta toimintaympäristöstä huolimatta varsin tavallisia, sillä suurinta osaa niistä on havaittu myös aiemmin eri aloilla toimivien yritysten datan laatuun liittyviä haasteita käsittelevässä kirjallisuudessa. Seuraavissa alaluvuissa pohditaan tarkemmin merkittävimpiä aineistosta nousseita ongelmateemoja ja pohditaan alustavasti kehitystoimenpiteitä kirjallisuutta hyödyntäen.

5.1.1 Datan ja järjestelmien hajanaisuus

Haastattelussa eniten mainintoja keräsivät ongelmat, joiden tulkittiin olevan seurausta hajanaisesta järjestelmäarkkitehtuurista. Eri tietojärjestelmiä on useita, ja niissä käsitellään usein samaa dataa. Tämä johtaa muun muassa tiedon löytämisen hankaloitumiseen, tiedonsiirtokatkoksiin sekä epätietoisuuteen datan luotettavuudesta ja käyttötarkoituksista. Strong et al. (1997) mainitsevat vastaavanlaisia ongelmia: useat lähteet samalle datalle aiheuttavat ristiriitoja, ja hajanaiset järjestelmät vaativat datan integrointi- ja yhdistelytyötä. Ristiriitainen tieto voidaan sivuuttaa kokonaan, ja epäyhteneväinen esitystapa hankaloittaa tietojen hyödyntämistä. Umar et al. (1999) toteavat järjestelmien

välisen epäjohdonmukaisuuden olevan korkean prioriteetin ongelma, sillä tietoa voi joko puuttua vaadituista järjestelmistä tai sitä ei esitetä yhtenäisessä muodossa.

Datan hajanaisuus voi myös johtaa datan siiloutumiseen ja paikallisiin korjaustoimenpiteisiin datan laadun kustannuksella. Silvola et al. (2011) mainitsevat myös ongelmina datan siiloutumisen sekä epäyhteneväisen muodon, mikä hidastaa datan yhdistämistä. Haastatteluista saadut tulokset vaikuttavat olevan linjassa näiden havaintojen kanssa: myös tässä tapauksessa samalle asialle löytyi erilaisia arvoja useammasta paikasta, mikä hämmensi tiedon käyttäjiä. Lisäksi hajanaisuudesta aiheutuva tiedon löytämisen, keräämisen ja yhdistelyn työläys koettiin ongelmalliseksi erityisesti harvemmin käytettyjen järjestelmien kohdalla, sillä datan löytäminen ja käsittely vaativat tällöin myös järjestelmän käytön opettelua. Myös Umar et al. (1999) toteavat erilaiset järjestelmät ja niiden tuntemattomuuden ongelmalliseksi. Kohdeorganisaation tapauksessa järjestelmien ja datalähteiden määrä voi hankaloittaa myös datan laadun ylläpitämistä, sillä datalle ei ole mitään yksittäistä järjestelmää, jossa sen laatua voitaisiin valvoa ja puutteita korjata. Jos laatutarkastelua ja kehitystoimenpiteitä halutaan toteuttaa mahdollisimman lähellä datan tuotantojärjestelmää, joudutaan työ tekemään useaan kertaan.

Hajanaisen datan vaatimat integraatiot nousivat haastatteluissa esiin suurena ongelmien lähteenä käytetyistä järjestelmistä tai tietoryhmästä riippumatta. Puuttuvat tiedot tekevät datasta käyttöön soveltumatonta, ja voivat pahimmassa tapauksessa vääristää esimerkiksi laskentojen tuloksia datavirran loppupäässä. Umar et al. (1999) mainitsevat datan virtaamisen järjestelmien läpi yhtenä syynä ristiriitaiselle datalle. Ratkaisuksi ehdotetaan ongelman huomioimista arkkitehtuurin suunnittelussa: puurakenteen suosiminen, datan virtaaminen korkeamman tarkkuuden järjestelmästä pienemmän tarkkuuden järjestelmään sekä datan oikeellisuuden tarkistaminen aina sen siirtyessä järjestelmästä toiseen. Myös Silvola et al. (2011) tutkimuksen kahdeksasta yrityksestä kuusi mainitsee integraatioiden tai tiedonsiirron aiheuttavan ongelmia, joiden ratkaisuksi esitetään järjestelmien ja niiden välisten integraatioiden määrän minimointia, datamallin yhtenäistämistä integraatioissa tehtävien muutosten vähentämiseksi sekä tietojen tuotanto-, käyttö- ja ylläpitoprosessien mallintamista.

Hajanaisuusongelmat kytkeytyvät toiseen tuloksissa havaittuun ongelmaan, eli tietovaraston vajaakäyttöön. Organisaatiolla olisi mahdollisuus hyödyntää laajemmin yhteistä tietovarastoa, joka voisi vähentää hajanaisuuden aiheuttamia ongelmia ja tiedon siirtelyä järjestelmien välillä, mutta haastatteluiden perusteella käyttäjät eivät ole joko tietoisia tästä mahdollisuudesta tai tietovaraston tekniset rajoitteet estävät datan tehokkaan hyödyntämisen. Kuten luvussa 2.3.1 todettiin, ydintiedon hallinnan teoriassa keskeisenä ajatuksena on tällainen yksi ydintiedon sijoituspaikka. Myös kirjallisuus esittää ratkaisuksi

tietovarastoa: Umar et al. (1999) toteavat tietovaraston auttavan datan yhtenäistämisessä ja standardisoinnissa, mutta huomauttavat myös, että se ei voi korvata operatiivisia tietojärjestelmiä hitautensa takia. Strong et al. (1997) mukaan tietovarasto voi helpottaa datan saatavuutta, sillä sen kautta käyttäjät voivat saada pienemmän määrän heille olennaista dataa. Myös Lee et al. (2006) mainitsevat tietovarastot suosittuna ratkaisuna hajautettujen järjestelmien ongelmiin, sillä ne parantavat tiedon saatavuutta tarjoamalla keskitetyn käyttöliittymän muuten hajallaan oleviin tietoihin.

Kohdeorganisaation tapauksessa tietovaraston laajamittaisempi hyödyntäminen voisi vauhdittaa myös itsepalveluanalytiikan käyttöönottoa, sillä nykyisen datanhallintamallin mukaisesti työntekijöiden hyödynnettävissä oleva Power BI -analytiikkaohjelmisto ja sillä tuotetut visualisoinnit sekä raportit nojaavat vahvasti tietovaraston dataan. Haastatteluiden perusteella raporteille olisi kysyntää, mutta niitä ei vielä juurikaan hyödynnetä. Laajamittaisempi visuaalisten raporttien hyödyntäminen voisi tarjota dataa tiiviimmässä ja helpommin ymmärrettävässä muodossa, mikä edistäisi tiedon hyödyntämistä päätöksenteon tukena.

Silvola et al. (2011) huomauttavat, että käytännössä datan keskittäminen yhteen järjestelmään ei ole yleensä realistinen ratkaisu integraatioiden vaatimien resurssien takia, ja ehdottavat ratkaisuksi myös dataprosessien mallintamista. Tällainen mallinnus voisi edistää datan saatavuutta myös kohdeorganisaatiossa, sillä useat haastateltavat korostivat oikean datan löytämisen vaikeutta ja totesivat sen perustuvan lähinnä kokemuksen myötä karttuvaan hiljaiseen tietoon. Myöskään datan alkuperäinen lähde ja sen mahdollinen käsittely tietovirran aikana eivät olleet aina haastateltavien tiedossa, mikä heikentää datan mainetta ja voi johtaa datan sivuuttamiseen.

5.1.2 Datan laadun valvonta

Kohdeorganisaatiossa ei haastatteluiden perusteella aktiivisesti valvota datan laatua, vaan virheet huomataan usein viiveellä ja joissain tapauksissa vasta ulkoisen sidosryhmän ilmoitettua asiasta. Tämä ei ole itsessään puute datan laadussa, mutta se voi vaikuttaa datan hyödyntämiseen liiketoiminnassa merkittävästi ongelmien kertautuessa niiden jäädessä huomaamatta. Asiakkaille tai muille sidosryhmille asti päätyvät virheet voivat vaikuttaa myös koko yrityksen maineeseen, joten kyseessä on korkean prioriteetin ratkaistava ongelma. Virheellisen tiedon havaitsemisen ohella valvonta voisi lisätä henkilöstön luottamusta käytettävään dataan, sillä tällä hetkellä dataa voidaan päätyä hyödyntämään ilman tarkkaa tietoa datajoukon laadusta, mikä huoletti haastateltuja datan käyttäjiä.

Aiemmin esitetyillä datan laadun aktiivisuustasoilla (Silvola et al. 2011) kohdeorganisaatio vaikuttaa sijoittuvan passiivisen ja reaktiivisen tason väliin: datan laatua ei valvota järjestelmällisesti, vaikka osa järjestelmistä tarjoaa siihen apuvälineitä hälytysten muodossa. Kuten alkuperäisen tutkimuksen kohdeyritykset, myös tässä tapauksessa kohdeorganisaatio nousee ajoittain reaktiiviselle tasolle korjaamaan havaittuja ongelmia, kunnes palautuu taas passiiviseen tilaan. Korkeammalle aktiivisuustasolle vaatisi ensin datan laadun reaaliaikaista seurantaa (aktiivinen taso) ja ideaalitilanteessa laatuongelmien ennaltaehkäisyä (proaktiivinen taso) (Silvola et al. 2011). Vallitsevan ymmärryksen mukaan proaktiivinen laatuongelmien ehkäisy on lähestymistapana parempi kuin reaktiivinen ongelmien korjaaminen niiden ilmaantuessa (Mahanti 2019, s. 319–321; Allen & Cervo 2015; Silvola et al. 2011; Redman 2008 s. 55–85). Toisaalta kohdeorganisaation datassa on erityistä sen nopea muuntautuminen: esimerkiksi kantaverkosta sekuntien välein mitattavat arvot vaihtelevat jatkuvasti, jolloin ongelmiin joudutaan puuttumaan myös reaktiivisesti. Tällöin paras vaihtoehto voi olla proaktiivisia ja reaktiivisia menetelmiä yhdistelevä hybridilähestymistapa (Mahanti 2019 s. 321).

Myös haastateltavat toivat selkeästi esiin toiveen datan laadun valvonnasta ja mittaamisesta. Tällä hetkellä osalle datasta ei ole määritelty mitään kriteeristöä, jolla sen laatua voitaisiin arvioida. Mittaamisen puute on myös kirjallisuudessa tunnistettu ongelma: Silvola et al. tutkimien yritysten ohella myös Umar et al. (1999) tutkimuksessa mittaamisen tarve nousi korkean prioriteetin ongelmaksi, jonka ratkaisuksi esitetään useita erilaisia mittareita suoraan itse datan, sen alustojen tai dataprosessien seurantaan. Myös Haug et al. (2013) tutkimuksen kohdeyrityksistä valtaosa kokee mittaamisen puutteen ongelmalliseksi.

Datan laadun mittareissa ja niiden suunnittelussa voidaan hyödyntää laadun ulottuvuuksia. Vaikka haastatteluiden kysymykset olivat luokiteltu etukäteen ulottuvuuksittain, tuloksista ei voida suoraan laskea ulottuvuuksien sisältämien ongelmien määrää, sillä haastateltavat sijoittivat kohtaamiaan samanlaisia ongelmia eri ulottuvuuksiin. Esimerkiksi tiedonsiirtokatkoksen seurauksena yhdessä järjestelmässä voi näkyä tyhjä arvo, kun taas toisessa tapauksessa tuo puuttuva arvo ei näy suoraan, mutta se muuntaa laskennan tuloksena saadun luvun selkeästi virheelliseksi. Vaikka näiden laatuongelmien syy on samassa puuttuvassa arvossa, ensimmäinen tapaus näkyy käyttäjälleen puutteena täydellisyydessä, kun taas toinen tarkkuudessa. Osa haastateltavista myös mainitsi samat ongelmat usean ulottuvuuden kohdalla joko sellaisenaan tai hieman eri näkökulmasta. Kysymysrungossa hyödynnetyn AIMQ-menetelmän ulottuvuuslistaus ohjaa myös vahvasti ongelmien luokittelua ulottuvuuksiin - esimerkiksi tiedonsiirtokatkoksien

ongelmat voitaisiin laskea myös oikeellisuuden tai eheyden ongelmiksi, jos luokitteluun käytettäisiin Sebastian-Colemanin (2013) DQAF-mallin ulottuvuuksia.

Yksi perinteisimmistä ja konkreettisimmista mitattavista ulottuvuuksista on tarkkuus, eli datan korreloiminen sen kuvaaman todellisen maailman ilmiön kanssa. Tutkimuksessa ei kuitenkaan havaittu juurikaan suoranaisia tarkkuusongelmia. Toisaalta osa haastateltavista toi esiin huolen siitä, että mittaustietojen tarkkuutta ei ole varmistettu. Voimajärjestelmien mittauksien tarkkuuden arviointi voi myös olla haastavaa, sillä mittauksia ei voi suoraan verrata mihinkään ennalta oikeaksi tiedettyyn arvoon. Tarkkuuden sijaan suurimmat ulottuvuuksin kuvattavat aineistosta esiin nousseet ongelmat liittyvät saatavuuteen ja täydellisyyteen, joten myös laadun valvonta- ja mittaustoimenpiteissä voisi olla kannattavaa keskittyä näihin.

Valvonnan, mittarien ja prosessien ohella laadun seurannassa on olennaisessa osassa tiedon omistajat, vastaavat ja muut mahdolliset vastuuhenkilöt. Kirjallisuudessa on tunnistettu puutteellisen vastuunjaon olevan merkittävä ongelma datan laadun hallinnassa (katso esimerkiksi Haug et al. 2013, Silvola et al. 2011, Umar et al. 1999). Kohdeorganisaatiossa on tunnistettu ydintiedot ja nimitetty niille tietovastaavat, mutta tietovastaavien aktiivisuus ja tietämys roolistaan on haastatteluiden perusteella hyvin vaihtelevaa. Olennaisilla tiedoilla on siis nimetty vastuuhenkilö, mutta tämä ei vielä tarkoita, että vastuuhenkilöt proaktiivisesti valvoisivat datan laatua ja edistäisivät sen hyödyntämistä läpi organisaation. Toisaalta heille ei ole myöskään annettu työkaluja valvontaan, sillä virallisia, yhdessä sovittuja laatukriteereitä tai -tavoitteita ei ole kehitetty eikä teknisiä ratkaisuja datan laadun seurantaan juuri ole.

Nykyinen datanhallintamalli on uusi, joten on luonnollista, että sen jalkautus on vielä kesken myös vastuuhenkilöiden toimenkuvan osalta. Mallin ollessa alkuvaiheessa myös lopulliset ydintiedot voivat vielä elää: tutkimuksen aikana niitä päivitettiin jo kertaalleen, ja haastatteluissa nousi esiin toive vielä uudelle täydennykselle. Organisaatiossa on käynnissä useita tietointensiivisiä kehityshankkeita, joten tilanteen eläminen on luonnollista – toisaalta ydintietojen jatkuva päivittäminen ja niiden vastuuttaminen uusille tietovastaaville ei todennäköisesti helpota datanhallintamallin jalkauttamista tai tietojen hallintaa ylipäätään. Toisaalta käyttö- ja tilatieto ovat liiketoimintakriittisyydestä huolimatta tyypiltään perinteisistä ydintiedoista poikkeavia muun muassa nopean uusiutuvuutensa takia, joten ydintiedon hallinnan menetelmät eivät välttämättä sovellu niiden hallintaan täydellisesti. Tämä on kohdeorganisaation sisälläkin uniikki ongelma, sillä muut tietoalueet sisältävät enimmäkseen hitaammin muuttuvaa asiakas- ja laitetietoa.

Datan hallinta ei myöskään ole sisältänyt tiukkaa standardisointia, sillä haastattelujen perusteella datan arvojen etumerkkeihin ja aikasarjojen nimiin ei ole olemassa yhteisiä sääntöjä. Joka tapauksessa pohja työlle on jo olemassa nimettyjen vastuualueiden muodossa, mikä helpottanee kehitystyötä jatkossa. Silvola et al. (2016) korostavat muutosjohtamisen ja henkilöstön osaamisen kehittämisen merkitystä datan laadun hallinnassa, joten lisäkoulutukset voisivat olla toimiva vaihtoehto tilanteen parantamiseksi.

5.1.3 Tietoryhmäkohtaiset haasteet

Ennusteiden laatua pidettiin ongelmallisena erityisesti niiden käyttäjien toimesta. Huomionarvoista on, että valtaosa käyttäjistä puhui ennusteista nimenomaan yhtenä nippuna, vaikka todellisuudessa ne jakautuvat useaan eri ydintietoon (katso liite A). Yksi selitys tälle on, että tuotanto- ja kulutusennusteet sijaitsevat samassa tietojärjestelmässä. Ennusteet voidaan ajatella useasta eri lähdedatasta muodostuvana tietotuotteena, sillä niiden laskennassa hyödynnetään muun muassa sääennusteita sekä historiallisia mittaustietoja. Aineiston perusteella on hankala arvioida, ovatko laatuongelmat peräisin viallisesta lähtödatasta vai ennusteen laskennasta.

Kohdeorganisaatiossa on aiemmin teetetty tutkimus ennusteiden laadusta, jonka mukaan tuulivoiman tuotantoennusteen tarkkuus sekä saatavuus ovat hyvällä tasolla. (Heikura 2020). Tähän tulokseen verrattuna on mielenkiintoista, että ennusteiden maine on käyttäjien keskuudessa heikko. Samalla tulos on hyvä esimerkki datan laadun *fitness for use* -määritelmän tärkeydestä: matemaattisesti arvioituna data on keskimäärin hyvälaatuista, mutta käyttäjien keskuudessa se voidaan jopa sivuuttaa kokonaan ajoittain. Strong et al. (1997) havaitsivat, että dataa ei käytetä, jos sen maine on kärsinyt esimerkiksi harkinnanvaraisen tuotantoprosessin takia. Tässä tapauksessa kyse voi olla samanlaisesta ilmiöstä, sillä ennusteiden tuotantoprosessi ei välttämättä ole käyttäjien tiedossa, ja huonot kokemukset ennusteiden tarkkuudesta jäävät mieleen. Ennusteet ovat myös hyvä esimerkki datasta, jota voidaan ajatella TDQM-menetelmän (Wang 1998) mukaisena tietotuotteena: kyseessä on informaatiota, joka on jalostettu eri lähteistä saatavasta datasta ja sillä on itsessään suuri arvo liiketoiminnalle. Näin ollen myös sen laatuun tulisi kiinnittää huomiota. Osa datasta ja itse ennusteet ovat riippuvaisia ulkoisista tahoista (Heikura 2020), joten niiden laadunhallinta ei ole täysin kohdeorganisaation vallassa.

Toinen vain yhtä ydintietoryhmää koskettava haaste on verkonhallinnan toiminnanohjausjärjestelmä ja sen sisältämät verkko-omaisuustiedot. Tämä teeman ongelmat ovat varsin suoraviivaisia vanhentuneita henkilötietoja ja puutteellisia laitteiden ominaisuuk-

sia, joiden korjaaminen vaatinee lähinnä resursseja tietojen keräämiseen ja päivittämiseen. Lisäksi itse toiminnanohjausjärjestelmän hitaus ja monimutkaisuus heikentävät datan saatavuutta, mutta tämän ongelman korjaaminen voi vaatia järjestelmän uusimista tai vähintään laajamittaista kehitystyötä.

5.2 Organisaation datan laadun kypsyystaso

Kuten luvussa 2.5.2 todettiin, datan laadun kehittämisessä voidaan hyödyntää organisaation kypsyyden arviointia erikseen tarkoitusta varten kehitettyjä malleja hyödyntämällä. Haastatteluaineistosta havaittujen ongelmien perusteella voidaan määrittää kohdeorganisaation nykyinen kypsyystaso, jolloin kirjallisuuden kypsyysmallien avulla voidaan kartoittaa ja priorisoida tarvittavia kehitystoimenpiteitä seuraavalle tasolle pääsemiseksi. Taulukossa 16 on visualisoitu kohdeorganisaation nykytilaa Mahantin (2019, s. 295) osiin pilkotun mallin mukaisesti. Mallista on jätetty pois saavutettu hyöty, sillä haastatteluaineisto ei mahdollista sen arviointia. Taulukon muodostamaan matriisiin on merkitty vihreällä taso, jonka kriteerit organisaatio täyttää.

Taulukko 16. Kohdeorganisaation kypsyystaso Mahantin (2019, s. 295) mallia mukaillen

	Alkeellinen	**Toistettava**	**Määritelty**	**Hallittu**	**Tehokas**
Tekniikka	Yleissovellus (esim. Excel), manuaalisia prosesseja, tarpeen mukaan toteutettuja rutiineja	Taktisen tason työkaluja sovellustasolla tai yksiköissä siiloutuneesti	**Laatutyökaluja profilointiin ja siivoamiseen, tietovarasto, BI-sovelluksia**	Tietovarastoa ja BI-sovelluksia pidemmälle vietyjä laatutyökaluja, metatiedon hallintatyökaluja	Työkalut on standardoitu organisaation läpi. Alustaratkaisu datan profilointiin, valvontaan ja visualisointiin.
Asenne	Datan laatu nähdään kustannuksena	**Alustava tietoisuus datan hallinnan merkityksestä**	Dataa käsitellään organisaatiotasolla tuloksen kannalta kriittisenä	Dataa käsitellään mahdollisena kilpailuedun lähteenä	Data nähdään kriittisenä ja datan laatu mahdollistajana
Lähestyminen	**Tulipalojen sammuttamista tarpeen vaatiessa, ei juurisyiden analysointia**	Datan laadun dokumentointi mahdollistaa toistettavuuden	Isommat ongelmat dokumentoitu, mutta ei kokonaan ratkaistu	Proaktiivinen ehkäisevä työ	Strategista optimointia
Ihmiset	Ei datarooleja, ei tietoisuutta datan hallinnan käytännöistä	Tietokannan ylläpitäjä, datan laatu IT:n vastuulla	**Datan ylläpitäjä, tietovastaava ja -omistajaroolit nousemassa**	Useamman tason tietovastaavaroolit käytössä	Keskeinen datarooli

Tekniikan ja järjestelmien näkökulmasta nykyinen taso on määritelty: organisaatiossa on käytössä BI-työkaluja ja tietovarasto, sekä lisäksi joissain järjestelmissä on jo mukana

datan laatua valvovia ominaisuuksia. Haastatteluiden perusteella näitä työkaluja ei kuitenkaan juuri käytetä, joten pelkkä teknisten ratkaisujen olemassaolo ei vielä ratkaise ongelmia. Loshinin (2011) mallin teknologiakomponentti ei huomioi analytiikkatyökaluja tai tietovarastoa lainkaan, jolloin määritelty taso vaatisi organisaatiolta muun muassa standardoituja menetelmiä laatutyökalujen käyttöön sekä teknologiaa datan validointiin. Mahantin (2019, s. 295) mallia sovellettaessa BI- ja tietovarastokyvykkyydet voivat nostaa kyvykkyyttä jopa hieman perusteettomasti, sillä noiden työkalujen olemassaolo ei vielä suoraan auta datan laadun kehittämisessä.

Asennetta ei voi mitata kovin tarkasti kerätyn aineiston perusteella, mutta organisaatiossa on havahduttu nykytilan puutteisiin ja pohdittu mahdollisia kehitysaskeleita sekä otettu käyttöön datanhallintamallia asteittain, joten alustava tietoisuus datan hallinnan merkityksestä on olemassa. Datanhallintamallin juurtuessa organisaatioon myös tämä osa-alue voi kehittyä tietoisuuden kasvaessa.

Organisaation lähestymistapa on haastatteluiden perusteella hyvin reaktiivinen, eikä juurisyitä ongelmien taustalla ole analysoitu tai dokumentoitu ennen tätä tutkimusta. Tämä alue on selkeästi muita mallin osia jäljessä. Haastatteluissa tuli esiin puutteita datan, tietojärjestelmien ja tiedonsiirron valvonnassa sekä lukuisia esimerkkejä ongelmista, jotka on havaittu myöhässä. Myös Loshin (2011) pitää reaktiivista toimintaa ja laatuodotusten puuttumista tunnusomaisena alkeelliselle kypsyystasolle. Mahantin (2019) malli ei ota kantaa suoraan laadun mittaamiseen ja seurantaan, mutta luvussa 2.5.1 todettiin valvonnan olevan olennainen tekijä aktiivisemmassa datan laadun hallinnassa. Muissa malleissa ylemmät kypsyystasot vaativatkin laatusääntöjen määrittelyä sekä valvontaa (Spruit & Pietzka 2015, Loshin 2011).

Dataroolit ovat organisaatiossa muita osa-alueita edellä, sillä ydintiedot on tunnistettu ja niille on määritelty vastuuhenkilöt. Spruit & Pietzkan (2015) mallissa nimetyt tietovastaavat täyttäisivät jo hallitun tason vaatimukset, joten nykyinen malli on oikein hyödynnettynä toimiva. Haastatteluiden perusteella osa rooleihin nimetyistä henkilöistä eivät kuitenkaan käytännössä kanna vastuuta datan laadusta, vaan tietovastaavuus on jäänyt nimityksen tasolle.

5.3 Kehitysehdotukset

Kypsyystasojen vaatimusten perusteella suurimmat puutteet ovat organisaation lähestymistavassa, jota pitäisi kehittää proaktiivisempaan suuntaan, kuten haastateltavatkin tiedostivat. Teknologian ja vastuiden osalta organisaatio vaikuttaa olevan määritellyllä kypsyystasolla, mutta käytännössä datanhallintamallin ja teknisten ratkaisujen puutteellinen

jalkautus heikentää tilannetta. Taulukkoon 17 tiivistetyt kehitysehdotukset pohjautuvat näihin haastatteluaineiston kautta saatuihin havaintoihin sekä aiempaan datan laadun kehittämistä koskevaan tutkimuskirjallisuuteen.

Taulukko 17. Toimenpide-ehdotukset kohdeorganisaatiolle

Toimenpide	Sisältö lyhyesti	Vastaa ongelmiin
Aktiivinen valvonta	- Määritetään liiketoimintakriittiselle datalle laatukriteerit ja seurataan niiden toteutumista	- Mittarien ja valvonnan puute - välillisesti myös muut
Datan keskittäminen	Parannetaan datan saatavuutta ja yhtenäistetään esitystapaa - Siirtämällä enemmän dataa tietovarastoon - Kiinnittämällä huomiota integraatioiden määrään ja tarkoituksenmukaisuuteen - Määrittämällä päälähteet datalle	- Hajanainen arkkitehtuuri - Tietovaraston ja raportoinnin vajaakäyttö
Tietovirtojen dokumentointi	- Parannetaan datan saatavuutta kuvaamalla datan elinkaari eri lähteiden, prosessien ja järjestelmien kautta	- Hajanainen arkkitehtuuri
Datanhallintamallin jalkautus	- Viestitään ja koulutetaan henkilöstöä datanhallintamallista ja datan merkityksestä - Osallistetaan tietovastaavia muissa kehitystoimenpiteissä	- Datan hallinnointi ja yhteiset käytännöt - Mittarien ja valvonnan puute - Välillisesti myös muut
Datan korjaustoimenpiteet	- Korjataan puutteet verkko-omaisuustiedoissa - - Korjataan mahdolliset muut ongelmat, jotka tulevat ilmi aktiivisen valvonnan toteutuksen myötä (esim. integraatiot, ennusteiden lähtötiedot)	- Verkonhallinnan toiminnanohjausjärjestelmä - Mahdollisesti myös muut

5.3.1 Datan laadun aktiivinen valvonta

Haastateltavien mukaan epätietoisuus datan laadusta ja virheiden huomaamatta jääminen ovat molemmat olennaisia haasteita organisaation päivittäisessä toiminnassa. Ideaalitilanteessa virheellisen datan syntyminen ehkäistäisiin organisaation proaktiivisella toiminnalla, mutta käyttötoiminnan dataa tuotetaan valtaosin automaattisilla mittauksilla, eikä kaikkia mahdollisia mittausvirheitä ja järjestelmäkatkoksia todennäköisesti voida poistaa. Yksi kehitysaskel voisi olla organisaation datan laadun valvonnan tason kehittäminen Silvolan et al. (2011) asteikolla reaktiivisesta aktiiviseksi, jolloin datan laatua seurattaisiin reaaliaikaisesti ja ongelmat myös havaittaisiin itse. Tällöin ongelmiin voitaisiin puuttua ennen kuin ne näkyvät sidosryhmille tai sekoittavat laskennallisia malleja.

Datan laatumenetelmien mukaan valvonta tulisi ottaa käyttöön datan korjaustoimenpiteiden jälkeen (katso esimerkiksi Loshin 2011, McGilvray 2008), mutta tässä tapauksessa

liiketoiminnalle olennaisin data uusiutuu niin nopeasti, että korjaavat toimenpiteet vaikuttaisivat lähinnä pitkän aikavälin historiadatan analysointiin. Lisäksi erilaisten mittarien käyttöönotto ja seuranta voisi paljastaa puutteita datan laadussa, jotka ovat voineet jäädä käyttäjiltä huomaamatta, eivätkä siten ole tulleet haastatteluissa esiin. Samalla organisaation olisi mahdollista kerätä tarkempaa tietoa tiedonsiirtokatkoksista ja selvittää, onko ongelma erityisen vakava tiettyjen järjestelmien tai ydintietojen kohdalla.

Kaiken datan valvominen voi olla liian kallista, joten organisaation kannattaa keskittää toimenpiteet liiketoiminnan kannalta olennaisimpiin tietoihin (Mahanti 2019, s. 326). Tämä vaatii eri tietojen tärkeyden analysointia. Samalla on hyvä pohtia datan käyttäjiä ja tietovastaavia osallistaen, mitä liiketoimintasääntöjä datalla on, mitä laadunvalvonnassa halutaan mitata ja mitä käyttäjät odottavat datan laadulta, jotta erilaiset laatukriteerit ja -mittarit voidaan muodostaa. Mittaaminen sisältää aina vertailua: vertailukohteena voi olla esimerkiksi aiemman tiedon pohjalta muodostetut raja-arvot tai historiallinen data (Sebastian-Coleman 2013, s. 49–53). Käyttötoiminnan datalle voi olla hankala saada suoria vertailuarvoja, sillä valtaosalle kantaverkon käyttö- ja tilatiedoista ei ole vaihtoehtoista lähdettä. Tällöin tarkkuusulottuvuutta (eli kuinka hyvin data kuvaa todellisuutta) ei välttämättä voida hyödyntää, mutta sen voi korvata Sebastian-Colemanin (2013, s. 62–63) käyttämällä oikeellisuudella, jossa mittaustietoja verrataan ennalta määritettyihin arvoihin. Kirjallisuudessa tärkeiksi ulottuvuuksiksi ovat nousseet myös täydellisyys, oikea-aikaisuus sekä yhtenäisyys (Batini et al. 2009). Nämä ulottuvuudet ovat todennäköisesti käyttökelpoisia myös kohdeorganisaation datan laadun mittaamisessa, sillä havaituista ongelmista tiedonsiirtokatkokset voivat vaikuttaa kaikkiin kolmeen ulottuvuuteen. Lisäksi ajoittain jumiutuvat mittaukset voidaan havaita näitä ulottuvuuksia valvomalla.

Haastatteluiden perusteella asiantuntijoilla on tarve saada tietoa käyttämänsä datan laadusta, joten mahdollisten valvonta- ja mittaustoimenpiteiden tulokset tulisi saada myös helposti jakoon tiedon käyttäjille. Allen & Cervo (2015) jakavat datan laatumittarit kahteen kategoriaan: valvontamittarit ilmoittavat välitöntä korjaamista vaativista poikkeamista, kun taas raportointinäkymät (engl. dashboards) ja tuloskortti (engl. scorecard) tarjoavat yleiskuvaa datan laadusta esimerkiksi laatu-ulottuvuuksien kautta kuvattuna. Raportointinäkymien hyödyntämistä suosittelevat myös Mahanti (2019, s. 326) ja Loshin (2011) kypsyysmallissaan. Toisaalta raportointinäkymät voivat olla liian hidas käyttäjien toivomaan laatuseurantaan datan hyödyntämisen ollessa parhaimmillaan lähes reaaliaikaista. Ennen teknisen toteutuksen miettimistä on siis pohdittava, mitä halutaan valvoa. Käytettävien mittarien tulisi olla ymmärrettäviä, toistettavissa sekä tarkoituksenmukaisia (Sebastian-Coleman s. 44–46) ja niissä on hyvä huomioida käyttäjien tarpeet (Cappiello et al. 2004).

5.3.2 Datan keskittäminen

Haastatteluissa eniten esiin nousseet ongelmat liittyivät datan hajautuneisuuteen eri järjestelmissä. Erityisesti saatavuuteen ja tiedonsiirtoihin liittyvät haasteet näkyivät haastateltujen asiantuntijoiden päivittäisessä työssä puuttuvan datan ja sen etsimiseen kulutetun ajan muodossa. Kirjallisuudesta tunnistetut ratkaisut keskittyvät tietojärjestelmäkokonaisuuden muokkaamiseen muun muassa vähentämällä integraatioita (Silvola et al. 2011) ja kiinnittämällä huomiota tiedon virtauksen järkevyyteen eri järjestelmien välillä (Umar et al. 1999). Lyhyellä aikavälillä kohdeorganisaation tietojärjestelmäarkkitehtuurin uusiminen ei ole realistista, mutta ydintiedon hallinnan teorian mukaisesti (Silvola et al. 2011; Dreibelbis et al. 2008) tietoja voisi pyrkiä siirtämään aiempaa laajemmin koko yhtiön laajuiseen tietovarastoon tai mahdollisesti jollekin muulle tarkoituksenmukaiselle data-alustalle, mikäli tietovaraston tekniset rajoitteet, kuten päivitystahti, muodostuvat liian suureksi esteeksi. Kaikki haastateltavat eivät olleet tietoisia tietovaraston olemassaolosta ja käytännöistä, vaikka saattoivat kaivata samaan aikaan lisää Power BI -raportteja tai muita visualisointeja datasta. Tietoisuuden lisääminen tietovarastosta ja sen sisältämän datamäärän kasvattaminen voisi siis edistää analytiikan hyödyntämistä ja siten kehittää tietojen esitystapoja saatavuuden parantamisen ohella.

Lisäksi haastatteluaineiston perusteella henkilöstöllä on haasteita luotettavien tietolähteiden hahmottamisessa, joten ydintietojen hallintaa voisi kehittää määrittämällä jokaiselle pääjärjestelmän, josta tieto pääsääntöisesti haetaan. Allen & Cervo (2015) toteavat, että tällaisen pääjärjestelmän (engl. system of record) määrittäminen on välttämätöntä ristiriitatilanteiden välttämiseksi. Ideaalitilanteessa jokaisella tiedolla on yksi pääjärjestelmä (Silvola et al. 2011), mutta esimerkiksi häiriötilanteissa tieto voidaan hakea jostain toisesta lähteestä (Allen & Cervo 2015). Vaikka kohdeorganisaation tapauksessa joitain tietoja ei voitaisi sijoittaa tietovarastoon sen teknisten rajoitteiden takia, pääjärjestelmän määrittäminen voisi auttaa selventämään ydintietojen elinkaarta.

5.3.3 Tietovirtojen kuvaaminen

Haastatteluissa datan saatavuus koettiin ongelmalliseksi järjestelmien suuren määrän takia. Datan virtaus eri järjestelmien läpi ei ole selkeä käyttäjille, ja sen mahdollinen tunteminen perustuu lähinnä kokemuksen kautta kerättyyn hiljaiseen tietoon. Myös alkuperäinen tietolähde voi olla epäselvä loppukäyttäjille. Tilannetta voisi parantaa datan elinkaaren visualisoiminen niin, että käyttäjät voivat nopeasti tarkistaa datan kulkeutumisen eri järjestelmien läpi aina tarvittaessa. Oikein toteutettuna mallinnus voisi myös edesauttaa aktiivisempaa dataongelmien korjaamista ja niiden ehkäisyä, sillä datavirtoja kuvaava dokumentaatio auttaa selvittämään myös syitä mahdollisten tulevien ongelmien

taustalla sekä esimerkiksi löytämään datan myös integraatiokatkosten yhteydessä. Tällainen dokumentaatio voisi edistää datan saatavuutta, kun oikean tiedon löytäminen ei olisi enää ainoastaan yksilöiden keräämän hiljaisen tiedon varassa.

Luvussa 2.5.1 esitelty IP-MAP-työkalu (Shankaranarayan et al. 2003) on yksi mahdollinen työkalu datavirtojen mallintamiseen. Toisaalta haastatteluiden perusteella mallinnuksen ei tarvitse olla yksityiskohtainen, vaan olennaista olisi visualisoida eri järjestelmät ja niiden väliset suhteet datan näkökulmasta.

5.3.4 Datanhallinnan jalkautus

Fingridin datanhallintamalli on vielä uusi, ja haastatteluiden perusteella sen jalkautuksen tasossa on suuria eroja eri asiantuntijoiden välillä. Olennaista dataa on tunnistettu ja vastuuhenkilöitä on nimetty, mutta tietovastaavien rooli ei vielä vaikuta olevan täysin selkeä heille itselleen. Tietovastaavien aktivoiminen ja osallistaminen nivoutuu myös aktiivisempaan valvontaan, sillä sekä kirjallisuuden (katso luku 2.3.2) että yhtiön datanhallintamallin mukaisesti vastuun laadun seurannasta tulisi olla tietovastaavilla.

Henkilöstön osaamistaso on merkittävä tekijä datan laadun hallinnassa (Haug & Albjørn 2011; Umar et al. 1999), joten yksi konkreettinen toimenpide voisi olla tietovastaavien ja muun henkilöstön koulutus datanhallintamallista sekä datan merkityksestä liiketoiminnassa ylipäätään. Samalla tämä voisi edistää tietämystä tietovarastosta ja raportointimahdollisuuksista, mikä sitoo tämän myös datan keskittämiseen.

Viestinnän ja koulutuksen ohella on hyvä huomioida vastuurooliin nimettyjen asiantuntijoiden käytössä olevat aikaresurssit. Tällä hetkellä tietovastaavan rooliin ei ole kuulunut juurikaan proaktiivisia tehtäviä tai aktiivista seurantaa, joten roolin sisällön terävöityessä myös vastaavilta vaadittu aikaresurssi voi kasvaa. Myös motivaatio tehtävien hoitamiseen voi muuttua, jos työn määrä tai sisältö vaihtuvat. Yksi kirjallisuudessa tunnistettu tekijä on henkilöstön palkitseminen hyvästä datan laadusta (Haug et al. 2013; Haug & Albjørn 2011; Umar et al. 1999), mikä voisi olla yksi keino herättää myös kohdeorganisaation henkilöstö ajattelemaan datan laadun merkitystä.

5.3.5 Datan korjaustoimenpiteet

Haastatteluissa ilmenneet puutteet verkonhallinnan toiminnanohjausjärjestelmän tiedoissa olisi hyvä korjata täydentämällä puuttuvat tiedot ja päivittämällä vanhentuneet henkilötiedot. Laitetietojen täydentäminen ja henkilötietojen päivittäminen vaativat todennäköisesti manuaalista työtä, sillä ajantasaisia tietoja tuskin löytyy mistään valmiina ratkaisuna automaattista täyttöä varten. Näiden harvemmin päivittyvien tietojen kohdalla

voisi olla hyvä pohtia päivitysprosessin ja -vastuiden tarkentamista, jotta puutteilta vältytään jatkossa.

Automaattisesti kerätty data vaatii korjaustoimenpiteitä, sillä vaikka virheellinen data voidaan havaita, sen ympäristöön pääsyn estäminen voi olla hyvin vaikeaa tai mahdotonta. (Loshin 2011) Aktiivisempi datan valvonta voi siis paljastaa nykyistä tehokkaammin virheitä, jotka vaativat datan siivoamista. Esimerkiksi haastatteluissa mainitut ennusteiden ongelmat voivat olla ainakin osittain peräisin virheellisestä lähtödatasta, jolloin sen korjaaminen parantaa myös ennusteiden tarkkuutta.

6. PÄÄTELMÄT

Tutkimuksen tarkoituksena oli selvittää kohdeorganisaation datan laadun ongelmia ja pohtia alustavia kehitystoimenpiteitä tilanteen parantamiseksi. Vastauksia haettiin sekä kirjallisuudesta että haastatteluaineistosta. Kirjallisuuskatsauksessa myös määriteltiin datan laadun käsite ja pohdittiin sen arviointi- ja kehitysmenetelmiä, datan hallinnointia osana datan laadun hallintaa sekä aiemmissa tutkimuksissa havaittuja laatuongelmia. Empiirisessä osiossa toteutettiin haastattelututkimus kohdeorganisaation datan käyttäjille, ja haastatteluista saatuja tuloksia peilattiin kirjallisuudesta löydettyihin havaintoihin.

Ensimmäinen tutkimuskysymys oli:

1. Mitä ongelmia käyttötoiminnan datan laadussa on tällä hetkellä?

Jotta kysymykseen voitiin vastata kattavasti, kirjallisuuskatsauksessa määriteltiin ensin datan laatu ja pohdittiin sen arviointimenetelmiä. Vakiintuneen määritelmän mukaan data on laadukasta, kun se on käyttötarkoitukseen soveltuvaa. Tämän hahmottamisen ja mittaamisen helpottamiseksi datan laatu voidaan jakaa erilaisiin subjektiivisiin ja objektiivisiin ulottuvuuksiin, joita eri mallit tunnistavat useita. Yleisimpiä ulottuvuuksia ovat tarkkuus, täydellisyys, oikea-aikaisuus ja yhtenäisyys. Vakiintuneita määritelmiä ulottuvuuksille ei kuitenkaan ole – eri mallit voivat määritellä ulottuvuudet eri tavalla samoista nimityksistä huolimatta, ja ne voivat olla myös osittain päällekkäisiä.

Datan laadun arviointiin on kehitetty useita erilaisia teoreettisia menetelmiä, joissa voidaan hyödyntää laadun ulottuvuuksia. Menetelmät voivat hyödyntää joko objektiivisia mittaustuloksia, subjektiivisempia datan käyttäjien omia kokemuksia tai yhdistellä näitä molempia. Olennaista on valita menetelmä ja arvioitavat ulottuvuudet kohteen tarpeiden perusteella, sillä kirjallisuus ei tunnista yhtä vakiintunutta menetelmää. Myös objektiivisten mittareiden suunnittelussa tulee huomioida datan käyttäjien tarpeet. Empiiristä osiota varten luvussa 2.2.2 esitellystä AIMQ-menetelmän kyselypohjasta muodostettiin tarkoitukseen sopiva haastattelurunko, jonka avulla tuotettiin katsaus datan käyttäjien havaitsemista nykytilan laatuongelmista.

Vastauksena ensimmäiseen tutkimuskysymykseen kirjallisuudessa havaittuja ongelmien aiheuttajia ovat muun muassa tietojärjestelmien väliset integraatiot, käyttäjien tekemät virheet sekä epäyhteneväisyydet järjestelmien välillä. Lisäksi hyvälaatuisen datan saavuttaminen vaatii organisaatiolta myös toimivaa datan hallinnointia, eli esimerkiksi dataan liittyvien sääntöjen ja vastuuroolien määrittelyä ja noudattamista. Vastuu datasta olisi hyvä olla liiketoimintayksiköissä, jotka dataa käyttävät, sillä heillä on myös paras

tieto siihen kohdistuvista tarpeista. Empiirisessä osiossa havaitut ongelmat olivat pääosin samanlaisia: erityisesti esiin nousivat hajanaisten järjestelmien aiheuttamat saatavuus- ja tiedonsiirto-ongelmat, valvonnan puute, tietovaraston ja raportoinnin vajaakäyttö, datanhallintamallin jalkautuksen keskeneräisyys sekä toiminnanohjausjärjestelmän tietojen puutteellisuus.

Toinen tutkimuskysymys koski datan laadun kehittämistä:

2. Miten käyttötoiminnan datan laatua voidaan kehittää?

Datan laadun kehittäminen jaetaan kirjallisuudessa kahteen näkökulmaan: reaktiivisessa lähestymistavassa laatupuutteita pyritään korjaamaan sitä mukaa kun niitä havaitaan, kun taas proaktiivisessa lähestymistavassa pyritään löytämään ongelmien juurisyitä ja ehkäisemään virheiden syntyminen jo ennalta. Reaktiivisessa lähestymistavassa virheiden etsinnässä voidaan hyödyntää datan profilointia eli metatietojen koostamista puutteellisen datan löytämiseksi. Ongelmat voidaan korjata esimerkiksi korvaamalla virheellinen data uudella tai standardoimalla data yhtenäiseen muotoon. Proaktiivisessa lähestymistavassa voidaan esimerkiksi suunnitella uudelleen datan tuotanto- ja käyttöprosesseja virheiden juurisyiden poistamiseksi. Nykytilan ja tavoitetason määrittelyssä voidaan hyödyntää organisaation datan laadun kypsyysmalleja. Kohdeorganisaation ongelmien korjaamiseksi määriteltiin viisi kehitystoimenpidettä kirjallisuuden ja kypsyysmallien pohjalta: datan laadun aktiivinen valvonta, datan keskittäminen, tietovirtojen dokumentointi, datanhallinnan jalkauttaminen sekä puutteellisen datan korjaaminen.

6.1 Tutkimuksen merkitys

Tutkimus täydentää datan laadun tutkimuskenttää uudella tapaustutkimuksella, jossa kohdeorganisaation hyödyntämä ydintieto on perinteisestä asiakas- tai tuotedatasta poikkeavaa lähes reaaliaikaisesti kerättävää aikasarjadataa. Kohdeorganisaatio on myös ottanut käyttöön oman datanhallintamallinsa, joten tutkimuksessa saatiin myös kevyt katsaus datan hallinnoinnin käyttöönottoon yhden yksikön näkökulmasta. Mainituista erityispiirteistä huolimatta havaitut ongelmat ovat kuitenkin samanlaisia kuin muissa tapaustutkimuksissa (katso esimerkiksi Haug et al. 2013; Silvola et al. 2011; Umar et al. 1999) tunnistetut haasteet erityisesti tietojärjestelmien hajanaisuuden ja jatkuvan valvonnan puutteen osalta. Näin ollen tutkimus vahvistaa myös aiempien tutkimusten tuloksia.

Kohdeorganisaatio saa tutkimuksesta uutta tietoa käytännöistään ja niihin liittyvistä haasteista. Haastattelutulokset muodostavat kattavan kuvauksen tietoalueen datan laatuongelmista sekä niiden juurisyistä, ja kirjallisuuden avulla muodostetut kehitysehdotukset on muotoiltu kohdeorganisaation tarpeisiin sopiviksi.

6.2 Tutkimuksen arviointi ja rajoitteet

Laadullista tutkimusta voidaan arvioida validiteetin (tutkitaan sitä, mitä väitetään tutkittavan) sekä reliabiliteetin (tulosten toistettavuus) kautta (Tuomi & Sarajärvi 2018). Tapaustutkimuksissa validiteetti jaetaan vielä kolmeen osaan: rakenteelliseen, sisäiseen ja ulkoiseen validiteettiin (Yin 2018, s. 42–43). Validiteetin ja reliabiliteetin todentamismenetelmät ja niiden soveltamistavat tässä tutkimuksessa on esitelty taulukossa 18.

Taulukko 18. *Yin (2018, s. 43) tapaustutkimuksen laatutestit ja niiden toteutus*

Testi	**Menetelmä**	**Toteutus työssä**
Rakenteellinen validiteetti	Todistusaineiston kerääminen useasta lähteestä	Haastateltavia yhteensä 12 neljästä eri tietoryhmästä
	Raporttiluonnoksen arviointi tärkeimpien tiedonantajien toimesta	Kohdeorganisaation projektin ohjausryhmä sai luonnoksen kommentoitavaksi
Sisäinen validiteetti	Toistuvien rakenteiden yhdistely	Analyysivaiheessa toistuvat toisiinsa liittyvät ongelmat nostettiin erillisiksi teemoiksi
	Selityksien rakentaminen	Analyysivaiheessa yhdisteltiin havaintoja muodostaen mahdollisia syy-seuraussuhteita
	Kilpailevien selityksien huomiointi	Pyritty huomioimaan epävarmoissa tilanteissa (esim. ennusteiden ongelmien juurisyyt, tiedonsiirto-ongelmien taustasyyt)
	Logiikkamallien käyttö	Ei hyödynnetty lyhyen aikahorisontin vuoksi
Ulkoinen validiteetti	Teorian hyödyntäminen yhden tapauksen tutkimuksissa	Tutkimuskysymyksiin haettiin vastausta myös kirjallisuudesta ja tuloksia vertailtiin keskenään. Haastattelukysymykset ja toimenpide-ehdotukset muotoiltiin aiemman kirjallisuuden pohjalta.
Reliabiliteetti	Tapaustutkimuksen protokollan käyttö	Tutkimus on toteutettu ennalta laaditun tutkimussuunnitelman pohjalta, jossa linjattiin mm. aineistonkeruun ja raportin pääpiirteet.
	Tapaustutkimuksen tietokannan kehittäminen	Haastattelutallenteet ja niiden litteroinnit tallennettiin tutkimuksen ja raportin kirjoittamisen ajaksi.
	Todistusketjun ylläpitäminen	Haastatteluaineistosta on poimittu suoria sitaatteja ja sisällönanalyysiä on havainnollistettu esimerkein.

Rakenteellinen validiteetti tarkastelee tutkimuksen menetelmien soveltumista tutkimuskohteeseen (Yin 2018, s. 42). Suurimpana rakenteellisena rajoitteena tutkimuksessa kerättiin aineistoa ainoastaan haastatteluilla, sillä esimerkiksi kvantitatiivinen datapoikkeamien analysointi olisi ollut hyvin työlästä valitulla tietoaluerajauksella. Haastattelemalla kerätty aineisto on muistinvaraista ja siinä voi näkyä haastateltavien ennakkoasen-

teet (Yin 2018, s. 121). Lisäksi videopuheluina toteutettujen haastatteluiden vuorovaikutus ei välttämättä ole ollut yhtä luontevaa kuin kasvotusten, mikä on voinut lyhentää vastauksia (Saunders et al. 2019, s. 473–474). Validiteetin parantamiseksi haastateltavia valittiin kolme jokaisen tietoryhmän sisältä, mutta kyseessä on yhteensä vain 12 ihmisen otanta. Vaikka käytetty data on samaa, jokaisella asiantuntijalla voi olla sille omat käyttötarkoituksensa ja -tapansa, joten kaikkia käyttäjien kohtaamia ongelmia ei välttämättä ole huomioitu haastatteluaineistossa. Toisaalta haastateltavien mainitsemat ongelmat esiintyivät pääosin kaikissa haastatelluissa ryhmissä, joten tuloksia voidaan pitää johdonmukaisina.

Sisäistä validiteettia tarkastellaan lähinnä selittävissä tapaustutkimuksissa (Yin 2018 s. 42), mutta vaikka tämän työn asetelma on kuvaileva, muodostettiin aineistoa analysoidessa myös syy-seuraussuhteita ja selityksiä ongelmille. Haastateltavat toivat esiin omia näkemyksiään ongelmien taustoista, ja näitä selityksiä pyrittiin yhdistelemään laajemmiksi kokonaisuuksiksi tulosten esittelyssä. Aineistosta ei pysty kuitenkaan löytämään varmasti selityksiä kaikille ongelmille, joten esimerkiksi ennusteiden epätarkkuuden taustatekijöiksi on pohdittu puutteellisen lähtödatan ohella ennustejärjestelmän laskelmien puutteita.

Ulkoista validiteettia eli tulosten yleistettävyyttä pyrittiin parantamaan vastaamalla tutkimuskysymyksiin myös kirjallisuuskatsauksessa (luvut 2.4 ja 2.5), muodostamalla haastattelukysymykset valmiin kirjallisuudessa esitellyn ja testatun AIMQ-menetelmän pohjalta sekä pohtimalla havaittujen ongelmien ratkaisuja peilaamalla niitä aiemmissa tapaustutkimuksissa havaittuihin epäkohtiin sekä niihin kehitettyihin ratkaisuehdotuksiin. Kyseessä on kuitenkin yhden tapauksen tutkimus, joten haastatteluaineiston havainnot eivät todennäköisesti ole sellaisenaan laajasti yleistettävissä. Toisaalta havaitut ongelmat olivat linjassa kirjallisuuskatsauksen havaintojen kanssa, eivätkä tiukasti sidoksissa kohdeorganisaation toimintaympäristöön tai dataan.

Tutkimus toteutettiin aineistolähteisenä laadullisena tutkimuksena, jolloin tutkijan omat ennakkoasenteet voivat näkyä tutkimuksessa tehtyjen rajausten ja valintojen myötä. Tällöin myös tutkimuksen toistettavuus voi kärsiä. Tätä on pyritty ehkäisemään muun muassa avaamalla valintojen perusteluita, nostamalla esiin esimerkkejä aineiston analysoinnista sekä käyttämällä suoria haastattelusitaatteja tulosten esittelyssä.

6.3 Jatkotutkimusalueet

Tutkimuksen rajoitteista voidaan johtaa useita jatkotutkimuslinjoja. Koska kyseessä on yhden tapauksen tapaustutkimus, voisi olla mielenkiintoista soveltaa samoja menetelmiä

toiseen sähköverkkoalan tai muuten samanlaista dataa hyödyntävään yhtiöön. Lisäksi datan laadun arviointiin ja kehittämiseen on kehitetty useita menetelmiä, mutta niitä ei ole juurikaan sovellettu käytännössä alkuperäisten tutkimusten testien jälkeen (Batini et al. 2009). Erilaisia menetelmiä voisi olla mielenkiintoista vertailla myös käytännössä, jotta niiden toimivuudesta saataisiin laajemmin empiirisiä tuloksia.

Kohdeorganisaation näkökulmasta jatkotutkimuksia voisi toteuttaa myös numeerisilla menetelmillä samaan aineistoon tai tiettyihin ydintietoihin tulosten vahvistamiseksi, mikä ei tämän työn resurssien puitteissa ollut mahdollista. Kokonaista tietoaluetta tutkiessa ei ollut mahdollista keskittyä kovin tarkasti yksittäisiin ydintietoihin, jolloin myöskään esitettävät kehitystoimenpiteet eivät kohdistu yksittäisiin tietojoukkoihin. Koko yrityksen datanhallintamallin toteutumista voisi tutkia myös muissa yksiköissä, sillä kohdeorganisaatiossa sen jalkauttaminen oli vielä puutteellista. Jos ongelma paljastuisi laajemmaksi, siihen voitaisiin puuttua myös ylemmän johdon toimesta. Datanhallintamallin jalkautusta voisi tutkia erikseen pitkittäistutkimuksena, jolloin sen etenemisestä esimerkiksi kahden ensimmäisen vuoden aikana voitaisiin muodostaa tarkempi kuva pelkän nykytilan kuvauksen sijaan.

LÄHTEET

Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: a survey. The VLDB Journal, 24(4), 557-581. http://doi.org/10.1007/s00778-015-0389-y

Allen, M., & Cervo, D. (2015). Multi-domain Master Data Management: Advanced MDM and Data Governance in Practice. Elsevier.

Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31(2), 150. https://doi.org/10.1287/mnsc.31.2.150

Batini, C., Cappiello, C., & Francalanci, C. (2009). Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, 41(3), 16:1-16:52. https://doi.org/10.1145/1541880.1541883

Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles. Springer International Publishing. https://doi.org/10.1007/978-3-319-44421-5_9

Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. ACM. http://doi.org/10.1145/1012453.1012465

Chen, W., Zhou, K., Yang, S., & Wu, C. (2017). Data quality of electricity consumption data in a smart grid environment. Renewable and Sustainable Energy Reviews, 75, 98-105. https://doi.org/10.1016/j.rser.2016.10.054

Dreibelbis, A., Milman, I., van Run, P., Hechler, E., Oberhofer, M., & Wolfson, D. (2008). Enterprise master data management: an SOA approach to managing core information. IBM Press/Pearson plc.

Ehrlinger, L., Rusz, E., & Wöß, W. (2019). A Survey of Data Quality Measurement and Monitoring Tools. https://arxiv.org/abs/1907.08138

Eskola, J., Lätti, J., & Vastamäki, J. (2018). Teemahaastattelu: lyhyt selviytymisopas. Teoksessa Raine Valli (toim.) Ikkunoita tutkimusmetodeihin 1, Metodin valinta ja aineistonkeruu virikkeitä aloittelevalle tutkijalle. PS-kustannus.

Ge, M., & Helfert, M. (2007). A review of information quality research-develop a research agenda. Paper presented at the 12th International Conference on Information Quality,

Haug, A., & Arlbjørn, J. S. (2011). Barriers to master data quality. Journal of Enterprise Information Management, 24(3), 288-303. https://doi.org/10.1108/17410391111122862

Haug, A., Arlbjørn, J. S., & Pedersen, A. (2009). A classification model of ERP system data quality. Industrial Management & Data Systems, 109(8), 1053-1068. https://doi.org/10.1108/02635570910991292

Haug, A., Arlbjorn, J. S., Zachariassen, F., & Schlichter, J. (2013). Master data quality barriers: an empirical investigation. Industrial Management & Data Systems, 113(1-2), 234-249. https://doi.org/10.1108/02635571311303550

Heikura, A. (2020). Ennusteiden laadun vaikutukset sähköjärjestelmän käyttötoimintaan. Diplomityö, Aalto-yliopisto. Saatavilla http://urn.fi/URN:NBN:fi:aalto-202005243235

Juhila, K. (2021a). Laadullinen tutkimus ja teoria. Teoksessa Jaana Vuori (toim.) Laadullisen tutkimuksen verkkokäsikirja. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Saatavilla https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/mita-on-laadullinen-tutkimus/laadullinen-tutkimus-ja-teoria/, viitattu 25.3.2021

Juhila, K. (2021b). Teemoittelu. Teoksessa Jaana Vuori (toim.) Laadullisen tutkimuksen verkkokäsikirja. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Saatavilla https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/analyysitavan-valinta-ja-yleiset-analyysitavat/teemoittelu/, viitattu 28.7.2021

Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. Journal of Network and Computer Applications, 73, 57-81. https://doi.org/10.1016/j.jnca.2016.08.002

Laihonen, H., Hannula, M., Helander, N., Ilvonen, I., Jussila, J., Kukko, M., Kärkkäinen, H., Lönnqvist, A., Myllärniemi, J., Pekkola, S., Virtanen, P., Vuori, V., & Yliniemi, T. (2013). Tietojohtaminen. Tampereen teknillinen yliopisto - Tiedonhallinnan ja logistiikan laitos.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). Journey to data quality. MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. Information & Management, 40(2), 133-146. https://doi.org/10.1016/S0378-7206(02)00043-5

Levitin, A., & Redman, T. (1995). Quality dimensions of a conceptual view. Information Processing & Management, 31(1), 81-88. http://doi.org/10.1016/0306-4573(95)80008-H

Liu, C., Nitschke, P., Williams, S. P., & Zowghi Didar. (2020). Data quality and the Internet of Things. Computing.Archives for Informatics and Numerical Computation, 102(2), 573-599. https://doi.org/10.1007/s00607-019-00746-z

Loshin, D. (2009). Master Data Management. Elsevier Science & Technology.

Loshin, D. (2011). The Practitioner's Guide to Data Quality Improvement (1st ed.). Elsevier Science & Technology.

Mahanti, R. (2019). Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. Quality Press.

Maydanchik, A. (2007). Data quality assessment. Technics Publications.

McGilvray, D. (2008). Executing data quality projects: ten steps to quality data and trusted information. Morgan Kaufmann.

Määttänen, M. (2020). Pohjoismainen tasehallintahanke vie kohti reaaliaikamarkkinoita. Fingrid-lehti. Saatavilla https://www.fingridlehti.fi/pohjoismainen-tasehallintahanke-vie-kohti-reaaliaikamarkkinoita/, viitattu 8.9.2021.

Otto, B., & Hüner, K. (2009). Functional Reference Architecture for Corporate Master Data Management. Working paper [BE HSG / CC CDQ / 21], Institute of Information Management, University of St Gallen, St Gallen.

Otto, B., Hüner, K. M., & Österle, H. (2012). Toward a functional reference model for master data quality management. Information Systems and E-Business Management, 10(3), 395-425. https://doi.org/10.1007/s10257-011-0178-0

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Association for Computing Machinery.Communications of the ACM, 45(4), 211. https://doi.org/10.1145/505248.506010

Redman, T. C. (1995). Improve Data Quality for Competitive Advantage. Sloan Management Review, 36(2), 99. https://link.gale.com/apps/doc/A16497691/ITOF?u=tampere&sid=bookmark-ITOF&xid=936898d5

Redman, T. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79-82. https://doi.org/10.1145/269012.269025

Redman, T. C. (2004). Barriers to successful data quality management. Studies in Communication Sciences: Journal of the Swiss Association of Communication and Media Research, 4(2), 53. https://doi.org/10.5169/seals-790974

Redman, T. C. (2008). Data driven: profiting from your most important business asset. Harvard Business Press.

Redman, T. C. (2013). Data's Credibility Problem. Harvard Business Review, 91(12), 84-88.

Saunders, M. N. K., Thornhill, A., & Lewis, P. (2019). Research Methods for Business Students. Pearson Education, Limited.

Sebastian-Coleman, L. (2013). Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Morgan Kaufmann.

Shankaranarayan, G., Ziad, M., & Wang, R. Y. (2003). Managing data quality in dynamic decision environments: an information product approach. Journal of Database Management, 14, 14+. https://doi.org/10.4018/jdm.2003100102

Silvola, R., Harkonen, J., Vilppola, O., Kropsu-Vehkapera, H., & Haapasalo, H. (2016). Data quality assessment and improvement. International Journal of Business Information Systems; Ijbis, 22(1), 62-81. https://doi.org/10.1504/IJBIS.2016.075718

Silvola, R., Jaaskelainen, O., Hanna Kropsu-Vehkapera, & Haapasalo, H. (2011). Managing one master data - challenges and preconditions. Industrial Management & Data Systems, 111(1), 146-162. https://doi.org/10.1108/02635571111099776

Smith, H. A., & McKeen, J. D. (2008). Developments in Practice XXX: Master Data Management: Salvation Or Snake Oil? Communications of the Association for Information Systems, 23, 63-72. https://doi.org/10.17705/1CAIS.02304

Spruit, M., & Pietzka, K. (2015). MD3M: The master data management maturity model. Computers in Human Behavior, 51, 1068-1076. https://doi.org/10.1016/j.chb.2014.09.030

Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. Communications of the ACM, 40(5), 103-110. https://doi.org/10.1145/253769.253804

Tayi, G., & Ballou, D. (1998). Examining data quality. Communications of the ACM, 41(2), 54-57. http://doi.org/10.1145/269012.269021

Tuomi, J., & Sarajärvi, A. (2018). Laadullinen tutkimus ja sisällönanalyysi. Tammi.

Umar, A., Karabatis, G., Ness, L., Horowitz, B., & Elmagardmid, A. (1999). Enterprise Data Quality: A Pragmatic Approach. Information Systems Frontiers, 1(3), 279-301. https://doi.org/10.1023/A:1010006529488

Vilminko-Heikkinen, R., & Pekkola, S. (2013). Establishing an Organization's Master Data Management Function: A Stepwise Approach. Paper presented at the - 2013 46th Hawaii International Conference on System Sciences, 4719-4728. https://doi.org/10.1109/HICSS.2013.205

Vilminko-Heikkinen, R., & Pekkola, S. (2017). Master data management and its organizational implementation: An ethnographical study within the public sector. Journal of Enterprise Information Management, 30(3), 454-475. https://doi.org/10.1108/JEIM-07-2015-0070

Vilminko-Heikkinen, R., & Pekkola, S. (2019). Changes in roles, responsibilities and ownership in organizing master data management. International Journal of Information Management, 47, 76-87. https://doi.org/10.1016/j.ijinfomgt.2018.12.017

Vuori, J. (2021). Tapaustutkimus. Teoksessa Jaana Vuori (toim.) Laadullisen tutkimuksen verkkokäsikirja. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Saatavilla https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/tutkimusasetelma/tapaustutkimus/, viitattu 12.4.2021

Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86-95. https://doi.org/10.1145/240455.240479

Wang, R. (1998). A product perspective on total data quality management. Communications of the ACM, 41(2), 58-65. https://doi.org/10.1145/269012.269022

Wang, R. W., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5-33. https://doi.org/10.1080/07421222.1996.11518099

Weber, K., Otto, B., & Österle, H. (2009). One Size Does Not Fit All---A Contingency Approach to Data Governance. ACM Journal of Data and Information Quality, 1(1), 1-27. http://doi.org/10.1145/1515693.1515696

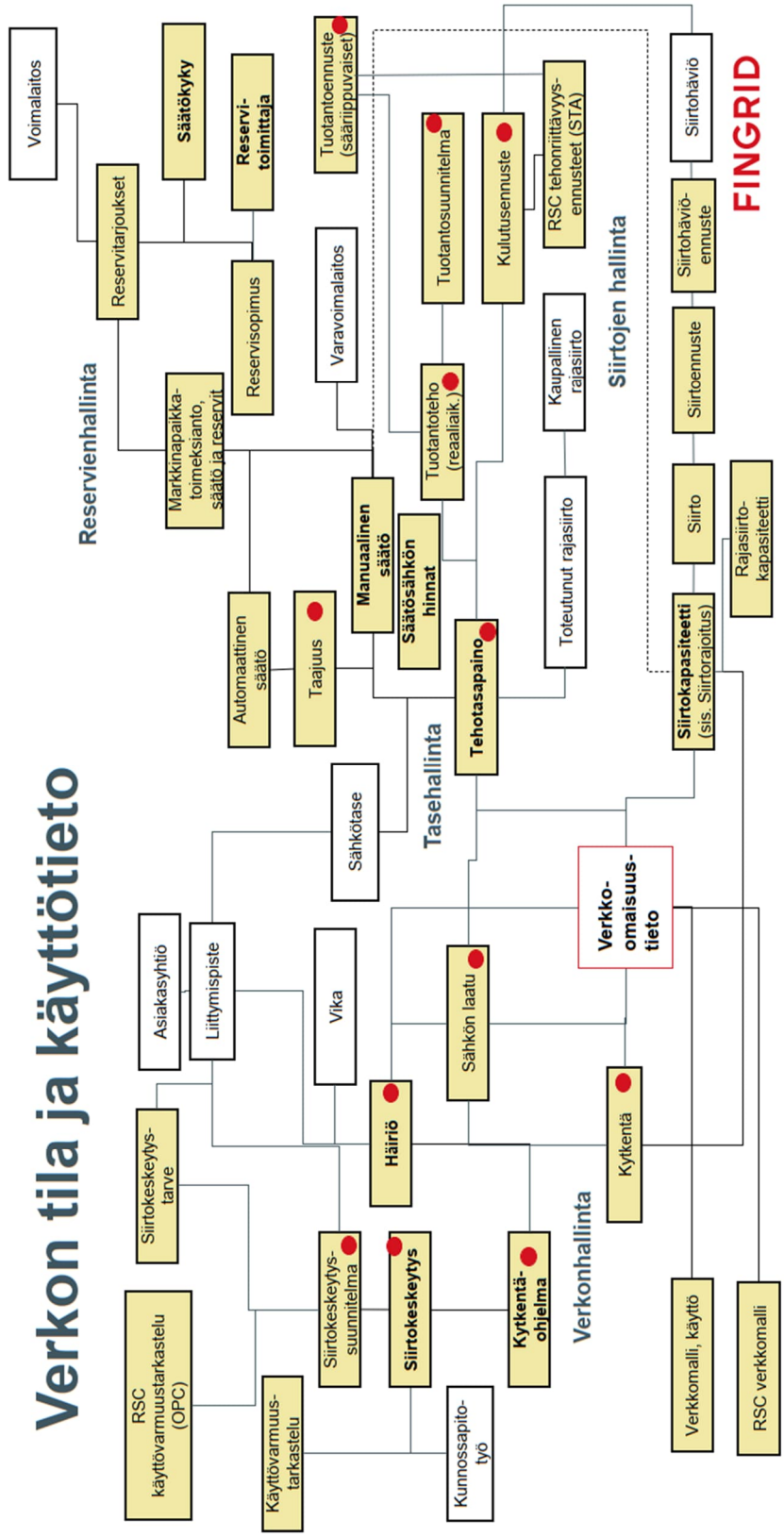
Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: The Hybrid Approach. Information & Management, 50(7), 369-382. https://doi.org/10.1016/j.im.2013.05.009

Xiao, Y., Lu, L. Y. Y., Liu, J. S., & Zhou, Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. Journal of Informetrics, 8(3), 594-605. https://doi.org/10.1016/j.joi.2014.05.001

Yin, R. K. (2018). Case study research and applications: design and methods (6th ed.). SAGE.

Yoon, V. Y., Aiken, P., & Guimaraes, T. (2000). Managing organizational data resources: Quality dimensions. Information Resources Management Journal, 13(3), 5-13. https://doi.org/10.4018/irmj.2000070101

LIITE A: TIETOALUEEN YDINTIEDOT



LIITE B: HAASTATTELURUNKO

Aloitus

- Kerrotko omin sanoin työnkuvastasi - mitä kaikkea siihen kuuluu?
- Kauanko olet ollut nykyisissä tehtävissä tai Fingridillä ylipäätään?
- Mitä ydintietoja hyödynnät eniten työssäsi?
- Mikä on roolisi suhteessa näihin ydintietoihin: oletko datan tuottaja, käyttäjä, sovellusvastaava ja/tai tietovastaava?

Laatu-ulottuvuudet

Saatavuus

- Onko data helposti saatavilla?
- Onko data saatavissa nopeasti aina tarvittaessa?

Täydellisyys

- Onko datassa mukana kaikki tarvittavat arvot?
- Täyttääkö data työtehtävien asettamat vaatimukset?

Tarkkuus/virheettömyys

- Onko data tarkkaa?
- Ovatko datan arvot oikein?
- Onko data luotettavaa?

Oikea-aikaisuus

- Onko data riittävän tuoretta työhösi?

Sopiva määrä

- Onko dataa sopiva määrä?

Tiivis esitystapa

- Onko data esitetty tiiviisti ja ytimekkäästi?

Johdonmukainen esitystapa

- Onko data esitetty johdonmukaisesti samassa muodossa?

Merkityksellisyys

- Onko data hyödyllistä työssäsi?

- Onko data sopivaa työhösi?

Helppokäyttöisyys

- Onko dataa helppo käsitellä tarkoitukseen sopivaksi?
- Onko dataa helppo yhdistää muihin tietoihin?

Tulkittavuus/ymmärrettävyys

- Onko datasta helppo tulkita mitä se tarkoittaa?
- Onko datan merkitys helppo ymmärtää?

Uskottavuus

- Onko data uskottavaa?

Maine

- Onko datan laadulla hyvä maine?
- Tuleeko data hyvistä lähteistä?

Lopetus

- Tuleeko vielä mieleen jotain muuta datan laatuun liittyvää mistä haluaisit kertoa?