Matti Virkkunen

BIG DATAN LAATUONGELMIEN TUTKIMUKSEN TEEMAT

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2021

TIIVISTELMÄ

Virkkunen, Matti
Big datan laatuongelmien tutkimuksen teemat
Jyväskylä: Jyväskylän yliopisto, 2021, 26 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja: Seppänen, Ville

Tässä tutkielmassa selvitetään big datan laatuongelmien tutkimuksessa esiintyviä teemoja. Tutkielma toteutettiin systemaattisena kirjallisuuskatsauksena. Tutkimusprosessi toteutettiin kahdessa osassa. Ensimmäiseksi etsittiin lähdeaineistoa valitusta tietokannasta. Tutkimuksen vähäisyyden vuoksi lähdeaineiston etsintä rajattiin IEEE Xplorer Standards -tietokantaan. Hakutuloksista suodatettiin tutkimukseen sopivimmat julkaisut. Tarkempaan tarkasteluun valikoitui 23 julkaisua. Toisessa vaiheessa valittujen julkaisujen tiivistelmistä kartoitettiin tutkimuksissa esiintyneitä teemoja. Tutkimuksessa löydettiin julkaisuista yhteensä kahdeksan erilaista teemaa. Eniten julkaisuissa oli teemana datan laadun parantaminen. Tämä teema löytyi kahdeksasta julkaisusta. Seuraavaksi eniten julkaisuissa oli teemana datan laatuongelmat big datan kontekstissa sekä datan laadun arviointi. Kumpikin oli teemana kolmessa julkaisussa. Kolmanneksi eniten julkaisuissa oli teemana big datan laadun parantaminen. Tämä oli teemana kahdessa julkaisussa. Muita teemoja löytyi julkaisuista vain yhden kerran. Niitä olivat: big datan laadun varmistaminen, big datan laadun auditointi, big datan laadun haasteet terveydenhuollossa sekä big datan laadun hallinta. Tutkimusta big datan laatuongelmista on tehty vähän ja jatkotukimusta tarvitaan kaikkiin tutkimuksessa löytyneisiin teemoihin. Jatkotukimusta olisi hyvä myös keskittää tutkimuksessa vähemmän esiintyneisiin teemoihin.

Asiasanat: big data, datan laatu, datan laatuongelmat

ABSTRACT

Virkkunen, Matti
Research of big data quality issues
Jyväskylä: University of Jyväskylä, 2021, 26 pp.
Information Systems, Bachelor's thesis
Supervisor: Seppänen, Ville

This bachelor's thesis explores the themes involved in the study of big data quality problems. The study was carried out as a systematic literature review. The research process was carried out in two parts. First, the source material was searched in the selected database. Due to the paucity of research, the search for source material was limited to the IEEE Xplorer Standards database. The most suitable publications for the study were filtered from the search results. 23 publications were selected for closer examination. In the second phase, the themes of the studies were mapped from the abstracts of the selected publications. The study found a total of eight different themes in the publications. Improvement of data quality was the main theme in the publications. This theme was found in eight publications. The next most published topics were data quality issues in the context of big data and data quality assessment. Each was the theme of three publications. The third most published topic was optimizing the improvement of big data quality. This was the theme of two publications. Other themes were found in the publications only once. These included: big data quality assurance, big data quality auditing, big data quality challenges in healthcare, and big data quality management. Little research has been done on big data quality problems and further research is needed on all the themes found in this study. It would also be a good idea to focus further support on less frequent themes in the study.

Keywords: Big data, data quality, data quality problem

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Digitalisaation myötä erimuotoisien datojen määrä on radikaalisti lisääntynyt ja big data -termin käyttö on yleistynyt. Big datalla tarkoitetaan massiivista tietojoukkoa, joka sisältää strukturoitua ja ei-strukturoitua tietoa, kuvia, äänitteitä ja videoita. Datan lähteitä ovat mm. sää- ja liikennedata, sensorit, kirjalliset dokumentit, kuvat, terveysdata, yritysten data jne. Sensoreita on eri laitteissa pian satoja miljardeja ja niiden määrän ennustetaan kasvavan vuoteen 2030 mennessä noin 100 triljoonaan (Neittaanmäki, 2021). Ihmisten lisäksi monenlaiset mittausjärjestelmät tuottavat automaattisesti dataa kiihtyvällä vauhdilla. Kun yhdistää monipuoliset datalähteet ja datan määrän kiihtyvän tahdin, saa otteen siitä ongelmasta, johon big data -ilmiössä haetaan vastausta (Salo, 2014).

Jo nykyisellään tallessa olevan datan määrä on valtaisa. Arvioiden perusteella siitä vain noin 20 % on jollakin tavoin saatu käsiteltyä. Datan asema ja merkitys yhteiskunnassa on radikaalisti muuttumassa. Datan määrä maailmassa kasvaa eksponentiaalisesti (Lehto & Neittaanmäki, 2016). Maailmastamme on tullut yhä monimutkaisempi ja moniin vaikeisiin ongelmiin etsitään kiireellisesti ratkaisuja. Big datalla on tässä kehitystyössä merkittävä rooli. Kyky jalostaa ja analysoida dataa tehokkaasti onkin yhä keskeisempi yhteiskunnan tuottavuutta ja kilpailukykyä voimistava tekijä (Lehto & Neittaanmäki, 2016).

Datan määrän kasvaessa datan laatuun on alettu kiinnittää yhä enemmän huomioita. On tunnistettu, että laadukkaalla datalla on merkitystä päätöksenteossa. Mutta mitä on datan laatu ja miten datan laatua mitataan? Voidaan todeta, että datan laadun määritelmä vaihtelee datan käytölle asetettujen vaatimusten mukaan. Datan laatu riippuu aina datan sisäisten ominaispiirteiden lisäksi liiketoimintaympäristöstä (Cai, 2015). Datan laadun arvioinnissa tavoitteena on ymmärtää datan laatu suhteessa odotuksiin tai käyttöön tai molempiin. Täyttääkö datan laatu odotuksia tai tietyn käyttötarkoituksen vaatimuksia (Sebastian-Coleman, 2013).

Big datan laadunhallinnassa on tunnistettavissa useita haasteita. Big data sisältää suurimmaksi osaksi jäsentämätöntä dataa, jonka muuttaminen jäsentyneeksi dataksi jatkokäsittelyä varten on hyvin aikaa vievää. Datavirta on

jatkuvaa, jonka seurauksena datan ajantasaisuus on lyhentynyt merkittävästi. Laatustandardien kehittäminen big datalle on vasta alkamassa. Kattavaa tutkimusta big datan laatustandardeista ja arviointimenetelmistä ei juurikaan ole tehty (Cai, 2015).

1.1 Tutkimuskysymys ja tutkimuksen rakenne

Tutkielmassa pyritään vastaamaan seuraavaan tutkimuskysymykseen: Millaisia teemoja esiintyy big datan laatuongelmia käsittelevässä tutkimuksessa? Kandidaatintutkielma etenee seuraavasti. Johdannon jälkeen luvussa kaksi tarkastellaan big dataa ilmiönä, esitellään datan laatua yleisesti ja käsitellään big datan laatuongelmia. Kolmannessa luvussa esitellään systemaattinen tiedonhakuprosessi. Informaatioteknologian tieteenalan valikoidusta tietokannasta haettiin systemaattisesti tutkimuskirjallisuutta tutkimuskysymykseen. Viimeinen luku käsittelee saatuja tuloksia, mitä päätelmiä voidaan tehdä ja mahdollisia jatkotutkimuskysymyksiä.

1.2 Tutkimusmenetelmä

Kandidaatintutkielma toteutettiin kirjallisuustutkimuksena. Tutkimuksen taustoittamiseksi aineistoa big datasta ilmiönä, datan laadusta ja datan laatuongelmista etsittiin laajasti Jyväskylän yliopiston kirjaston tarjoamien palvelujen kautta kansainvälisistä e-aineistotietokannoista, kirjaston kokoelmatietokannoista sekä avoimesta verkosta. Kansainvälisistä e-aineistotietokannoista löytyi aihetta käsitteleviä tieteellisiä artikkeleita ja konferenssijulkaisuja. Tietokantahakujen lisäksi tutkimustaustakirjallisuutta etsittiin löytyneiden lähteiden viittauksien joukosta. Kirjastojen kokoelmatietokannoista valikoitui lähteiksi big dataa käsitteleviä e-kirjoja sekä painettuja kirjoja. Avoimesta verkosta löytyneiden julkaisujen tieteellisyyttä arvioitiin julkaisijan ja julkaisupaikan mukaan. Tiedonhankintasuunnitelmassa valikoitiin hakutermeiksi big data, data quality ja data quality problems. Hakutermejä myös yhdisteltiin hakuvaiheessa hakulausekkeiksi. Työn edetessä hakustrategiaa muokattiin tekemällä mm. rajauksia sekä julkaisuvuoden että vertaisiarvioinnin mukaan. Tavoitteena oli löytää mahdollisimman relevantteja lähteitä.

Tutkimuskirjallisuutta tutkimuskysymyksestä, mitä ovat big datan laatuongelmat, haettiin systemaattisesti informaatioteknologian alan IEEE Xplorer Standards -tietokannasta. Hakutermeinä käytettiin termejä big data, data quality, problems, issue. Myös tässä haussa hakustrategiaa muokattiin tekemällä rajauksia. Haun tulokset käytiin systemaattisesti läpi tarkastelemalla tiivistelmiä, tuloksia ja johtopäätöksiä. Saadut hakutulokset lajiteltiin teemoittain. Hakua ja saatuja tuloksia esitellään luvussa kolme.

2 MÄÄRITELMIÄ

Tässä luvussa käydään läpi tutkielmassa esiintyvät keskeiset käsitteet big data, big datan ominaisuudet ja datan laatu. Ensimmäiseksi pohditaan käsitettä big data ja toiseksi big datan laatua. Kolmanneksi käsitellään datan laatuongelmia.

2.1 Big data

Useimmat meistä ovat viime vuosina joko kuulleet tai lukeneet tiedotusvälineistä aiheesta big data. Big data ei ole kuitenkaan käsitteenä uusi. Ensimmäisiä kertoja se on esiintynyt erilaisissa tutkimuskirjallisuuden yhteyksissä jo 1990-luvulla. Big data on terminä kuvannut esimerkiksi suuria tietoaineistoja (Cox & Ellsworth, 1997) tai tallennettavan datan määrää tiedonlouhinnan yhteydessä (Weiss & Indurkhya, 1998).

Erilaisia big datan määritelmiä löytyy kirjallisuudesta paljon. Ylijoki on omassa väitöskirjassaan tutkinut käsitteen esiintymistä tutkimuskirjallisuudessa. Useimmat käsitteen määrittelyt ajoittuvat vuosille 2014–2016 (Ylijoki, 2019). Ylijoen mukaan ongelmana big datan määritelmissä on se, että ne ovat usein epäjohdonmukaisia. Yleinen epäkohta on yhdistää määritelmään tapausriippuvaisia elementtejä kuten arvo ja totuudenmukaisuus. Tämä herättää heti kysymyksen "arvoa kenelle?" tai "totuudenmukaisuus mihin tarkoitukseen?". Näihin ei löydy vastausta vain dataa tarkastelemalla (Ylijoki, 2019).

Tunnetuimpana big datan määritelmänä pidetään yleisesti Meta Group -nimisessä yrityksessä työskennelleen Doug Laneyn vuonna 2001 julkaisemassa muutaman sivun raportissa esiintynyttä määritelmää (Gandomi & Haider, 2015). Laney kuvaili raportissa ensimmäisenä big datalle kolme ominaisuutta: nopeus (engl. velocity), määrä (engl. volume) sekä moninaisuus (engl. variety). Big data kiinnostuksen kohteensa lähti uudelleen nousuun 2011. Kolmesta mainitusta ominaisuudesta määrä viittaa datan suureen määrään ja erheellisesti moni

suomentaakin sen näin. 2000-luvulla dataa on tullut valtavasti ja datan määrän paljous on ollut keskeinen tekijä big data -käsitteen synnylle (Salo, 2014).

Big datan englanninkieliset vastineet alkavat v-kirjaimilla ja kolmen v:n määritelmä Big datalle onkin vakiintunut tutkimuskirjallisuuteen (Gandomi & Haider 2015). Käsittelen seuraavaksi näitä big datan kolmea ominaisuutta.

Big datan ensimmäisellä ominaisuudella nopeus tarkoitetaan datan muodostumisen nopeutta. Dataa syötetään kiihtyvällä nopeudella tietojärjestelmiin. Kuinka nopeasti syötettyä dataa pitäisi pystyä analysoimaan ja kuinka nopeasti analysoinnin tuottamiin tuloksiin pitäisi reagoida? Älypuhelimien ja muiden digitaalisten laitteiden sekä sensorien määrän suuri kasvu on aiheuttanut datan muodostumisen räjähdysmäisen kasvun. Tämä on aiheuttanut kasvavan tarpeen datan ajantasaiselle prosessoinnille ja analysoinnille. Perinteiset datan hallintajärjestelmät eivät pysty prosessoimaan tuhansien datan lähteiden virtaa (Gandomi & Haider, 2015).

Toisella ominaisuudella määrä tarkoitetaan yleensä käsiteltävän aineiston kokoa. Big data-aineiston koko voi vaihdella teratavuista eksatavuihin. Big datan määrä on kuitenkin asiayhteydestä riippuvaista ja se voi vaihdella eri tekijöiden mukaan. Näitä tekijöitä voivat olla mm. aika ja tallennusformaatti. Tallennusmäärien kasvaessa se, mitä tänä päivänä pidetään big datana ei tulevaisuudessa välttämättä enää ole sitä. Myös erilaiset tallennusformaatit vievät erilaisen määrän tilaa. Tämän takia on hankalaa määritellä tiettyä raja-arvoa sille, minkä kokoista datamäärää voidaan pitää big datana. (Gandomi & Haider, 2015)

Kolmas big datan ominaisuus, moninaisuus viittaa tallennettavan datan rakenteen vaihtelevuuteen. Dataa voidaan tallentaa jäsennettynä datana (engl. strucured data), osittain jäsennettynä datana (engl. semi-structured data) sekä jäsentämättömänä datana (engl. unstructured data). Alla kuvataan jäsentelyä tarkemmin.

1. Jäsennetyllä datalla tarkoitetaan relaatiotietokantoihin tai laskentataulukoihin tallennettua dataa. Osittain jäsennettyä dataa ovat esimerkiksi erilaiset XML- ja JSON- dokumentit.

2. Osittain jäsennetty data ei edellytä tiukkoja standardeja ja se on myös yleensä koneluettavaa. Osittain jäsennettyjä dataformaatteja käytetään erityisesti internetissä käytävässä tiedonvälityksessä. Jäsentämätön data sisältää mm. teksti-, ääni- sekä videotiedostoja.

3. Jäsentämätöntä dataa kerätään useista lähteistä: sosiaalisesta mediasta, älypuhelimista, sensoreista, kuvista, verkkosivuista (Gandomi & Haider, 2015). Esim. sensoreita on eri laitteissa kohta satoja miljardeja ja niiden määrän ennustetaan kasvavan vuoteen 2030 mennessä noin 100 triljoonaan (Neittaanmäki, 2021).

Jaottelu ei kuitenkaan tee oikeutta datan monimuotoisuudelle. Parempi on puhua jatkumosta, jossa ääripäiden väliin mahtuu paljon välimuotoja (Salo, 2013).

Kolmea mainittua ulottuvuutta, nopeus, koko ja moninaisuus, on käytetty yleisesti kuvaamaan suurten, vaihtelevien ja nopeasti kasvavien datamäärien aiheuttamaa painetta. Mitä paremmin data täyttää nämä kolme ominaisuutta, sitä selkeämmin voidaan puhua big datasta (Salo, 2014).

Ajan myötä big datalle on määritelty lisää ominaisuuksia. Näitä ovat esimerkiksi arvo ja todenmukaisuus sekä vaihtelevuus ja monimutkaisuus (Gandomi & Haider, 2015). Tässä tutkielmassa käytetään big datan määrittelyssä Laneyn kolmen V:n ominaisuuksien lisäksi arvoa ja todenmukaisuutta.

Arvo (engl. value) on Oracle-yrityksen esittelemä ominaisuus big datalle. Heidän mukaansa big datalla on suhteellisen pieni arvo verrattuna datan määrään. Alkuperäisessä muodossaan oleva prosessoimaton data on suhteellisen arvotonta. Arvoa saadaan datalle enemmän, kun eri lähteistä saatua dataa yhdistellään ja analysoidaan. Data arvo konkretisoituu yritykselle siinä vaiheessa, kun dataa on muokattu ja yhdistelty siten, että sitä voidaan hyödyntää liiketoiminnassa (Gandomi & Haider, 2015).

Monet järjestelmät tuottavat niin suunnattomat määrät dataa, ettei pysyvä talteenotto ole edes mahdollista, saati tarpeellista. Big datan kohdalla havahdutaan tosiasiaan, että datassa piilee valtavasti arvoa, joka vain odottaa löytäjäänsä. Tämä edellyttää tilastomatemaattista osaamista sekä analysointityökalujen hallintaa (Salo, 2014).

Big datasta saatava hyöty nähdään suomalaisissa yrityksissä vielä rajoittuneesti. Hyöty suuntautuu enemmän sisäisen tehokkuuden parantamiseen kuin täysin uuden arvon tai uusien liiketoimintamallien luomiseen. Big datasta on mahdollista luoda uutta arvoa ja liiketoimintaa, mutta se on ennen kaikkea liiketoiminnan kehittämistä (Ylijoki, 2019).

Todenmukaisuus (engl. veracity) on IBM-yhtiön esittelemä termi datan epävarmuudelle. Todenmukaisuudella tarkoitetaan erilaisten datatyyppien luotettavuutta. Data voi olla puutteellista ja ristiriitaista. Siihen voi myös liittyä viivettä tai sen käsittely tuottaa arviointeja tai likiarvoja. Big datassakin tulisi kuitenkin tavoitella mahdollisimman korkeaa datan laatua. Tietyn tyyppisestä datasta ei kuitenkaan ole mahdollista puhdistaa sen luontaista epävarmuutta. Esimerkkejä tällaisesta datasta ovat sää ja asiakkaan ostospäätökset. Paljon epäluotettavaa dataa tuottavat sosiaalisen median jäsentymätön data sekä esineiden internet (Al-Jepoori, 2018).

Datan epävarmuus tulee siis ottaa aina huomioon, kun käsitellään big dataa. (Schroeck ym., 2012). Big dataan liittyy usein myös epävarmuutta datan laadun suhteen.

2.2 Datan laatu

Korkea datan laatu on yrityksille tärkeä voimavara, jonka avulla voidaan saada merkittävää kilpailuetua. Samalla huono datan laatu voi aiheuttaa yritykselle suuria ongelmia prosesseissa ja nostaa operatiivisia kustannuksia. Korkea datan

laatu voidaan määritellä monella tavalla eikä sille ole löydetty yhtä merkittävää määritelmää. Esimerkiksi korkea datan laatu voi tarkoittaa, että data vastaa käyttäjiensä odotuksia. Käyttäjiä voivat olla tietokoneet tai ihmiset (Sebastian-Coleman, 2012). Toinen laajasti hyväksytty korkea datan laatumääritelmä on se, että data on käyttäjien mielestä käyttöön sopivaa (Strong ym., 1997). Usein datan laatu jaetaan erilaisiin datan laatuominaisuuksiin, kuten virheettömyys, oikeellisuus ja ajantasaisuus. Datan laadun tutkimus pohjautuu monesti Juranin & Godfreyn (1999) määritykseen laadusta (Alshikhi & Abdullah, 2018).

Juran ja Godfrey (1999) kuvailevat laadulle kaksi tärkeää määritelmää. Laadulla tarkoitetaan tuotteen ominaisuuksia, jotka täyttävät asiakkaan vaatimukset näin kasvattaen asiakastyytyväisyyttä. Juranin toisen määritelmän mukaan laatu tarkoittaa myös vapautta puutteista. Nämä puutteet aiheuttavat työn uudelleen tekemistä, joka vähentää asiakastyytyväisyyttä. (Juran & Godfreay, 1999).

Wang ja Strong (1996) jaottelevat datan laatuominaisuudet neljään laatuulottuvuuteen: sisäiseen datan laatuun (engl. Intrinsic Data Quality), kontekstuaaliseen datan laatuun (engl. Contextual Data Quality), representationaaliseen datan laatuun (engl. Representational Data Quality) sekä datan saatavuuden laatuun (engl. Accessibility Data Quality).

Sisäisellä datan laadulla tarkoitetaan, että datassa itsessään on laatua. Oikeellisuudella tarkoitetaan, että datassa ei ole virheitä kuten puuttuvia arvoja tai kirjoitusvirheitä. Objektiivisuudella tarkoitetaan, että data on puolueetonta sekä ennakkoluulotonta. Käyttäjän näkökulmasta sisäisiin laatuominaisuuksiin kuuluvat myös maine sekä uskottavuus, koska ne ovat hyvin riippuvaisia oikeellisuudesta sekä objektiivisuudesta. (Wang & Strong, 1996).

Kontekstuaalisella datan laadulla tarkoitetaan, että dataa täytyy arvioida aina sen käyttötarkoituksen mukaan. Datan käyttötarkoituksen konteksti voi muuttua ajan myötä, mikä tekee korkean kontekstuaalisen datan laadun saavuttamisesta vaikeaa. Kontekstuaaliseen laatuun kuuluvat: lisäarvo, asiaankuuluvuus, ajantasaisuus, valmius sekä datan sopiva määrä. Datan lisäarvolla tarkoitetaan, kuinka hyödyllistä data on ja tuoko sen käyttö hyötyä käyttäjälleen (Wang & Strong, 1996). Datan asiaankuuluvuudella mitataan kuinka käytettävää sekä hyödyllistä data on kontekstiinsa nähden (Wang & Strong, 1996). Ajantasaisuudella taas tarkoitetaan sitä, missä määrin datan ikä on käyttötarkoitukseen sopivaa (Wang & Strong, 1996). Ajantasaisuus kuvaa myös, kuinka nopeasti tiedonmuutokset reaalimaailmassa näkyvät tietojärjestelmässä (Wand & Wang, 1996). Datan valmiudella mitataan sitä, että data on käyttötarkoitukseensa tarpeeksi laajaa ja syvällistä. Kuinka paljon datassa on virheitä kuten puuttuvia arvoja ja vääriä arvoja ja onko vaaditut arvot täytetty. Tietojärjestelmän rakenteen täytyy olla sellainen, että se pystyy datan avulla kuvaamaan reaalimaailman järjestelmää (Batini ym., 2009).. Dataa täytyy myös olla käyttötarkoitukseensa nähden sopiva määrä.

Representationaalinen datan laatu sisältää datan muotoon sekä datan merkitykseen liittyviä laatuominaisuuksia. Wang ja Strongin (1996) mukaan tähän laatu-ulottuvuuteen kuuluvat seuraavat laatuominaisuudet: tulkittavuus, ymmärrettävyys, yhdenmukaisuus sekä esittämisen tiiviys. Tulkittavuudella sekä

ymmärrettävyydellä tarkoitetaan, että datassa esiintyvät symbolit, yksiköt sekä määritykset ovat ymmärrettäviä. Yhdenmukaisuudella ja tiiviydellä tarkoitetaan, että datan muoto pysyy yhdenmukaisena läpi aineiston. Data tulee esittää myös tarpeeksi ytimekkäästi (Wang & Strong, 1996).

Viimeisenä laatu-ulottuvuutena Wang ja Strong määrittelivät saatavuuden. Tähän ulottuvuuteen he sisällyttivät saatavuuden sekä saatavuuden tietoturvan. Saatavuudella tarkoitetaan, että datan on käyttäjälleen helposti saatavilla sekä käytettävissä. Saatavuuden tietoturvalla tarkoitetaan, että pääsyä dataan on tarpeeksi rajoitettu, jotta ulkopuolisten eivät pääse dataan käsiksi. Nyky-yhteiskunnassa saatavuus ja tietoturva korostuvat, koska data tallennetaan sähköisesti tietokantoihin ja niitä käytetään tietokoneilla. Seuraavaksi käsitellään datan eri laatuongelmia.

TAULUKKO 1 Datan laatu-ulottuvuudet (Wang & Strong, 1996)

Laatu-ulottuvuus	Laatuominaisuudet
Sisäinen laatu	Uskottavuus Oikeellisuus Objektiivisuus Maine
Kontekstuaalinen laatu	Lisäarvo Asiaankuuluvuus Ajantasaisuus Valmius Datan sopiva määrä
Representationaalinen datan laatu	Tulkittavuus, Ymmärrettävyys Yhdenmukaisuus Esittämisen tiiviys
Datan saatavuuden laatu	Saatavuus Pääsyn tietoturva

2.3 Datan laatuongelmat

Kirjallisuudessa datan laatuongelmia on lajiteltu useasta eri näkökulmasta. Useammassa tutkimuksessa laatuongelmia on pyritty lajitelemaan hierarkkiseen järjestykseen. (Laranjeiro ym., 2015)

Rahm ja Don jakavat laatuongelmat neljään eri luokaan niiden alkuperän sekä instanssitason (engl. instance level) mukaan. Laatuongelmien alkuperää on määritelty sen mukaan, tuleeko data yhdestä tai useammasta lähteestä. Ilmentymätasot puolestaan on jaettu sen mukaan, ilmeneekö ongelma instanssi- vai skeemamatasolla. Yhdestä lähteestä tulevia skeematason ongelmia ovat esimerkiksi väärät tietotyypit tai yksilöivän kentän rajoitteiden rikkominen (engl. uniqueness vialation). Skeematason ongelmat johtuvat yleensä huonosta tietokannan

suunnittellusta tai liian vähäisistä kenttien yhtenäisyysrajoituksista (engl. integrity constraints). Instanssitasoon kuuluvat laatuongelmat, joita ei voi vähentää paremmalla tietokannan suunnitellulla. Näitä ovat esimerkiksi kirjoitusvirheet, puuttuvat arvot sekä epäselkeät ja kryptiset arvot. Instanssitason virheet voivat johtua datan lähteestä tai esimerkiksi käyttäjästä (Rahm & Do, 2000).

Datan tuleminen useammasta lähteestä pahentaa yksitäisestä lähteestä esiintyneitä laatuongelmia. Eri lähteistä tuleva data voi sisältää aikaisemmin mainittuja virheitä, data voi olla ristiriidassa keskenään tai datassa voi esiintyä päällekkäisyyksiä. Tämä johtuu yleensä siitä, että datalähteitä kehitetään ja ylläpidetään erillään toisistaan vastaamaan tiettyihin tarpeisiin. Esimerkkeinä usean lähteen laatuongelmista Rahm ja Don listaavat nimeämiskonfliktit ja rakenteelliset konfliktit. Nimeämiseen liittyviä ristiriitaisuuksia esiintyy, kun sama dataa sisältävä kenttä on nimetty kahdessa lähteessä eri tavalla, tai samoja nimiä on käytetty kuvaamaan eri dataa. Edelleen, jos esimerkiksi objektin rakenne on kahdessa lähteessä eri, tuottaa se rakenteellisia ristiriitaisuuksia (Rahm & Do, 2000).

Laranjeiro (2015) laajentaa tutkimuksessaan Rahm ja Don (2000) lajittelua datan laatuongelmista sijoittamalla laatuongelman tutkimuksessaan tunnistettuihin laatu-ulottuvuuksiin. (Taulukko 2). Tämä antaa kuvan laatuongelmien yhteydestä laatu-ulottuvuuksiin. Lajitteluun he sisällyttivät seuraavat laatu-ulottuvuudet: saatavuus, oikeellisuus, täydellisyys, yhtenäisyys sekä ajantasaisuus.

TAULUKKO 2 Laranjeiron (2015) luokittelu datan laatuongelmista Rahm ja Dohn (2000) luokittelun pohjalta

Ongelmatyypit		Datan laatuongelmat	Saatavuus	Virheettömyys	Täydellisyys	Yhdenmukaisuus	Ajantasaisuus
Lähde	Taso						
Yksittäinen	Instanssi	Puuttuva data		x	x		
		Virheellinen data		x			
		Kirjoitusvirheet		x			
		Epäselvä data	x				
		Asiaankuulumaton data	x			x	
		Vanhentunut data		x			x
		Arvot väärissä kentissä	x	x	x	x	
		Virheelliset viitteet		x			
		Duplikaatit	x				
	Skeema	Toimialueen rikkomus		x			
		Toiminnallisen riippuvuuden rikkomus		x			
		Väärä tietotyyppi	x				
		Viite-eheyden rikkomus	x	x			
		Kaksoisarvojen eston rikkomus					
Monta	Instanssi	Rakenteelliset ristiriidat	x			x	
		Erilaiset sanajärjestykset	x			x	
		Erilaiset aggregaatiotasot	x	x		x	
		Ajallinen yhteensopimattomuus		x		x	x
		Erilaiset yksiköt	x			x	
		Erilaiset esittämistavat	x			x	
	Skeema	Synonyymien käyttö	x				
		Homonyymien käyttö	x				
		Erikoismerkkien käyttö	x				
		Erilaiset merkistökoodaukset	x			x	

3 TUTKIMUKSEN TOTEUTUS

Tutkielmassa toteutettiin systemaattinen kirjallisuuskatsaus, jonka tarkoituksena oli selvittää big datan laatuongelmia käsittelevissä tutkimuksissa esiintyviä teemoja. Tutkimus toteutettiin kahdessa osassa. Ensimmäiseksi tutkimusainestoa etsittiin valituilla hakusanoilla IEEE Xplorer Standards -tietokannasta. Hakutuloksista suodatettiin pois otsikon perusteella tutkimusaiheeseen sopimattomat julkaisut. Tämän jälkeen valituista julkaisuista pyrittiin tiivistelmien pohjalta selvittämään tutkimuksissa esiintyviä teemoja.

Hakusanoiksi aineiston etsintään valikoitui seuraavat sanat: ”Big Data”, ”Data Quality”, ”Problem” sekä ”Issues”. Tietokantahaut toteutettiin marraskuun 2021 loppupuolella. Ensimmäisellä hakukerralla hakulausekkeessa käytettiin termejä ”Big Data”, ”Data Quality”, problem. Haku palautti tuloksiksi 74 osumaa (Taulukko 3). Toiseen hakuun ”problem” -sanan tilalle vaihdettiin sana ”issues”. Tämä palautti tuloksiksi 64 osumaa. Haussa ei käytettu tietokannan tarjoamia suodattimia osumien vähäisyyden takia. Alla olevassa taulukossa kolme on esitelty tietokantahakujen tulokset. Tulokset on myös jaoteltu julkaisun tyypin mukaan.

TAULUKKO 3 Hakulausekkeet ja niiden tulokset jaoteltuna tyypin mukaan

Hakulauseke	**Artikkelien yhteismäärä**	**Konferenssi-julkaisu**	**Tieteellinen aikakausijulkaisu**	**Ennakko-julkaisu**	**Tiedelehti**
”Big Data” ”Data Quality” problem	**74**	**57**	**15**	**1**	**1**
”Big Data” ”Data Quality” issue	**64**	**49**	**10**	**4**	**1**

Alustavasti aineiston hauen perusteella voidaan todeta, että tutkimusta datan laatuongelmista on vähän. Suurin osa löydetyistä julkaisuista on konferenssijulkaisuja. Haut myös palauttivat osittain samoja tuloksia. Tämän jälkeen hakutuloksista seulottiin otsikon ja julkaisupaikan perusteella. Hakutulos ohitettiin, mikäli oli selvää, ettei julkaisu vastannut tutkimuskysymykseen. Lopulta tarkempaa analysointia varten jäi 23 artikkelia. Valitut artikkelit on julkaistu vuosien 2015–2020 välisenä aikana. Taulukkoon 4 on listattu niiden artikkelien perustiedot, joista löydettiin teema.

TAULUKKO 4 Tarkempaan tarkasteluun valitut artikkelit

Otsikko	**Kirjoittajat**	**Sisältö**	**Vuosi**
Big data, big data quality problem	Becker, D., Dunn King, T., McMullen, B.	Esittelee neljän tapaustutkimuksen pohjalta, eroavatko big datan laatuongelmat normaalin datan laatuongelmaista.	2015
Sakdas: A Python Package for Data Profiling and Data Quality Auditing	Loetpipatwanich, S., Vichitthamaros, P.	Esittelee Python-paketin datan profilointia sekä auditointia varten	2020
A Study of Handling Missing Data Methods for Big Data	Ezzine, I., Benhlima, L.	Tarjoaa yleiskatsauksen joihinkin menetelmiin ja lähestymistapoihin puuttuvan datan käsittelyyn big datassa	2018
Data Preprocessing Method For The Analysis Of Incomplete Data On Students In Proverty	Huang, H., Wei, B., Dai, J., Ke, W.	Ehdottaa esikäsittelymenetelmää huonolaatuisen datan parantamiseksi.	2020
Big RDF data cleaning	Tang, N.	Käsittelee RDF Datassa esiintyviä datan laatuongelmia.	2015
Wind power generation forecasting and data quality improvement base on big data with multiple temporal-spatual scale	Qiao, L., Chen, S., Bo, J., Liu, S., Ma, G., Wang, H., Yang, J.	Tuulienergian kapasiteetin ennustamisen datan laadun parantaminen neuroverkon ja Newtonin interpolointifunktion avulla.	2019
Cluster-Based Best Match Scanning for Large-Scale Missing Data Imputation	Yu, W., Zhu, W., Liu, G., Kan, B., Zhao, T., Liu, H.	Uusi tehokkaampi algoritmi puuttuvien arvojen löytämiseen big datasta.	2017

(jatkuu)

Taulukko 4 (jatkuu)

Data Cleaning Optimization for Grain Big Data Processing using Task Merging	Ju, X., Lian, F., Zhang. Y.	Esittelee uuden optimointitekniikan, mikä perustuu tehtävien yhdistämiseen.	2019
An Automated Big Data Accuracy Assessment Tool	Mylavarapy, G., Thomas, J., Viswanathan, K.	Esittelee koneoppimista hyödyntävän datan oikeellisuuden arviointityökalun.	2019
A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends Towards Cleaning Dirty Data	Hossen, M., Goh, M., Hossen, A., Rahman, A.	Kirjallisuuskatsaus verkkoliiketoiminnassa esiintyviin big datan laatuongelmiin.	2020
Enhancing data quality by cleaning inconsistent big RDF data	Benbernou, S., Ouziri, M.	Käsittelee menetelmiä huonolaatuisen RDF datan puhdistamista varten	2017
Online anomaly detection over Big Data streams	Rettig, L., Khayati M., Cudr é-Mauroux, P., Piórkowski, M.	Käsittelee menetelmää, mikä helpottaa big datassa esiintyvien poikkeavuuksien löytämistä.	2015
Data quality issues in big data	Rao, D., Gudivada, V., Raghavan, V.	Tarjoaa näkökulmia datan laatuongelmiin big datan kontekstissa.	2015
The quality concerns in health care Big Data	Molinari, A., Nollo, G.	Käsittelee terveydenhuollon analytiikassa käytettävän big datan laatuongelmia.	2020
Data quality in big data processing: Issues, solutions, and open problems	Zhang, P., Xiong, F., Gao, J., Wang, J.	Esittelee big datan laatuongemia, jotka esiintyvät big datan prosessoinnin eri vaiheissa.	2017
A platform Solution of Data-Quality Improvement for Internet-of-Vehicle Services	Zhang, M., Wo, Z., Xie, T.	Esittelee alustatason ratkaisun datan laadun parantamiseen, mikä on suunniteltu varmistamaan palveluiden käyttövarmuus ajoneuvopalveluiden internetille.	2018

(jatkuu)

Taulukko 4 (jatkuu)

Big data and quality: A literature review	Lakshen, G., Vraneš, S., Janev, V.	Käsittelee big datan laatuongelmia sekä niiden tilaa.	2016
Big Data Validation and Quality Assurance – Issues, Challenges, and Needs	Gao, J., Xie, C., Tao, C.	Esittelee big datan validointia sekä laadun varmistamista sekä niihin liittyviä konsepteja ja prosesseja.	2016
Computing data quality indicators on Big Data streams using CEP	Yang, W., Da Silva, A., Picard M	Esittelee lähestymistavan älymittarien lähettämän datan laatuindikaatt	2015

3.1 Tulokset

Hakutulosten seulonnan jälkeen valittujen julkaisuiden tiivistelmistä pyrittiin päättelemään tutkimuksissa esiintyviä teemoja. Julkaisuista löydettiin yhteensä kahdeksan erilaista teemaa. Selvästi eniten tutkimuksia oli tehty teemasta datan laadun parantaminen. Näitä julkaisuja löydettiin yhteensä kahdeksan kappaletta. Datan laadun parantamista käsittelevät julkaisut käsittelivät laadun parantamista datan puhdistamisen näkökulmasta. Esimerkiksi Ezzine ja Behlima (2015) esittelevät tutkimuksessaan tapoja puuttuvien arvojen käsittelyyn big datassa.

Datan laadun parantamisen teemaan liittyy oleellisten myös teema laadun parantamisprosessin optimointi. Datan laadun parantamisen optimointia käsittelevää tutkimusta löydettiin kaksi kappaletta. Ju, Feiyu ja Zhang (2019) käsittelevät tutkimuksessaan datan puhdistamisen optimointia. Julkaisussaan he ehdottavat menetelmää, jonka avulla toistuvia tehtäviä voidaan vähentää ja näin nopeuttaa datan puhdistusprosessia (Ju ym., 2019).

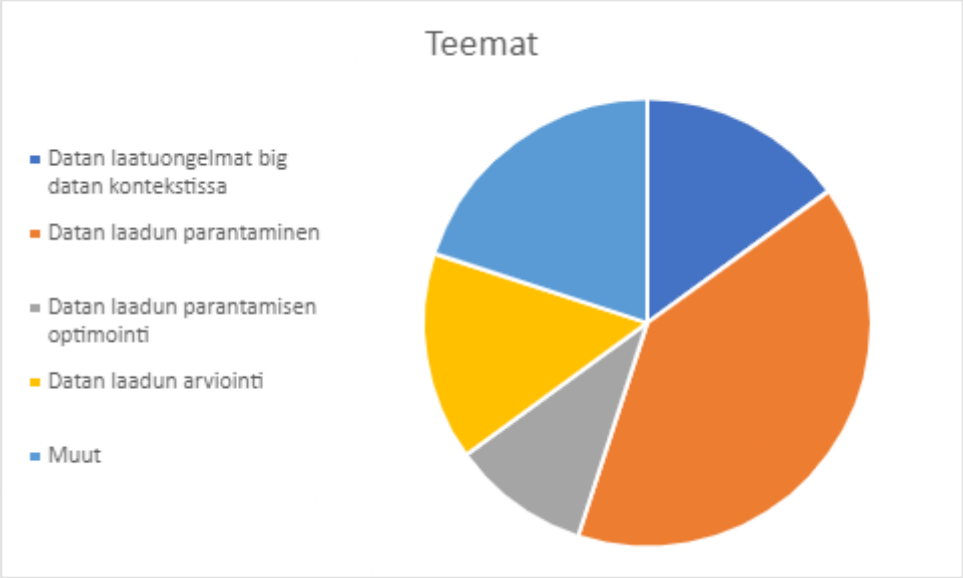
Datan laadun arviointi on kolmas teema. Laadun arviointia käsittelevää tutkimusta löydettiin yhteensä 3 kappaletta. Artikkelit esittelivät työkaluja ja malleja, joiden avulla datan laatua pystytään arvioimaan. Esimerkiksi Mylavarapu, Thomas ja Viswanathan ehdottavat datan oikeellisuuden arviointiin käytettävää työkalua, joka on kehitetty Apache Spark -analytiikkamoottorin päälle.

Neljännestä teemasta, datan laatuongelmat big datan kontekstisissa, käsitteleviä artikkeleja löydettiin kolme kappaletta. Artikkeleissa tavoitteena oli esitellä big datassa esiintyviä laatuongelmia. Eroavatko laatuongelmat perinteisessä datassa esiintyvistä laatuongelmista?

Julkaisuista löydettiin myös teemoja, joita oli käsitelty vain yhdessä artikkelissa. Näitä olivat big datan laadun varmistaminen, big datan laadun

auditointi, big datan laadun haasteet terveydenhuollossa sekä big datan laadun hallinta. Kolme viimeisintä artikkelia on julkaistu vuonna 2020, joten voidaan olettaa, että aikaisempaa tutkimusta ei teemojen osalta ole tehty.

Kolme artikkelia hylättiin julkaisujen tarkemmassa tarkastelussa. Kaksi julkaisua ei tiivistelmän mukaan käsitellyt datan laatua tai big dataa. Yhden julkaisun tiivistelmästä ei selvinnyt tutkimuksen teema.
Kuvioon 1 on kuvattu eri teemojen osuuksia löydettyjen teemojen kokonaismäärästä. Teemat, joita käsittelevää tutkimusta oli vain yksi kappale, on koottu ”Muut” osuuden alle. Näitä ovat big datan laadun varmistaminen, big datan laadun auditointi, big datan laadun haasteet terveydenhuollossa sekä big datan laadunhallinta.



KUVIO 1 Eri teemojen osuudet löydetyistä teemoista

3.2 Pohdinta

Big dataan liittyvät tutkimukset ovat lähestyneet aihetta valtaosin teknisestä näkökulmasta. Tämä on ymmärrettävää, sillä big datan tulee olla hyödynnettävässä muodossa ja laadultaan riittävän hyvää. Tämä taas edellyttää osaamista datan keräämisessä. Tarvitaan analysointitaitoja ja algoritmista ajattelua (Ylijoki, 2019). Big datan ominaisuudet kuten määrä, nopeus ja erityisesti moninaisuus tuovat omat haasteensa big datan laadun varmistamiseen. Datan laatu on yksi merkittävistä big datan hyödyntämisessä kohdatuista haasteista. Kirjallisuuskatsauksen tuloksista huomataan, että suurin osa big datan laatuongelmia käsittelevästä tutkimuksesta pyrkii tuomaan ratkaisuja siihen, miten datan laatua voidaan parantaa tehokkaasti.
Datan laadun parantamista käsittelevät julkaisut käsittelivät laadun parantamista datan puhdistamisen näkökulmasta sekä laadun parantamisprosessin

optimointia. Kelleher ja Tierney (2021) painottavat, miten tärkeää on varmistaa, että käytetty aineisto on hyvälaatuista, sillä tehdyn selvityksen mukaan big dataa hyödyntävissä projekteissa noin 80 prosenttia työajasta kuluu tietoaineiston luomiseen, puhdistamiseen ja päivittämiseen. Onkin varmistettava, että datan käytölle on selkeästi määritelty tavoite, käytetty aineisto on tarkoitukseen sopivaa ja sitä on riittävä määrä (Kelleher & Tierney, 2021). Monet lähteet painottavat korkean datan laadun merkitystä big datan käytössä. Big datan tuoma suuri datan määrän hyödyntäminen ja analysointi mahdollistavat palvelun laadun parantamisen, asiakkaiden paremman ymmärtämisen sekä riskien ennustamisen ja estämisen. Kuitenkin datan analysoinnin täytyy perustua hyvälaatuiseen dataan. Huonolaatuinen data aiheuttaa vähäistä datan hyödyntämistä sekä voi aiheuttaa merkittäviä virheitä päätöksenteossa (Cai, 2015).

Datan laadun arviointia käsittelevät tutkimukset esittelivät työkaluja ja malleja, joiden avulla datan laatua pystytään arvioimaan. Laadun arvioinnilla varmistetaan, että data on käyttökohteeseen sopivaa mm. yhtenäistä ja oikeellista. Laadun arviointityöllä on merkitystä, sillä sen avulla pystytään vähentämään aineiston valmisteluun käytettävää työmäärää.

Datan laatuongelmia big datan kontekstissa käsitteleviä tutkimuksissa tavoitteena oli esitellä big datassa esiintyviä laatuongelmia. Eroavatko laatuongelmat normaalikokoisessa datassa esiintyvistä laatuongelmista? Beckerin, Dunn Kingin ja McMullenin (2015) tutkimuksessa ilmeni, että big datan laatuongelmat eivät merkittävästi eroa perinteisistä datan laatuongelmista.

Muita tutkimuksessa löydettyjä teemoja ei ole tutkittu kuin muutamissa julkaisuissa. Tämä johtunee siitä, että big data on teknologiana vielä suhteellisen uusi. Kirjallisuuskatsauksessa toteutetussa tietokantahauista huomataan, että tutkimus laatuongelmista on vielä vähäistä.

Yritysten sovellukset voivat olla huonosti suunniteltuja, aineistomallit huonosti toteutettuja eikä henkilökuntaa ole koulutettu varmistamaan, että käytetään hyvää aineistoa. Lukemattomat tekijät voivatkin tuoda huonolaatuista aineistoa järjestelmiin. Kelleherin ja Tierneyn mukaan hyvälaatuisen aineiston tarve on niin tärkeää, että yritykset ovat palkanneet työntekijöitä, jotka jatkuvasti tarkastavat aineistoa, arvioivat sen laatua sekä esittävät parannusehdotuksia syötetyn aineiston laadun parantamiseksi (Kelleher & Tierney, 2021).

4 YHTEENVETO

Tässä tutkielmassa pyrittiin selvittämään, millaista tutkimusta big datan laatuongelmista on tehty. Tutkielma toteutettiin systemaattisena kirjallisuuskatsauksena. Tutkimuksen vähäisyyden takia lähdeaineiston haku rajattiin IEEE Xplorer Standards -tietokantaan. Löydetyn aineiston avulla pyrittiin vastaamaan seuraavaan tutkimuskysymykseen.

- Millaisia teemoja esiintyy big datan laatuongelmia käsittelevässä tutkimuksessa.

Ensimmäisessä sisältöluvussa käydään läpi tutkielmassa esiintyvät keskeiset käsitteet. Luku jaettiin kolmen alalukuun, joista jokainen käsitteli kirjallisuuden pohjalta yhtä keskeistä käsitettä. Käydyt käsitteet olivat big data, datan laatu sekä datan laatuongelmat. Ensimmäisessä alaluvussa käytiin läpi, mitä on big data ja esiteltiin sen keskeiset ominaisuudet. Seuraavaksi esiteltiin, mitä tarkoitetaan datan laadulla. Lopuksi käytiin läpi datassa esiintyviä laatuongelmia.

Toisessa sisältöluvussa käsiteltiin tutkimuksen toteutusta ja tuloksia. Tutkimus toteutettiin kahdessa osassa. Luku on jaettu kolmeen alalukuun. Ensimmäisessä alaluvussa esitellään aineiston hakuprosessi sekä tietokantahakujen tulokset. Toisessa alaluvussa esitellään tutkimuksen tulokset. Viimeisessä alaluvussa pohditaan, mitä johtopäätöksiä voidaan tehdä löydettyjen tulosten perusteella. Ensimmäiseksi aineistoa haettiin valitusta tietokannasta ja siitä suodatettiin tarkempaan tarkasteluun tutkimuksen kannalta oleellisimmat julkaisut. Tarkempaan tarkasteluun valittiin yhteensä 23 julkaisua. Seuraavaksi valittujen tutkimusten tiivistelmistä pyrittiin kartoittamaan tutkimuksissa esiintyviä teemoja.

Valituista julkaisuista löydettiin yhteensä kahdeksan erilaista teemaa. Eniten artikkeleissa käsiteltiin datan laadun parantamista. Tätä teemaa käsitteli yhteensä kahdeksan artikkelia. Seuraavaksi eniten käsiteltiin datan laadun parantamista sekä datan laatuongelmia big datan kontekstissa. Kumpaakin aihetta käsitteli kolme artikkelia. Kolmanneksi eniten käsiteltiin datan laadun parantamisen optimointia. Tätä aihetta käsitteli kaksi artikkelia. Näiden teemojen lisäksi löydettiin neljä erilaista teemaa, joita oli käsitelty vain yhdessä artikkelissa. Niitä

ovat big datan laadun varmistaminen, big datan laadun auditointi, big datan laadun haasteet terveydenhuollossa sekä big datan laadunhallinta.

Tutkimuksen suurimpana rajoitteena ja haasteena oli big datan laatuongelmia ja datan laatuongelmia käsittelevän tutkimuksen vähäinen määrä. Tämä vaikeutti lähteiden löytämistä joihinkin osioihin. Big data on käsitteenä uusi, joten voidaan myös olettaa, että tutkimusta aiheesta ei ole vielä ehditty tekemään. Big data on alkanut kerätä suosiota vasta vuonna 2011 (Gandomi & Haider, 2015). Tutkimuksessa tarkempaan tarkasteluun valitut artikkelit oli julkaistu vuosina 2015–2020.

Tutkimuksen vähäisen määrän takia jatkotutkimusta selvästi tarvitaan kaikkiin tutkielmassa löydettyihin teemoihin. Aikaisempi tutkimus on eniten käsitellyt laadun parantamista, joten tutkimusta olisi hyvä enemmän painottaa vähemmän tutkittuihin teemoihin kuten big datan laadunhallintaan sekä big datan laadun varmistamiseen.

LÄHTEET

Al-Jepoori, M., & Al-Khanjari, Z. (2018). Framework for handling data veracity in big data. International Journal of Computer Science and Software Engineering, 7(6), 138–141.

Alshikhi, O. A., & Abdullah, B. M. (2018). Information quality: Definitions, measurement, dimensions, and relationship with decision making. European Journal of Business and Innovation Research, 6(5), 36–42.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52. https://doi.org/10.1145/1541880.1541883

Becker, D., King, T. D., & McMullen, B. (2015). Big data, big data quality problem. 2015 IEEE International Conference on Big Data (Big Data), 2644–2653. https://doi.org/10.1109/BigData.2015.7364064

Benbernou, S., & Ouziri, M. (2017). Enhancing data quality by cleaning inconsistent big RDF data. 2017 IEEE International Conference on Big Data (Big Data), 74–79. https://doi.org/10.1109/BigData.2017.8257913

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(0), 2. https://doi.org/10.5334/dsj-2015-002

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. Proceedings. Visualization '97 (Cat. No. 97CB36155), 235-244,. https://doi.org/10.1109/VISUAL.1997.663888

Ezzine, I., & Benhlima, L. (2018). A Study of Handling Missing Data Methods for Big Data. 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), 498–501. https://doi.org/10.1109/CIST.2018.8596389

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gao, J., Xie, C., & Tao, C. (2016). Big Data Validation and Quality Assurance—Issuses, Challenges, and Needs. 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE), 433–441. https://doi.org/10.1109/SOSE.2016.63

Hossen, M. I., Goh, M., Hossen, A., & Rahman, Md. A. (2020). A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends Towards Cleaning Dirty Data. 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), 209–213. https://doi.org/10.1109/ICSGRC49013.2020.9232648

Huang, H., Wei, B., Dai, J., & Ke, W. (2020). Data Preprocessing Method For The Analysis Of Incomplete Data On Students In Poverty. 2020 16th International Conference on Computational Intelligence and Security (CIS), 248–252. https://doi.org/10.1109/CIS52066.2020.00060

Ju, X., Lian, F., & Zhang, Y. (2019). Data Cleaning Optimization for Grain Big Data Processing using Task Merging. 2019 6th International Conference on Information Science and Control Engineering (ICISCE), 225–233. https://doi.org/10.1109/ICISCE48695.2019.00053

Juran, J. M., & Godfrey, A. B. (Toim.). (1999). Juran's quality handbook (5th ed). McGraw Hill.

Kelleher, J. D., & Tierney, B. (2018). Data science. The MIT Press.

Lakshen, G. A., Vranes, S., & Janev, V. (2016). Big data and quality: A literature review. 2016 24th Telecommunications Forum (TELFOR), 1–4. https://doi.org/10.1109/TELFOR.2016.7818902

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A Survey on Data Quality: Classifying Poor Data. 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), 179–188. https://doi.org/10.1109/PRDC.2015.41

Lehto, M., & Neittaanmäki, P. (2016). Big datan ja data-analyysin tutkimus ja opetus vahvistavat kansallista digiloikkaa. Futura 35 (2016): 2.

Loetpipatwanich, S., & Vichitthamaros, P. (2020). Sakdas: A Python Package for Data Profiling and Data Quality Auditing. 2020 1st International Conference on Big Data Analytics and Practices (IBDAP), 1–4. https://doi.org/10.1109/IBDAP50342.2020.9245455

Molinari, A., & Nollo, G. (2020). The quality concerns in health care Big Data. 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), 302–305. https://doi.org/10.1109/MELECON48756.2020.9140534

Mylavarapu, G., Thomas, J. P., & Viswanathan, K. A. (2019). An Automated Big Data Accuracy Assessment Tool. 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 193–197. https://doi.org/10.1109/ICBDA.2019.8713218

Neittaanmäki, P., Lehto, M., & Savonen, M. (2021). Yhteiskunnan digimurros.

Qiao, L., Chen, S., Bo, J., Liu, S., Ma, G., Wang, H., & Yang, J. (2019). Wind power generation forecasting and data quality improvement based on big data with multiple temporal-spatual scale. 2019 IEEE International Conference on Energy Internet (ICEI), 554–559. https://doi.org/10.1109/ICEI.2019.00104

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3–13.

Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. 2015 IEEE International Conference on Big Data (Big Data), 2654–2660. https://doi.org/10.1109/BigData.2015.7364065

Rettig, L., Khayati, M., Cudre-Mauroux, P., & Piorkowski, M. (2015). Online anomaly detection over Big Data streams. 2015 IEEE International Conference on Big Data (Big Data), 1113–1122. https://doi.org/10.1109/BigData.2015.7363865

Salo, I. (2014). Big data & pilvipalvelut. Jyväskylä: Docendo.

Schroeck, M., Shockley, R., Smart, J., Romero Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report.

Sebastian-Coleman, L. (2012). Measuring data quality for ongoing improvement: A data quality assessment framework. Newnes.

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

Tang, N. (2015). Big RDF data cleaning. 2015 31st IEEE International Conference on Data Engineering Workshops, 77–79. https://doi.org/10.1109/ICDEW.2015.7129549

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95.

Wang, R. Y., & 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

Weiss, S. M., & Indurkhya, N. (1998). Predictive data mining: A practical guide. Morgan Kaufmann Publishers.

Wenlu Yang, Da Silva, A., & Picard, M.-L. (2015). Computing data quality indicators on Big Data streams using a CEP. 2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), 1–5. https://doi.org/10.1109/IWCIM.2015.7347061

Ylijoki, O. (2019). Big Data–Towards Data-driven Business.

Yu, W., Zhu, W., Liu, G., Kan, B., Zhao, T., & Liu, H. (2017). Cluster-Based Best Match Scanning for Large-Scale Missing Data Imputation. 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), 232–238. https://doi.org/10.1109/BIGCOM.2017.48

Zhang, M., Wo, T., & Xie, T. (2018). A Platform Solution of Data-Quality Improvement for Internet-of-Vehicle Services. 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom), 1–7. https://doi.org/10.1109/PERCOM.2018.8444581

Zhang, P., Xiong, F., Gao, J., & Wang, J. (2017). Data quality in big data processing: Issues, solutions and open problems. 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 1–7. https://doi.org/10.1109/UIC-ATC.2017.8397554