METADATAN LAATU JULKAISUARKISTOISSA

Annukka Ruotsalainen

Tampereen yliopisto
Informaatiotieteiden yksikkö
Informaatiotutkimus ja
interaktiivinen media
Pro gradu -tutkielma
Marraskuu 2016

TAMPEREEN YLIOPISTO, Informaatiotieteiden yksikkö
Informaatiotutkimus ja interaktiivinen media
RUOTSALAINEN, ANNUKKA: Metadatan laatu julkaisuarkistoissa
Pro gradu -tutkielma, 67 s., 4 liites.
Marraskuu 2016

Julkaisuarkistot ovat tieteen avoimuuden kannalta olennaisia järjestelmiä, joiden kautta organisaatiossa tuotettuja aineistoja voidaan tarjota vapaasti saataville. Metadatan – aineistoa kuvailevan tiedon – avulla julkaisuarkistojen aineistoja haetaan ja hallinnoidaan. Metadata mahdollistaa myös järjestelmien yhteentoimivuuden ja aineistojen uudelleenkäytön. Julkaisuarkistojen toimivuus ja hyödyllisyys on siis osaltaan riippuvaista metadatasta.

Tutkielmassa selvitetään, millaista on suomalaisten julkaisuarkistojen opinnäytetöiden metadatan laatu. Metadatan laatu on moniulotteista ja siihen vaikuttaa useampi tekijä. Laadukas metadata on muun muassa johdonmukaista, tarkkaa ja antaa kuvailemastaan kohteesta tietoa kattavasti. Laatua mitataan täydellisyyden ja painotetun täydellisyyden näkökulmasta. Mielenkiinnon kohteena on se, kuinka kattavasti kuvailukenttiä on metadatatietueissa käytetty.

Analyysimenetelmänä tutkielmassa käytetään kuvaileva tilastoanalyysia, ja tutkimusote on kvantitatiivinen. Laatua mitataan automaattisin menetelmin, jolloin tutkimukseen kohteeksi voidaan ottaa kaikki opinnäytetöiden metadata. Tutkimusaineisto koostuu 14 julkaisuarkiston 209407 metadatatietueesta. Tarkoituksena ei ole etsiä syitä metadatan laadun hyvään tai huonoon tasoon, vaan kuvata metadatan laatua tietyllä hetkellä kussakin julkaisuarkistossa. Tutkielmassa myös testataan, onko metadatan laadulla yhteyttä metadatatietueen tallennusvuoteen tai kuvailemansa opinnäytetyön tasoon.

Tulokset osoittavat, että julkaisuarkistojen opinnäytetöiden metadatan laatu vaihtelee hyvin paljon paitsi julkaisuarkistojen sisällä myös eri arkistojen välillä. Kuvailukenttien käytössä on siis paljon hajontaa, mikä viittaisi siihen, että metadatan käyttöä ohjaavaa standardia ei ole noudatettu johdonmukaisesti. Tuloksien perusteella vaikuttaisi siltä, että metadatan laatu on yhteydessä siihen, minkä tasoista opinnäytetyötä se kuvailee.

Avainsanat: metadata, laatu, laadunarviointi, julkaisuarkistot, Dublin Core

Sisällysluettelo

1 JOHDANTO....1

2 METADATATUTKIMUS....3

2.1 Laadun analysointi osana metadatatutkimusta....3
2.2 Metadata....5
2.3 Metadatan laatu....6
2.4 Yhteenvetoa laatumalleista....8

3 METADATAN LAADUN MITTAAMINEN....11

3.1 Manuaalinen tutkimus....12
3.2 Automaattinen tutkimus....15
3.3 Metadatastandardin vaikutus laatuun....20
3.4 Yhteenvetoa tutkimuksista....22

4 TUTKIMUSASETELMA- JA MENETELMÄT....24

4.1 Julkaisuarkistot....24
4.2 Aineiston keruu ja aineiston kuvaus....26
4.3 Tutkimusmenetelmänä kuvaileva tilastoanalyysi....30
4.4 Metadatan laadun mittaaminen....32

5 JULKAISUARKISTOJEN METADATAN LAATU....37

5.1 Metadatan laatu täydellisyyden näkökulmasta....37
5.2 Metadatan laatu painotetun täydellisyyden näkökulmasta....41
5.3 Metadatan laatu suhteessa aikaan ja opinnäytteen tasoon....50

6 JOHTOPÄÄTÖKSET....60

Lähteet ….....65

LIITE 1 Sähköpostikysely julkaisuarkistoille
LIITE 2 Julkaisuarkistojen kuvailukenttien vastaavuudet
LIITE 3 Yhtenäistetyt opinnäytetasot

1 JOHDANTO

Avoimuus on tieteen ja tutkimuksen keskeinen periaate, joka on noussut merkittäväksi tavaksi edistää tieteen vaikuttavuutta yhteiskunnassa. Avoimuus antaa mahdollisuuden tutkimustulosten entistä laajempaan todentamiseen, läpinäkyvyyteen ja toistettavuuteen. Avoimuuden periaate kattaa koko tutkimusprosessin: niin tutkimustulosten kuin -aineistojenkin pitäisi avoimen tieteen periaatteiden mukaan olla kaikkien saatavilla ja käytettävissä myös pitkällä aikavälillä. (Avoin tiede ja tutkimus 2016).

Julkaisuarkistot ovat yksi väline tieteellisen tiedon avoimen saatavuuden mahdollistamiseksi. Julkaisuarkistot ovat järjestelmiä, joiden tarkoituksena on aineistojen avoimen saatavuuden kautta edistää ja tukea tiedon jakamista ja uudelleenkäyttöä. Avoimien arkistojen todellinen vahvuus on mahdollisuus niihin tallennettujen aineistojen yhdistelyyn. Puhutaan järjestelmien yhteentoimivuudesta (interoperability), niiden kyvystä niin sanotusti kommunikoida keskenään. Mahdollisuus metadatan – aineistoa kuvailevan tiedon – vapaaseen haravointiin ja käyttöön uudessa kontekstissa, on yhteentoimivuuden keskeinen edellytys. (Confederation of Open Access Repositories 2011.)

Metadatan merkityksen korostuessa nousee kysymys sen laadusta tärkeäksi seikaksi. Voidaan sanoa, että metadatan laatu korreloi vahvasti sen kanssa, kuinka hyödyllisiä erilaiset järjestelmät käyttäjilleen ovat. Järjestelmiin tallennettua tietoa haetaan ja hallitaan metadatan avulla, joten metadatan laatu vaikuttaa osaltaan koko järjestelmän käytettävyyteen. Jos metadata on laadultaan huonoa, voi se pahimmassa tapauksessa kokonaan estää paitsi aineistojen löytymisen myös järjestelmien sujuvan käytön ja yhteentoimivuuden. Ei siis ole sama, onko avoimiin arkistoihin tallennettu aineisto kuvailtu monipuolisesti vai ei, millaista metadata on muodoltaan tai miten valitun metadatastandardin suosituksia on noudatettu. Olisikin tärkeää, että järjestelmien hallinnoijat olisivat tietoisia, millaista heidän aineistoihinsa liittyvä metadata on laadultaan.

Tässä tutkielmassa selvitetään, kuinka laadukasta suomalaisten julkaisuarkistojen metadata on. Julkaisuarkistoissa metadatan pitäisi olla mahdollisimman yhdenmukaista ja standardeja noudattavaa, jotta aineistoja voitaisiin käyttää myös uusien järjestelmien ja palvelujen muodostamiseen ja jotta tiedon löytyminen olisi sujuvaa. Avoin tiede ja tutkimus -sivustolla (2016) suositellaan julkaisuarkiston metadatan osalta esimerkiksi, että metadatan pitää olla riittävän yksityiskohtaista ja laajaa, aineiston tyyppi täytyy käydä ilmi ja metadatan pitää perustua yhteisesti määriteltyyn ontologiaan tai sanastoon ja aineistojen yhteydessä pitäisi hyödyntää pysyviä tunnisteita.

Tutkielman mielenkiinnon kohteena on, kuinka kattavasti kuvailukenttiä on arkistoihin tallennetuissa opinnäytetöiden metadatatietueissa käytetty. Kuvailukenttien käyttö kertoo metadatan laadun osalta sen yksityiskohtaisuudesta ja täydellisyydestä. Mitä enemmän kuvailukenttiä metadatatietueessa on, sitä monipuolisemmin ja kattavammin metadatan voidaan olettaa antavan tietoa kuvailemastaan aineistosta. Metadatan laatua mitataan automaattisin menetelmin, jolloin tutkimuksen kohteeksi voidaan ottaa kaikki julkaisuarkistoihin tallennetut metadatatietueet.

Tutkielman toisessa luvussa esitellään metadatatutkimuksen eri ulottuvuuksia keskittyen erityisesti metadatan laadun tutkimukseen sekä määritellään tutkielman kannalta olennaiset käsitteet. Kolmannessa luvussa perehdytään metadatan laatua mittaaviin tutkimuksiin, joista on erotettavissa kaksi päälinjaa: laadun mittaus manuaalisesti ja laadun mittaus hyödyntäen automaattisia menetelmiä. Luvussa myös luodaan katsaus tutkimuksiin, joissa on selvitetty, miten valittu metadatastandardi vaikuttaa metadatan laatuun.

Neljäs luku keskittyy tutkimusasetelman- ja menetelmien avaamiseen. Luvussa esitellään tutkimusympäristö eli julkaisuarkistot, kuvataan tutkimusaineiston lataaminen avoimen rajapinnan kautta ja esitellään aineisto tarkemmin. Luvussa myös esitetään katsaus tilastolliseen tutkimuksen keskeisiin periaatteisiin ja kuvataan, miten tässä tutkielmassa metadatan laatua mitataan ja analysoidaan.

Viidessä luvussa esitellään julkaisuarkistojen opinnäytetöiden metadatan laadun mittauksessa esiin nousseet tulokset. Viimeisessä luvussa esitetään johtopäätökset metadatan laadun mittauksessa esiin nousseista seikoista.

2 METADATATUTKIMUS

Tässä luvussa esitellään metadatatutkimuksen yleisiä linjoja ja erityisesti laadun arviointia osana sitä. Luvussa paneudutaan myös tämän tutkielman keskeisiin käsitteisiin metadataan ja metadatan laatuun.

2.1 Laadun analysointi osana metadatatutkimusta

Metadataa on tutkittu monesta näkökulmasta ja monessa eri kontekstissa. Tutkimusta on tehty pääasiassa informaatiotutkimuksen ja tietojenkäsittelytieteen aloilla. Näkökulmia metadatan tutkimuksessa on useita. Sicilian (2014) listauksen mukaan metadataa on tutkittu ainakin seuraavista näkökulmista: digitaaliset kirjastot ja arkistot sekä niiden aineistojen metadata, webbi ja metadata, metadastandardit tai -skeemat ja niiden soveltuvuus eri konteksteihin, metadatan luomisen eri menetelmät, metadatan laatu, metadatan vaikutus järjestelmien yhteentoimivuuteen, metadatan rooli ontologioissa ja yhteys semanttiseen webiin, metadatan merkitys linkitetyn avoimen datan muodostamisessa. (Sicilia 2014, 3–4.)

Moulaisonin ja muiden (2012) esitelmässä mainitaan metadatatutkimuksen keskeiseksi kohteeksi tällä hetkellä linkitetyn avoimen datan tuomat haasteet metadatan luomiselle ja käytölle. Tähän liittyen metadatan laadun mittaamisen eri menetelmät, metadatan luominen automaattisesti ja tiedon jakaminen sekä metadataskeemoihin liittyvä tutkimus on kirjoittajien mukaan ajankohtaista ja relevanttia. (Moulaison ym. 2012.)

Metadatan laadun analysoiminen ja mittaaminen on siis yksi näkökulma metadatan tutkimiseen, eikä sen merkitys ole ainakaan vähentynyt viime vuosina. Metadatan laatu vaikuttaa kaikkiin niihin toiminnallisuuksiin, joissa metadata on osallisena. Esimerkiksi käy linkitetty avoin data: Hoolandin ja Verborghin (2014) mielestä linkitetyn datan myötä metadatan laatu on noussut parrasvaloihin ja saanut sille kuuluvan huomion. Epäjohdonmukaisen, muodottoman ja rakenteettoman metadatan epäsuotavat vaikutukset ovat suuria ja laaja-alaisia. (Van Hooland & Verborgh 2014, 71.)

Metadatatutkimuksessa aihetta lähestytään metadatan laatua yleisesti käsitteellistävän mallin avulla, jonka pohjalta luodaan erilaisia laadun mittaamisen tapoja. Näiden avulla metadatan laadusta tietyssä kontekstissa pyritään saamaan kokonaiskuva: millaista tämän järjestelmän sisältämä metadata on laadultaan. Tutkimusta on pidetty toisaalta liian teoreettisena ja toisaalta taas liian kapea-alaisena. Hooland ja Verborgh kirjoittavat:

> "Many publications develop theoretical models and typologies (or present critiques and extensions of existing ones), without making it explicit how they can be put into practice to actually help you enhance your metadata." (Van Hooland & Verborgh 2014, 72.)

Metadatan laatua esittelevät mallit voivat Hoolandin ja Verborghin mielestä olla hyödyllisiä, koska ne tarjoavat paremman ymmärryksen metadatan laadusta. Mutta mallien pitäisi pystyä kertomaan myös, mitä käytännön hyötyä niistä on. Kuitenkin tässä teorian ja käytännön yhteensovittamisessa metadatan laadun tutkimus on kirjoittajien mielestä epäonnistunut. Sicilia (2014) perää puolestaan tutkimuksen monipuolistamista. Hänen mielestään metadatan laadun tutkimus on keskittynyt lähinnä metadatan täydellisyyden arviointiin ja jossain määrin myös asiasanastojen käyttöön. Sicilian mukaan tutkimuksen pitäisi kuitenkin kiinnittää huomiota myös metadatan välittämän tiedon runsauteen eli siihen missä määrin metadata välittää hyödyllistä tietoa sekä siihen, miten metadatan avulla mahdollistetaan kokoelmien linkittyminen toisiinsa. (Sicilia 2014, 5.)

Diane Hillmannin mielestä metadatan laatuanalyysin pitäisi pystyä vastaamaan ainakin kysymyksiin: mitä metadatakenttiä on käytetty; kuinka suuressa osassa tietueista tiettyjä kenttiä on käytetty; kuinka johdonmukaista metadata on; millaisia kaavoja tai rakenteita metadatan käytössä voidaan havaita (Hillmann 2008).

Miksi metadatassa on puutteita ja ongelmia? Tähän syynä on usempi tekijä. Metadatan laadukkuuteen vaikuttaa niin järjestelmän käyttäjien tarpeet ja heidän tapansa käyttää järjestelmää kuin itse järjestelmänkin piirteet. Metadatan laatu on siis mitä suurimmassa määrin kontekstisidonnaista: yhdessä yhteydessä laadukkaaksi katsottu metadata voi olla toisessa yhteydessä huonolaatuista. Laadukas metadata kuvaa aineistoa monipuolisesti, mikä tarkoittaa, että kuvailussa käytetään kuvailuelementtejä kattavasti. Kuitenkin käytännön kuvailutyössä joudutaan usein valitsemaan, kuvataanko paljon aineistoja, mutta ei niin yksityiskohtaisesti vai kuvataanko vähemmän aineistoja monipuolisesti ja tarkasti. Tähän valintaan ovat syynä rajalliset resurssit: yksityiskohtaisen ja ymmärrettävän metadatan luominen on kallista asiantuntijatyötä, joka vie aikaa. Laatua heikentävät myös metadatan epätarkkuudet. Esimerkiksi ihmistyölle luonnolliset kirjoitusvirheet vaikuttavat metadatan laatuun. Metadatan johdonmukaisuutta taas vähentää kuvailuelementtien ja arvojen epäyhtenäinen käyttö järjestelmän eri aineistojen välillä. (Hider 2012, 77–81.)

Van Hoolandin ja Verborghin (2014) kritisoimia metadatan laadun teoreettisia malleja voidaan kuitenkin soveltaa käytäntöön, ja näin on tutkimuksessa myös tehty. Erilaisia laadun mittaamisen tapoja on kehitelty ja testattu olemassa oleviin aineistoihin. Metadatan laatu ja sen analysointi on myös tämän tutkielman näkökulma metadatan tutkimiseen. Laatua tutkitaan ja analysoidaan julkaisuarkistojen kontekstissa.

Julkaisuarkistojen metadatan laatua on tutkittu hyvin vähän, vaikka nimenomaan julkaisuarkisto järjestelmänä ja avoimena aineistojen julkaisuympäristönä tarjoaa tutkimukselle metadatan laadun näkökulmasta mielenkiintoisia aspekteja. Avoimen tieteen periaatteiden mukaan julkaisuarkistojen metadataa pitäisi pystyä vapaasti käyttämään ja yhdistelemään esimerkiksi uusissa hakuportaaleissa. Julkaisuarkistojen aineistojen pitäisi myös olla avoimesti haettavissa ja saavutettavissa. Jotta nämä periaatteet toteutuisivat, ei metadatan laadussa saisi olla ongelmia, sillä aineistojen hallinnointi ja haku tapahtuu metadatan avulla.

2.2 Metadata

Metadata – tietoa tiedosta – on rakenteellista, koodattua dataa, joka kuvailee aineiston kontekstia, sisältöä ja rakennetta (Greenberg 2005). Metadatan tehtävä on mahdollistaa järjestelmään tallennetun aineiston hakeminen, tunnistaminen, valitseminen ja paikallistaminen. Se on siis olennaista aineiston tehokkaan käytön, löytämisen ja hallinnoinnin kannalta. (Hider 2012, 18.) Termiä metadata käytetään yleensä puhuttaessa digitaalisten aineistojen kuvailutiedoista; rakenteellisella tiedolla viitataan siihen, että data on tietokoneen luettavissa.

Metadata voidaan jakaa eri tyyppeihin sen käytön mukaan. Jaottelut poikkeavat jossain määrin toisistaan, mutta useimmissa tehdään jako kuvailevaan, rakenteelliseen ja hallinnolliseen metadataan. Malleissa myös korostetaan metadatan yhteyttä toimintaan: sen tärkein tehtävä on mahdollistaa erilaiset toiminnallisuudet, kuten tiedonhaku. (Greenberg 2005.)

Metadatan yhteydessä käytetään termejä elementti, kenttä, arvo ja muoto. Kenttä ja elementti tarkoittavat samaa asiaa, ja niillä viitataan piirteisiin, joilla tallennetta kuvataan. Esimerkiksi kuvattavan kohteen nimi merkitään *Nimeke*-kenttään ja tekijä *Tekijä*-kenttään. *Nimeke* ja *Tekijä* ovat siis kuvailukenttiä. Arvo on se, mitä kuvailukenttään tallennetaan, kentän tietosisältö. Arvo voi olla tekstiä tai numeroita. Muoto taas tarkoittaa sitä, missä muodossa arvo merkitään: merkitäänkö esimerkiksi tekijän nimi muotoon sukunimi, etunimi vai toisinpäin tai missä muodossa päivämäärä ja kieli merkitään. Usein metadatastandardi antaa suosituksia siitä, millaista muotoa missäkin kentässä pitäisi käyttää. Esimerkiksi kielen merkintään suositellaan tiettyä ISO-standardia. Erilaiset epäjohdonmukaisuudet ja poikkeukselliset ratkaisut kaikissa näissä metadatan rakenteissa vaikuttavat metadatan laatuun heikentävästi. Laatua arvioidaankin usein sekä kenttien, arvojen että muotojen osalta. (Hider 2012, 6–7, 77.)

Metadatan käyttöä ohjaavat metadatastandardit (puhutaan myös skeemoista, malleista tai formaateista). Standardissa muun muassa määritellään, mitä piirteitä tallenteesta kuvataan eli mitä kuvailukenttiä käytetään, mitä ja millaisia arvoja kentät saavat ja millaisessa muodossa arvot pi-

täisi esittää. Standardien noudattaminen on suositeltavaa, koska se vaikuttaa monin tavoin metadatan laatuun. Metadata on todennäköisesti yhdenmukaisempaa ja sen jakaminen eri järjestelmien välillä on ongelmattomampaa, kun eri järjestelmissä on noudatettu samaa standardia. Standardi voi olla organisaation sisäinen, kansallinen tai kansainvälinen. Metadatastandardit on yleensä kehitetty tietynlaista käyttökontekstia varten. Esimerkiksi Dublin Core on tarkoitettu nimenomaan digitaalisten tallenteiden kuvailuun. (Hider 2012, 103–104.)

2.3 Metadatan laatu

Millaista on laadukas metadata? Mitä tarkoittaa metadatan laatu? Yksiselitteistä määritelmää ei ole olemassa. Määrittelyn vaikeuden syystä ollaan kuitenkin yksimielisiä: koska metadata on monialaista ja kontekstisidonnaista, on mahdotonta määritellä laatu niin, että se kattaisi kaikki käyttöyhteydet. (Tani ym. 2013.)

Useassa tapauksessa metadatan laatu määritelläänkin fitness for use -ajatuksena. Eli metadatan laatu on sidoksissa sen käyttöön ja käyttöympäristöön. Käsitettä on pyritty systematisoimaan luomalla malleja, joiden avulla voidaan tunnistaa ja arvioida laatupiirteitä ja erilaisia laadun mittaamisen tapoja. Malleilla pyritään vähentämään subjektiivisuutta ja hallitsemaan laadun käsitteen moniulotteisuutta. (Tani ym. 2013.) Esittelen seuraavassa kolme tällaista mallia.

1. Brucen ja Hillmannin malli

Brucen ja Hillmannin (2004, 242–249) kehittämä metadatan laatumalli on yksi käytetyimmistä. Siinä metadatan laatua tarkastellaan seitsemän piirteen kautta. Näitä ovat:

- täydellisyys (completeness): metadatan pitäisi kuvailla kohdettaan mahdollisimman kattavasti. Tähän kuuluu muun muassa se, että metadatakenttiä pitäisi olla täytettynä mahdollisimman paljon.

- tarkkuus (accuracy): tiedon täytyisi olla oikeellista. Esimerkiksi kirjoitusvirheet tai faktuaaliset virheet heikentävät laatua.

- odotuksenmukaisuus (conformance to expectations): metadatan pitäisi vastata käyttäjäkunnan odotuksiin. Tätä piirrettä voidaan pitää yhtenä suurena metadatan laadun mittarina.

- looginen johdonmukaisuus ja koherenssi (logical consistency and coherence): metadatan pitäisi noudattaa standardeja johdonmukaisesti ja käyttää yleisesti hyväksyttyjä käsit-

teitä. Koherenssi on myös tietueen sisäistä: kuvailukenttien pitäisi kuvailla samaa kohdetta.

- saavutettavuus (accessibility): metadatan pitäisi olla luettavissa ja ymmärrettävissä, niin ihmisten kuin koneidenkin. Saavutettavuus voi siis olla luonteeltaan fyysistä tai kognitiivista.

- ajantasaisuus (timeliness): jos kuvailun kohteena oleva tietoresurssi muuttuu, täytyisi metadatankin muuttua.

- alkuperä (provenance): joskus metadatan laatua voi arvioida sen mukaan, kuka metadatan on luonut. Onko metadata esimerkiksi luotu ihmisvoimin vai automaattisesti? Alkuperä-piirteeseen kuuluu myös, millaisia muutoksia metadata on olemassaolonsa aikana kokenut: onko arvoja esimerkiksi lisätty tai poistettu.

Brucen ja Hillmannin esittelemä piirrejoukko ei ole sellaisenaan tarkoitettu laadun mittaamiseen, koska se on liian abstrakti. Pelkkien piirteiden luettelo ei vielä kerro, miten laatua oikeastaan pitäisi mitata. Malli tarjoaa kuitenkin hyvän lähtökohdan, josta käsitettä voi lähteä operationalisoimaan.

2. Stvilia et al. malli

Stvilian ja muiden (2007) laatumallin ensimmäinen versio ilmestyi vuonna 2004. Tutkijat ovat myöhemmin kehittäneet malliaan. Tässä esitelty versio on vuodelta 2007.

Stvilian ja muiden *Information Quality Assessment Framework* on teoreettisempi kuin Brucen ja Hillmannin malli. Sen tarkoituksena on pohtia informaation laadun ulottuvuuksia laajemmin kuin vain metadatan kohdalla. Mallissa laatuulottuvuksia on kaikkiaan 38, jotka on jaettu kolmeen kategoriaan: sisäinen laatu (intrinsic IQ), suhteellinen/kontekstuaalinen laatu (relational/contextual IQ) ja maineeseen perustuva laatu (reputational IQ).

Sisäisellä laadulla tarkoitetaan sellaisia piirteitä, joita voidaan arvioida ottamalla huomioon vain aineisto itsessään ja sen suhde erilaisiin määrityksiin (esimerkiksi kirjoitusvirheet suhteessa sanakirjaan/kielioppiin). Sisäisen laadun piirteet eivät ole riippuvaisia kontekstista, joten niitä voidaan mitata enemmän tai vähemmän objektiivisesti.

Suhteellisen/kontekstuaalisen laadun piirteet puolestaan mittaavat informaation laatua suhteessa johonkin sen käyttökonteksteista (esimerkiksi onko metadatan kertoma nimeke oikein suhteessa tallenteen varsinaiseen nimekkeeseen).

Maineeseen perustuva laatu pohjautuu informaation paikkaan kulttuurisissa rakenteissa. Esimerkiksi informaatio katsotaan laadukkaammaksi, koska sen on luonut tietty taho.

Mallia on arvosteltu sen teoreettisesta monimutkaisuudesta ja jatkuvasta muuttumisesta. Tutkijat itse ovat operationalisoineet malliaan joidenkin piirteiden osalta nimenomaan metadatan laadun arvioimiseen, mutta mallia kokonaisuudessaan ei ole hyödynnetty.

3. Margaritopoulos et al. malli

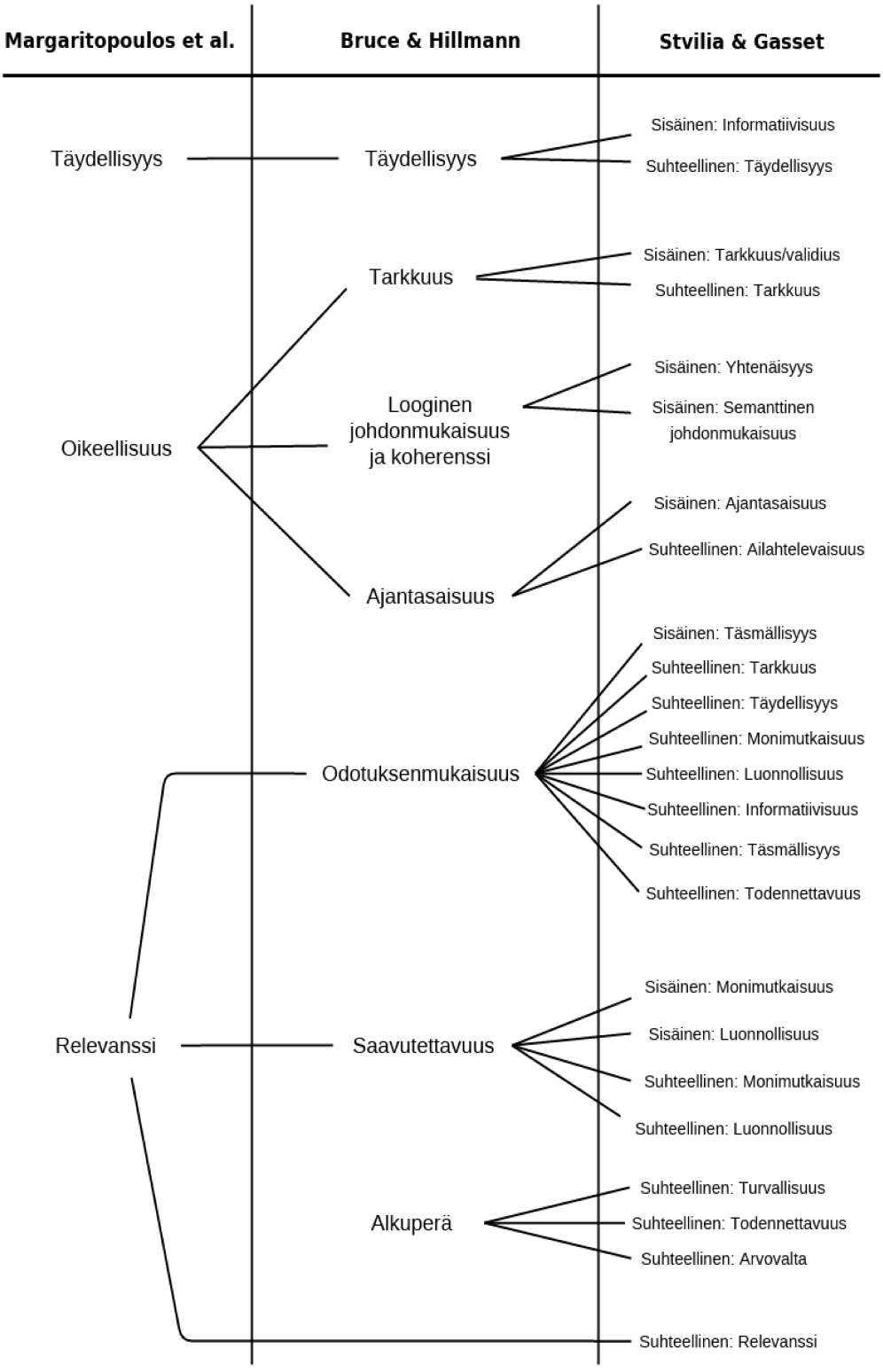
Margaritopoulosin ja muiden (2008) laatumalli *A Conceptual Framework for Metadata Quality Assessment* perustuu oikeussalimetaforaan. Mallissa metadatatietueen laatua verrataan todistajan antaman lausunnon laatuun oikeudessa. Kun todistajaa velvoitetaan kertomaan totuus, kokonaisuudessaan eikä mitään muuta, tämä käännetään Margaritopoulosin ja muiden mallissa kolmeksi metadatan laatupiirteeksi: oikeellisuus (correctness), täydellisyys (completeness) ja relevanssi (relevance). Kirjoittajien mukaan näiden kolmen laatupiirteen avulla saadaan tarpeeksi tietoa metadatan laadusta missä tahansa kontekstissa.

Malli on luonteeltaan hyvin abstrakti eikä se tarjoa konkreettisia välineitä laadun mittaamiseen.

2.4 Yhteenvetoa laatumalleista

Metadatan laatua käsitteellistävät mallit eroavat toisistaan muun muassa laatupiirteiden määrän suhteen. Tanin ja muiden (2013) mukaan eri malleissa on myös yhteneväisyyksiä. Esimerkiksi samat laatupiirteet toistuvat useammassa mallissa. Ylipäätään se, että metadatan laatua lähestytään nimenomaan tiettyjen piirteiden kautta, yhdistää eri malleja. Ochoa (2014) on artikkelissaan esittänyt kaikki kolme mallia yhdessä, jolloin niiden risteäminen tulee paremmin esille (kuva 1).

Koska metadatan laatu on aina riippuvaista kontekstista, ei voidakaan päästä yhteisymmärykseen, mitä laatupiirteillä yksiselitteisesti tarkoitetaan ja mitkä niistä ovat olennaisimpia, Tani ja muut (2013) toteavat. Ochoan arvion mukaan mitä tahansa mallia hyödyntämällä saa kuvan metadatan laadusta.



Kuva 1: Metadatan laatumallien yhteneväisyydet (Ochoa 2014)

Suurin ero malleissa on niiden yksityiskohtaisuudessa: käyttämällä Stvilian ja muiden 38 laatupiirrettä saa väistämättä laajemman kuvan metadatan laadusta kuin Margaritopoulosin ja muiden kolmen piirteen perusteella. Toisaalta taas mitä useampaa piirrettä metadatasta haluaa tutkia, sitä enemmän siihen täytyy panostaa resursseja. Mallin valinta on siis tasapainoilua sen suhteen, kuinka yksityiskohtaisia tuloksia haluaa ja paljonko työhön on valmis panostamaan resursseja. (Ochoa 2014, 71.)

3 METADATAN LAADUN MITTAAMINEN

Metadatan laatuun ja laadun mittaamiseen on tärkeää kiinnittää huomiota järjestelmissä, joiden tarkoitus on edistää ja tukea tiedon jakamista ja uudelleenkäyttöä. Järjestelmiin tallennettua tietoa hallitaan ja haetaan metadatan avulla, joten metadatan laatu vaikuttaa osaltaan koko järjestelmän hyödyllisyyteen ja käytettävyyteen. Jos metadata on laadultaan huonoa, on järjestelmiin tallennettujen aineistojen löytäminen ja uudelleenkäyttö hankalaa. (Tani ym. 2013.) Myös Barton ja muut (2003) korostavat laadukkaan metadatan tärkeyttä paitsi järjestelmien yhteentoimivuudelle myös koko järjestelmän käytölle. Parhaimmillaan metadata on tehokas työkalu, joka mahdollistaa aineistojen löytämisen nopeasti ja helposti. Pahimmassa tapauksessa huonolaatuinen metadata johtaa tilanteeseen, jossa aineisto on tallennettu järjestelmään, mutta se jää pimentoon ja käyttämättä, koska käyttäjä ei sitä löydä. (Barton & ym. 2003.) Hider (2012, 86) tiivistää metadatan laadun tärkeyden seuraavasti: ”Ultimately, what makes for good metadata is its potential to support effective information retrieval.” Ja koska hyvälaatuinen metadata suuressa määrin vaikuttaa tiedonhaun onnistumiseen, on järkeenkäypää arvioida metadatan laatua ja tarpeen mukaan parantaa sitä, Hider (2012, 87) jatkaa.

Ochoa (2014) jakaa metadatan laadun tutkimuksen käytettyjen menetelmien mukaan manuaaliseen ja automaattisen tutkimukseen.

Manuaalisessa tutkimuksessa tutkitaan tilastollisesti merkittävää otosta metadatakokoelmasta. Tutkimuksen suorittaa ammattilainen käyttäen jotain ennalta valittua laatumallia tai -kriteeristöä. Tuloksena on eri laatupiirteiden keskiarvo. (Ochoa 2014, 73.) Manuaalisen tutkimuksen huonona puolena on, että kun kokoelmaan lisätään uusia tietueita, eivät tulokset enää ole päteviä, vaan arvio on tehtävä uudestaan. Otoksen perusteella ei voida sanoa mitään yksittäisistä metadatatietueista (paitsi niistä, jotka otoksessa ovat mukana). Manuaalisen tutkimuksen tekeminen on myös kallista. Kun digitaalisen aineiston ja sitä kautta metadatan määrän oletetaan vain kasvavan tulevaisuudessa, ei ihmistyön käyttäminen laadunarviointiin enää ole kannattavaa. (Ochoa & Duval 2009.)

Automaattisessa tutkimuksessa kerätään numeerista ja tilastollista dataa kaikista kokoelman metadatatietueista. Tutkimuksen keskiössä ovat metriikat, laskukaavat, jotka voidaan suorittaa tietokoneella. Metriikoissa voidaan ottaa huomioon paitsi metadata myös sen käyttö ja muu kokoelmaan liittyvä kontekstuaalinen tieto. Automaattisella tutkimuksella saadaan laatuarvio kokoelman jokaisesta metadatatietueesta huomattavasti halvemmalla kuin manuaalisessa tutkimuksessa. Vaikein osuus tutkimuksessa on metriikoiden luominen eli sen päättäminen, miten laatupiirteet muutetaan koneen ymmärtämään muotoon. Automaattista tutkimusta on kritisoitu siitä,

että se ei tarjoa yhtä merkityksellistä tietoa laadusta kuin manuaalinen arviointi. (Ochoa 2014, 73.)

Metadatan laadun tutkiminen ei ole ongelmatonta toteutetaanpa se sitten manuaalisesti tai automaattisesti, sillä laatu on moniulotteista. Metadatatietueessa on useita itsenäisiä piirteitä, jotka vaikuttavat laatuun: kirjoitusvirheet ja kuvailukenttien käytön taajuus vaikuttavat molemmat laatuun, mutta eivät ole välttämättä riippuvaisia toisistaan. Metadatan laatu on sidoksissa kokoelman käyttäjiin ja näiden suorittamiin tehtäviin. Yhden järjestelmän metadatan laadun arviointi ei välttämättä kerro mitään toisen järjestelmän metadatan laadusta, jos aineiston käyttö on niissä erilaista. Laatu ei ole staattista: tietueen vanheneminen, uusien metadatatietueiden lisääminen järjestelmään tai muutokset käyttötavoissa voivat vaikuttaa siihen, kuinka hyvin tietueet mahdollistavat järjestelmän erilaiset käytöt. (Ochoa 2014, 65–66.)

3.1 Manuaalinen tutkimus

Suuri osa metadatan laatua arvioivista tutkimuksista on toteutettu manuaalisesti tutkimalla ja analysoimalla tietyn suuruista otosta kohteena olevasta kokoelmasta. Tutkimuksissa on analysoitu esimerkiksi metadatan luomista ja sen vaikutusta metadatan laatuun. Tätä ovat tutkineet muun muassa Currier ja muut (Currier ym. 2004), Barton ja muut (Barton ym. 2003) sekä Greenberg ja muut (Greenberg ym. 2001). Joissakin tutkimuksissa taas on keskitytty nimenomaan yhteisarkistojen metadatan laatuun ja siihen, kuinka laatu vaikuttaa yhteentoimivuuteen ja tiedonhakuun, muun muassa Shreeves ja muut (Shreeves ym. 2005)

Esittelen seuraavaksi tarkemmin kaksi manuaalista tutkimusta. Tutkimukset ovat mielenkiintoisia tämän työn kannalta, koska niissä tutkitaan nimenomaan yliopistojen julkaisuarkistojen metadatan laatua.

Eun G. Parkin ja Marc Richardin (2011) tutkimuksen tavoitteena oli tutkia, miten ETD-MS-metadatastandardia on käytetty aineistojen kuvailuun yliopistojen julkaisuarkistoissa. Vaikka tutkimuksessa ei suoraan mainitakaan metadatan laadun arviointia, on metadatastandardin johdonmukainen käyttö yksi merkittävä tekijä metadatan laatua arvioitaessa.

Aineistona Park ja Richard käyttivät kymmenen kanadalaisen yliopiston julkaisuarkistojen metadataa. Tutkittu metadata kuvaili maisteritason opinnäytteitä ja akateemisia väitöskirjoja. Tutkijat pyysivät 15 yliopistolta 10 tietueen suuruista otosta heidän julkaisuarkistostaan. Otokseen pyydettiin ottamaan mukaan tietueita, jotka on luotu vuoden 2000 jälkeen ja jotka havainnollistavat suosituimpia kuvailuun käytettyjä kenttiä mahdollisimman hyvin. Mukaan valikoitui kaikkiaan 10 yliopiston julkaisuarkistot eli yhteensä 100 tietueen metadata.

Julkaisuarkistojen käyttämä ETD-MS-metadatastandardi perustuu Dublin Coreen (DC). ETD-MS on kehitetty erityisesti elektronisten opinnäytteiden ja väitöskirjojen kuvailuun. ETD-MS:ssä on käytössä samoja kuvailukenttiä kuin DC:ssa, mutta siihen on lisätty kenttä *Thesis* ja tälle neljä tarkennetta (degree name, level, discipline, grantor). Standardissa on 21 kuvailukenttää: 14 peruskenttää (muun muassa *Nimeke*, *Tekijä*, *Kieli*, *Aihe*, *Kuvailu*) ja näistä johdetut 7 kenttää, joissa on mukana kenttätarkenne.

Analysoiduissa tietueissa kuvailuelementtejä oli käytetty standardista poikkeavasti hyvin monella tavalla. Park ja Richard jakavat nämä poikkevat käytöt kolmeen luokkaan: 1) tapaukset, joissa on käytetty standardin mukaista kuvailukenttää, mutta siihen lisätty oma tarkenne. Kentän merkitys on pysynyt kuitenkin standardinmukaisena; 2) tapaukset, joissa kenttiä käytetty standardista poikkeavalla tavalla; 3) tapaukset, joissa käytetty tarkenteita ja joissa kentän merkitys muuttunut standardista poikkeavaksi.

Joitakin kohdan 1 tapauksia esiintyi hyvin johdonmukaisesti arkistojen välillä. Esimerkiksi *Kieli*-kentässä oli käytetty tarkennetta language.iso tai *Formaatti*-kentässä format.mimetype vaikka kumpikaan ei ole standardin määrittelemä tarkenne. Tämän luokan poikkeamat olivat kaikkein yleisimpiä. Kohdan 2 poikkeamat ilmenivät esimerkiksi siten, että kuvailukenttään oli tallennettu tietoa, joka kuuluisi toiseen kenttään (esimerkiksi *Kuvailu*-kentässä tutkielman tiivistelmän sijaan tieto tallenteen lajista). Kohdan 3 poikkeamat ilmenivät esimerkiksi siinä, että oli käytetty kenttää contributor.author *Tekijä*-kentän sijaan ilmoittamaan työn tekijä.

Erityisesti viimeisen kohdan tapaukset olivat tutkijoiden mielestä hyvin ongelmallisia, koska kuvailukenttiä oli niissä käytetty aivan vääränlaisen tiedon ilmoittamiseen. Tähän luokkaan kuuluivat myös poikkeamat, joissa saman tiedon ilmoittamiseen oli käytetty monta eri kenttää eri arkistoissa. Esimerkiksi laitos tai ohjelma, johon opinnäytetyö kuului, oli tallennettu neljään eri kenttään, vaikka tiedolle on olemassa standardissa selkeästi oma kenttänsä. Poikkeavia olivat myös käänteiset tapaukset: samaa kenttää oli käytetty tallentamaan eri tietoa. Esimerkiksi *Kuvailu*-kentässä oli ilmoitettu päivämäärä, työhön liittyvät muistiinpanot, oppiaste tai taloudelliset tukijat. Myös metadatan merkintätavoissa oli vaihtelua. Huomattavinta tämä oli päivämäärän ilmoittamisessa.

Parkin ja Richardin (2011) analyysin tuloksena oli, että vaikka kaikissa julkaisuarkistoissa on käytössä sama metadatastandardi ja näin ollen samat säännöt ja suositukset metadatan käytölle, on sen soveltamisessa suuria eroja. Parkin ja Richardin mukaan heidän analyysinsa viittaa siihen, että julkaisuarkistot yksinkertaisesti eivät välitä tai ota huomioon ohjeita ja kuvauksia, miten metadatastandardia pitäisi käyttää. Julkaisarkistot ovat kuitenkin yhä merkittävämpi väylä tarjota yliopistossa tuotettuja aineistoja vapaasti kaikkien saataville, joten olisi suotavaa, että niiden etsi-

minen ja löytäminen olisi suhteellisen vaivatonta. Poikkeamat metadatastandardin suosittelemasta käytöstä voivat vaikeuttaa näitä toimia huomattavasti.

Mary Kurtz (2010) puolestaan tutki kolmen yhdysvaltalaisen yliopiston julkaisuarkiston metadatan laatua. Kurtzin tutkimuksen keskiössä on nimenomaan Dublin Core -metadatastandardin ja digitaalisen arkiston luomiseen ja hallintaan tarkoitetun DSpace-ohjelmiston yhteiskäyttö. Erityisen kiinnostavaa tietoa artikkeli tarjoaa paitsi tutkimustulosten muodossa myös kattavassa selvityksessään, miten DSpace-ohjelmisto toimii ja millaisia ominaisuuksia siinä on.

Kurtz arvioi metadatan laatua täydellisyyden, tarkkuuden ja johdonmukaisuuden näkökulmasta. Nämä kolme oli valittu sillä perusteella, että Parkin (2009) mukaan ne ovat yleisimmin hyväksyttyjä metadatan laadun kriteereitä. Tutkimuksessa aineistona oli 20 tietueen suuruinen otos jokaisesta julkaisuarkistosta. Julkaisuarkistot jakautuivat erilaisia aineistoja sisältäviin kokoelmiin. Kustakin kokoelmasta oli mukana yksi tietue. Kaikkien kolmen julkaisuarkiston taustalla oli DSpace-ohjelmisto.

Täydellisyydellä Kurz tarkoittaa, kuinka laajasti olennaisia metadatakenttiä on käytetty. Olennaisiksi kentiksi Kurz valitsi *Nimeke*-, *Tekijä*-, *Aihe*- ja *Kuvailu* -kentät. Koska DSpace määrittää automaattisesti *Nimeke*-kentän pakolliseksi, oli tätä kenttää täytetty kaikissa tapauksissa. Suurimmat erot eri arkistojen välillä syntyivät *Aihe*- ja *Kuvailu* -kenttien käytössä. *Aihe*-kentän käyttö vaihteli sadasta prosentista neljäänkymmeneen, *Kuvailu*-kentän puolestaan 75:stä neljäänkymmeneen. Näiden kenttien osalta sama julkaisuarkisto sai alimmat prosenttiluvut, mikä Kurtzin mukaan viittaa, että kyseisen arkiston metadata ei ole täydellisyyden osalta laadukasta.

Tarkkuudella Kurz tarkoittaa metadatakenttien tietojen oikeellisuutta. Tarkkuuteen vaikuttavat muun muassa kirjoitusvirheet. Ongelmia tarkkuudessa oli kaikissa arkistoissa, ja ne ilmenivät monin eri tavoin. Esimerkiksi puutteellisuudet asiasanojen käytössä (sanaa *the* ei voida katsoa oikeelliseksi asiasanaksi, vaikka käytettäisiin kuinka kattavaa asiasanan määritelmää tahansa, Kurtz huomauttaa), keskenään ristiriidassa olevat arvot (esimerkiksi kaksi merkintää tallenteen kielestä, joista vain toinen on oikeellinen) ja epätietoisuus *Tekijä*- ja *Muu tekijä* -kenttien käytöstä (joissakin tietueissa oli sama tieto molemmissa kentissä). Kurtz huomasi myös, että asiasanojen yksityiskohtaisuudessa ja kontrolloitujen asiasanastojen käytössä oli suurta vaihtelua eri kokoelmien välillä.

Johdonmukaisuus on määritelty metadatan yhdenmukaisuudeksi esimerkiksi sen suhteen, millaisessa muodossa kuvailukenttien arvot on esitetty. Koska DSpace generoi automaattisesti tiettyjen kenttien metadatan suoraan tallenteesta (esimerkiksi päiväykseen liittyvät tiedot ja *Tekijä*-kentän arvot) ei näiden kenttien osalta ilmennyt ongelmia johdonmukaisuudessa. Ongelmallisia

johdonmukaisuuden kannalta olivat esimerkiksi henkilöiden nimet, kun niitä oli käytetty *Aihe*-kentässä: sukunimen ja etunimen järjestys vaihteli. Myös suurakkosten kirjava käyttö tässä kentässä aiheutti epäjohdonmukaisuuksia. Tällaiset epäjohdonmukaisuudet voivat aiheuttaa ongelmia muun muassa aineiston löytymisessä. Epäjohdonmukaista oli myös dc.description.sponsorship-kentän käyttö. Kenttää oli käytetty hyvin erilaisten tietojen ilmoittamiseen jopa saman arkiston sisällä: se saattoi kertoa opinnäytetyön ohjaajan, rahoituksen lähteen tai lahjoittajan nimen.

Koska Kurz tutki vain kolmea laatupiirrettä, rajaa tämä pois monia mahdollisia laadun ongelmia, joita aineistoissa voi olla. Kurtzin tutkimuksessa tuli kuitenkin ilmi samanlaisia ongelmia metadatan laadussa kuin Parkin ja Richardinkin edellä esitellyssä tutkimuksessa. Esimerkiksi ongelmat *Tekijä-* ja *Muu tekijä* -kenttien käytössä, saman kuvailukentän käyttö hyvin erilaisen tiedon ilmoittamiseen, jolloin standardin mukainen merkitys hämärtyy ja epäjohdonmukaisuudet metadatan muodossa tulivat ilmi molemmissa tutkimuksissa.

3.2 Automaattinen tutkimus

Metadatan laadun mittaaminen automaattisin menetelmin on saanut erityisesti viime vuosina yhä enemmän huomiota. Buin ja Parkin (2006) metadatan laatua automaattisin menetelmin mittaava tutkimus perustuu aineiston tilastolliseen analyysiin. Ochoan ja Duvalin (2009) kehittämät automaattiset mittausmenetelmät antavat mahdollisuuden metadatan arviointiin hyvin monipuolisesti. Reiche ja Höfig (2013) puolestaan hyödyntävät omassa tutkimuksessaan Ochoan ja Duvalin metriikoita käytännössä. Automaattisia menetelmiä ovat kehittäneet ja testanneet myös muun muassa Hughes (2005) sekä Gavrilis ja kumppanit (2015).

Yen Bui ja Jung-ran Park (2006) tutkivat National Science Digital Library -palvelun (NSDL) metadatan laatua. Digitaalisessa kirjastossa on eri organisaatioiden tieteellistä aineistoa usealta eri alalta. Aineisto jakautuu organisaatioiden mukaisiin kokoelmiin, joita tutkimuksen teon aikaan olit yli sata.

Tutkimuksen aineistona oli 1040034 metadatatietuetta eri kokoelmista. NSDL käyttää Dublin Core -metadatatostandardin tarkennettua muotoa. Tutkimuksessa arvioitiin NSDL:n metadatan laatua yleensä ja laatuvaihteluita eri kokoelmien metadatan välillä. Laadun mittareina käytettiin viittä piirrettä: metadatan taajuus/lukumäärä (frequency), johdonmukaisuus (consistency), täydellisyys (completeness), tarkkuus (accuracy) ja datan tuottajien omat lisäykset (local additions of data providers). Artikkelissa ei tosin selvennetä sitä, mitä nämä piirteet oikeastaan tässä yhteydessä tarkoittavat. Metadata muutettiin analyysia varten taulukkomuotoon, jossa yksi rivi vastasi

yhtä tietuetta ja sarakkeet käytettyjä kuvailukenttiä. Tutkimuksessa keskityttiin arvioimaan metadatan laatua täydellisyyden näkökulmasta.

Täydellisyyttä mitattiin laskemalla, kuinka paljon kuvailukenttiä on käytetty. Olennaisiksi kentiksi, jotka tietueessa pitäisi ainakin olla, määriteltiin *Nimeke*, *Kuvaus*, *Tekijä*, *Aihe*, *Identifiointitunnus* ja *Laji*. Nämä kentät valittiin, koska ne katsottiin tärkeiksi tiedonhaun kannalta.

Nimekkeen ja *Identifiointitunnuksen* osalta käyttö oli lähes sataprosenttista. *Tekijä*-kentän käytön tutkijat odottivat olevan korkeampi kuin noin 83 prosenttia. *Aihe*-kentän käyttö oli puolestaan epätasaista: joissakin kokoelmissa tallenteen kuvailuun oli käytetty useita kuvailutermejä, mutta useimmissa kokoelmissa oli tyydytty vain muutamiin kuvailutermeihin. Buin ja Parkin yleishavainto kenttien käytöstä oli, että se on hyvin hajaantunutta: tiettyjä kenttiä (*Nimeke*, *Identifiointitunnus*) oli käytetty eri kokoelmien lähes kaikissa metadatatietueissa, mutta joidenkin kenttien käyttö taas oli vain muutaman prosentin luokkaa (esimerkiksi *Muu tekijä*, *Kattavuus*).

Artikkeli jää hieman puolitiehen metadatan laatua tutkiessaan, koska luvattua jatkotutkimusta, jossa metadatan laatua olisi arvioitu myös muiden laatupiirteiden osalta, en ainakaan itse löytänyt. Buin ja Parkin tilastollinen analyysi on kuitenkin ensimmäisiä, jossa aineistona on hyvin suuri määrä metadatatietueita eikä vain pieni otos. Tutkimus antoi samantyyppisiä tuloksia kuin esimerkiksi Kurtzin analyysi julkaisuarkiston metadatan laadusta täydellisyyden osalta. Tallenteen aihetta kuvailevien termien käytössä ilmeni myös molemmissa tutkimuksissa suurta hajontaa.

Xavier Ochoa ja Erik Duval (2009) vievät metadatan laadun mittaamisen automaattisin menetelmin tilastollista analyysia pidemmälle. Tutkijat esittelevät artikkelissaan joukon mittaustapoja, joilla laatua voidaan mitata täysin koneellisin menetelmin. Ochoan ja Duvalin metriikat pohjautuvat Brucen ja Hillmannin (2004) mallissa esiteltyihin 7 laatupiirteeseen. Kyseinen laatumalli on valittu, koska se Ochoan ja Duvalin mielestä kattaa myös esimerkiksi Stvilian ja muiden (2007) mallissa esitellyt laatupiirteet, mutta on tarpeeksi kompakti operationalisoitavaksi.

Koska esitellyt metriikat perustuvat samoihin laatupiirteisiin kuin useat manuaalisestikin suoritetut arvioinnit, ja ne voidaan toteuttaa koneellisesti, uskovat tutkijat, että laadun arviointi niillä on skaalautuvaa ja tarjoaa merkityksellisiä tuloksia. Jotta metadatan eri laatupiirteitä voidaan mitata automaattisesti, täytyy kukin piirre muuttaa tietokoneen ymmärtämään muotoon. Ochoan ja Duvalin mittausmenetelmät ovat joukko laskukaavoja eli metriikoita, joilla kullekin laatupiirteelle saadaan lukuarvo (useimmissa metriikoissa luku väliltä 0 ja 1), jolloin korkeampi luku tarkoittaa parempaa laatua.

Täydellisyydellä Ochoa ja Duval tarkoittavat sitä, kuinka paljon kuvailukenttiä on käytetty verrattuna siihen, kuinka paljon niitä standardissa määritetään. Painotettu täydellisyys (weighted completeness) ottaa huomioon, kuinka paljon niin sanottuja olennaisia kenttiä (esimerkiksi tiedonhaun kannalta) on käytetty, ja antaa näille mittauksessa korkeamman painoarvon. Tarkkuudella tarkoitetaan metadatan ja sen kuvaileman kohteen semanttista etäisyyttä toisistaan. Ochoa ja Duval mittaavat etäisyyttä *Nimeke*- ja *Kuvailu*-kenttien osalta käyttämällä tiedonhaun tutkimuksessa yleistä vektorimallia. Odotuksenmukaisuutta arvioidaan metadatan välittämän informaation rikkauden ja yksilöllisyyden kautta. Kenttien osalta, joihin tallennetaan kategorista tietoa (esimerkiksi asiasanat *Aihe*-kentässä) odotuksenmukaisuus mitataan laskemalla, kuinka paljon on käytetty yksilöllisiä termejä. Johdonmukaisuutta mitataan laskemalla, kuinka paljon on käytetty kenttiä, joita standardissa ei ole, missä määrin on käytetty standardin määrittämiä pakollisia kenttiä ja tarkastelemalla, sisältävätkö kentät oikean muotoista dataa. Koherenttiutta mitataan vertaamalla vapaata tekstiä sisältävien kuvailukenttien sisältöä ja laskemalla niiden semanttinen etäisyys; mitä lähempänä kuvailukenttien arvot ovat semanttisesti toisiaan, sitä koherentimpaa metadata on. Saavutettavuutta mitataan laskemalla, missä määrin metadatatietueet ovat linkittyneinä toisiinsa. Ajantasaisuudella Ochoa ja Duval mittaavat metadatan laatua tietyllä hetkellä: tietyllä hetkellä mitattua kaikkien muiden metriikoiden keskiarvoa verrataan esimerkiksi vuoden päästä mitattuun keskiarvoon. Alkuperän mittaamiseksi lasketaan kaikkien muiden metriikoiden tuottaman keskiarvon keskiarvo suhteuttamalla se kokoelmassa olevien metadatatietueiden määrään. Tässä ajatuksena on, että tietyn organisaation metadatan laatu kokonaisuudessaan kertoo siitä, kuinka luotettava metadatan tuottaja kyseinen organisaatio on.

Ochoa ja Duval testaavat metriikoitaan kolmessa tutkimuksessa. Ensimmäisessä ammattilaiset arvioivat metadatietueita manuaalisesti. Kohteena oli otos oppimaateriaaleja sisältävästä ARIADNE-arkistosta. Otos koostui englanninkielisistä informaatioteknologiaa käsittelevistä metadatatietueista. Kymmenen tietueen metadatan oli luonut kyseisen materiaalin kirjoittaja. Toiset kymmenen tapausta olivat sellaisia, että metadatan oli luonut automaattinen indeksoija. Kokeessa käytettiin Brucen ja Hillmannin mallin laatupiirteitä.

Arvioijat saivat ohjeet siitä, mihin seikkoihin kunkin piirteen kohdalla pitäisi kiinnittää huomiota. Koe toteutettiin webbilomakkeen avulla. Osallistujat kävivät läpi 20 tietuetta satunnaisessa järjestyksessä ja arvioivat metadatan laatua jokaisen piirteen osalta. Arviointiin käytettiin seitsemän portaista asteikkoa (Extremely low quality – Extremely high quality). Arvioijat eivät tienneet, oliko metadata ihmisen vai koneen luomaa. Analyysissa otettiin huomioon vain sellaiset arviot, joissa oli arvioitu kaikki tapaukset. Näitä oli 22.

Sama otos arvioitiin myös käyttäen automaattisia mittaustapoja. Tässä laskettiin laatuarvot samoille seitsemälle piirteelle, jotka oli arvioitu manuaalisesti. Tuloksista selvisi, että ihmisten tekemät ja automaattisilla menetelmillä suoritetut laatumittaukset eivät korreloi keskenään lukuunottamatta yhtä piirrettä (odotuksenmukaisuus). Ihmiset arvioivat laadun kaikkien piirteiden osalta korkeammaksi kuin automaattinen arviointi. Ochoa ja Duval eivät kuitenkaan hylkää hypoteesiaan, että automaattisilla metriikoilla voidaan arvioida metadatan laatua. Ihmisten suorittamista arvioinneista kävi ilmi, että arvioijat eivät kaikkien piirteiden kohdalla noudattaneetkaan annettuja ohjeita eli he arvioivat eri piirteitä kuin oli ohjeistettu. Varmoja johtopäätöksiä tutkimuksesta ei siis voinut tehdä, koska korrelaatio eri arviointitapojen välillä oli hyvin pientä.

Toisessa tutkimuksessa testattiin, kuinka metriikat erottavat laadukkaan ja huonolaatuisen metadatan toisistaan. Koska saatavana ei ollut metadataa, jonka laatu tiedettäisiin (eli onko se hyvää vai huonoa), valittiin kohteeksi tutkijoiden omaan arvioon perustuen kaksi laadultaan ääripäätä edustavaa metadatakokonaisuutta. Toisen kokonaisuuden metadatan olivat luoneet ammattilaisindeksoijat. Kyseessä oli LOM-standardia noudattavat 4426 oppimateriaalia pdf-muodossa. Toinen kokonaisuus muodostui samasta materiaalista, mutta metadata luotiin automaattista indeksointia käyttäen. Ennakko-oletus oli, että ihmisen luoma metadata olisi korkealaatuista verrattuna automaattisesti luotuun. Laadunarvioinnissa käytettiin samoja laskukaavoja kuin aikaisemmassakin tutkimuksessa pienin muutoksin. Laatuarvot laskettiin jokaiselle metadatatietueelle.

Tuloksena oli, että manuaalisesti tuotetun metadatan laatuarvot olivat yleisesti korkeammat kuin automaattisesti tuotetun. Esimerkiksi metadatan täydellisyyteen vaikutti se, että ihmiset täyttävät enemmän ja olennaisempia kuvailukenttiä kuin automaattinen indeksoija. Tarkkuuden osalta automaattisesti tuotettu metadata sai sen sijaan huomattavasti korkeamman tuloksen. Tämä johtuu tutkijoiden mukaan laskentatavasta: tarkkuus lasketaan vertaamalla kuvailukentän tekstin ja kuvailun kohteen sisällön (eli esimerkiksi artikkelin varsinainen sisältö) semanttista etäisyyttä. Koska automaattinen indeksoija ottaa kuvailukenttään tekstiä suoraan kuvailun kohteesta, on tämä väistämättä semanttisesti hyvin lähellä sitä, mistä kohteessa on kyse. Ihmiset käyttävät myös monipuolisemmin kuvailutermejä, mikä johtaa laadukkaampaan metadataan. Toisaalta taas kuvailutietojen informatiivisuutta mitattaessa molemmilla tavoilla tuotetut metadatat saivat korkeat laatuarvot.

Tutkijoiden mukaan koe osoitti, että automaattisilla metriikoilla tehty laatuarviointi kykenee erottamaan laadukkaan ja huonolaatuisen metadatan toisistaan. Tutkimus todisti myös sen, että metriikoita pystytään soveltamaan isoon aineistoon.

Kolmannessa tutkimuksessa testattiin, miten metriikat tunnistavat huonolaatuista metadataa kokoelmasta. Jos metriikat toimivat odotusten mukaan, huonolaatuinen metadata saa arvioinnissa

matalamman laatutuloksen. Aineistoksi valittiin osa aikaisemmassa tutkimuksessa käytetyistä manuaalisesti ja automaattisesti luoduista metadatatietueista. Laatuarvon perusteella aineisto jaettiin neljään luokkaan (huonolaatuisin–hyvälaatuisin). Näistä luokista valittiin sattumanvaraisesti tietueet, jotka esiteltiin arvioijille. Heidän tehtävänään oli valita metadatan laadultaan huonoin tietue. Arvioijille annettiin ohjeet, millä perusteella mitäkin piirrettä pitää arvioida (esimerkiksi "Valitse tapaukset, jotka tarjoavat vähiten informaatiota"). Jos vähintään kolme neljästä arvioijasta valitsi saman tapauksen, katsottiin, että valinta oli johdonmukainen. Tämän jälkeen näitä valintoja verrattiin metriikoiden antamiin laatuarvoihin ja katsottiin valitsivatko arvioijat sellaiset tapaukset, joilla oli matalin laatuarvo.

Tulos oli osin ristiriitainen. Kolmen laatupiirteen (täydellisyys, painotettu täydellisyys, odotuksenmukaisuus ja kaikkien laatupiirteiden yhteenlaskettu keskiarvo) osalta arvioijien valinta ja metriikoilla automaattisesti laskettu laatuarvo kohtasivat. Toisaalta tiettyjen piirteiden laatuarvo ei korreloinut arvioijien valinnan kanssa. Näitä piirteitä olivat tarkkuus, koherenttius ja saavutettavuus. Tämä viittaisi siihen, että metriikat eivät mittaa samaa laatupiirrettä kuin ihmisarvioijat. Tutkijat tulivat siihen johtopäätökseen, että tiettyjen laatupiirteiden osalta automaattiset metriikat pystyvät tunnistamaan huonolaatuisen metadatan. Erityisen yllättävää oli, että kaikkien laatupiirteiden keskiarvo -metriikka kuitenkin korreloi arvioijien tekemän valinnan kanssa.

Kolmen testin lopputulemana oli, että erityisen hyödyllistä automaattisten metriikoiden käyttö on tunnistettaessa huonolaatuista metadataa. Automaattiset metriikat kykenevät myös mittaamaan sellaisia laatupiirteitä, joita manuaalisesti ei voida arvioida. Esimerkiksi sitä, kuinka laajasti kategorisia arvoja on käytetty tai missä määrin metadatatietueet ovat linkittyneinä toisiinsa, on lähes mahdotonta arvioida muuten kuin koneellisesti.

Ochoan ja Duvalin metriikat joutuvat käytännön testiin Reichen ja Höfigin artikkelissa (2013), jossa he tutkivat julkisen hallinnon tuottaman aineiston metadataa. Aineistoksi on valittu kolme avointa dataa sisältävää arkistoa: GovData.de Saksasta, data.gov.uk Isosta-Britanniasta ja yleiseurooppalainen publicdata.eu. Saksalainen arkisto sisälsi tutkimuksen aikaan 2642, brittiläinen 9382 ja eurooppalainen 20099 metadatatietuetta. Publicdata.eu haravoi dataa myös kahdesta muusta tutkimuksessa mukana olleesta arkistosta.

Seitsemässä Ochoan ja Duvalin metriikasta käyttöön on valittu viisi: täydellisyys, painotettu täydellisyys, tarkkuus, informaation runsaus ja saavutettavuus. Loppuja metriikoita ei käytetty, koska ne olisivat vaatineet sellaista tietoa, jota kyseisistä arkistoista ei saanut. Metadatan laatua mitattiin kaikkien metadatatietueiden osalta. Metriikoita implementoidessaan Reiche ja Höfig tekivät niihin pieniä muutoksia. Tarkkuutta mitattaessa otettiin huomioon vain *Formaatti*-kenttä, johon tallennetaan aineiston fyysinen tai digitaalinen ilmiasu. Tämä siksi, että kuvailukentän käy-

tössä on huomattu toistuvia virheitä: *Formaatti*-kentän arvo on usein ristiriidassa tallenteen todellisen ilmiasun kanssa.

Suurimmat laatuarvot kaikissa kolmessa arkistossa saatiin täydellisyyden ja painotetun täydellisyyden osalta. Painotettu täydellisyys on sikäli tärkeä mittari, että metadatatietue voi saada siinä korkean laatuarvon, vaikka kuvailukenttiä olisikin käytetty vähän (eli täydellisyyden arvo on matala), koska käytetyt kuvailukentät ovat olleet olennaisia. Tarkkuuden osalta tuloksissa ilmeni paljon hajontaa. Kaikissa kolmessa arkistossa noin puolessa metadatatietueista *Formaatti*-kentän arvo oli väärä verrattuna tallenteen oikeaan ilmiasuun.

Tutkijoiden mielestä tulevissa tutkimuksissa olisi tärkeää laajentaa tarkkuuden mittaamista myös muihin kuvailukenttiin, jolloin metriikka kertoisi laajemmin metadatan laadusta. Matalin laatuarvo kaikissa arkistoissa saatiin informaation runsaus -metriikalla. Tähän Reiche ja Höfig näkevät useita syitä: *Kuvailu*-kentässä käytetään hyvin lyhyitä tekstinpätkiä tai kuvailu koostuu vain avainsanoista eikä yhtenäisestä kuvailutekstistä. Informaation runsauteen vaikuttaa yhtenä osatekijänä tekstin pituus, joten lyhyet kuvailut laskevat laatuarvoa.

Reiche ja Höfigin mielestä käytetyt metriikat osoittautuivat käyttökelpoisiksi huonolaatuisen metadatan tunnistamisessa. Niiden avulla voidaan esimerkiksi laatia lista tietueista, joissa metadata on huonolaatuista, jolloin niiden täydentäminen ja korjaaminen on helpompaa. Kolmen arkiston metadatan laadun välillä ei sen sijaan havaittu suuria eroja. (Reiche & Höfig 2013.)

3.3 Metadatastandardin vaikutus laatuun

Käytetyn metadatastandardin vaikutusta metadatan käyttöön ja metadatan laatuun on tutkittu muun muassa Dublin Core -metadataformaatin (DC) osalta. Formaatti on kehitetty erityisesti digitaalisten aineistojen kuvailuun. Sen kehitystyö alkoi vuonna 1995, ja siinä on ollut mukana kansainvälinen, eri alojen ammattilaisista koostuva yhteisö. DC on yksi käytetyimmistä metadatastandardeista, ja se on hyväksytty NISO (National Information Standard Organization) standardiksi (Z39.85) ja ISO-standardiksi (ISO 15836:2009).

DC-formaattia luotaessa perusajatuksena on ollut, että se mahdollistaisi aineistojen kuvailun mahdollisimman laajasti, huolimatta siitä, minkä alan aineistosta on kyse. Tavoitteena on ollut lisäksi, että se soveltuisi niin dokumentin kaltaisten tallenteiden, verkkoresurssien kuin myös esimerkiksi ilmiöiden tai tapahtumien kuvailuun. DC:n avulla aineistojen kuvailun halutaan olevan helppoa kaikille, ei vain kuvailun ammattilaisille. Merkittävä tavoite on ollut myös yleisesti ymmärrettävä semantiikka: käyttämällä elementtejä, joiden semantiikka on yleisesti tunnettua, voidaan tiedonhakua helpottaa. (Stenvall 2002.)

DC:ssa on kaksi tasoa. Niin sanottu perustaso eli DCMES (The Dublin Core Metadata Element Set) määrittelee 15 kuvailukenttää. Näitä ovat: Contributor (*Muu tekijä*), Coverage (*Kattavuus*), Creator (*Tekijä*), Date (*Aikamääre*), Description (*Kuvailu*), Format (*Formaatti*), Identifier (*Identifiointitunnus*), Language (*Kieli*), Publisher (*Julkaisija*), Relation (*Suhde*), Rights (*Oikeudet*), Source (*Lähde*), Subject (*Aihe*), Title (*Nimeke*) ja Type (*Laji*). Kaikki kentät ovat toistettavissa ja mikään kenttä ei ole pakollinen. Toinen taso on Qualified Dublin Core, joka määrittelee 15 peruskentän lisäksi kentät Audience (*Yleisö*), Provenance (*Provenienssi*) ja RightsHolder (*Oikeudenomistaja*) sekä joukon niin sanottuja kenttätarkenteita, joiden avulla kuvailusta tulee yksityiskohtaisempaa ja rajatumpaa. Suurinta osaa peruselementeistä voidaan tarkentaa niiden avulla. Esimerkiksi *Aikamääre*-kenttään voidaan liittää tieto, mistä päivästä on kyse (merkintä date.available tarkentaa ajankohdan tai aikajakson, jolloin tallenne on käytettävissä). Kuvailukenttien ja tarkenteiden lisäksi DC sisältää suosituksia esimerkiksi käytettävien sanastojen ja arvojen merkintäjärjestelmien suhteen. (Stenvall 2002; Dublin Core Metadata Initiative 2016.)

Ward (2004) tutki DC:n kuvailukenttien käyttöä avoimissa arkistoissa. Tarkoituksena oli selvittää, onko kuvailukenttiä käytetty siinä määrin kuin olisi mahdollista. Ward selvitti, mitä kuvailukenttiä on ylipäätään käytetty ja mitä ei; mitä kenttiä käytetty eniten ja mitä vähiten. Tutkimuksen kohteena oli kaikkiaan 82 arkistoa, joissa oli yhteensä 910919 tietuetta.

Analyysissa selvisi, että keskimäärin tietueissa oli käytetty 8 kuvailukenttää. Eniten käytettyjä olivat *Tekijä*, *Identifiointitunnus*, *Nimeke*, *Aikamääre* ja *Laji*. Näitä viittä kuvailukenttää oli käytetty 71 prosentissa kaikista tietueista. Tutkituista arkistoista 54 prosenttia käytti vain *Tekijä* ja *Identifiointitunnus* -kenttiä noin puolessa kaikista metadatatietueistaan.

Ward huomauttaa, että vaikka kaikkia 15:tä kenttää ei ole tarkoituskaan käyttää kaikissa tietueissa, olisi odotettavaa, että kuvailukenttiä olisi käytetty kattavammin. Ward jättää tulevan tutkimuksen tehtäväksi selvittää, miksi formaatti on niin sanotusti alikäytetty. Syy tähän saattaa Wardin mukaan olla metadataformaatissa itsessään tai tahoissa, jotka metadatan luovat. (Ward 2014.)

DC:n vaikutusta metadatan laatuun on tutkittu erilaisin tutkimusasetelmin (muun muassa kyselytutkimuksilla ja käyttäjätesteillä). Park ja Childress (2009) huomauttavat, että DC:n tiettyihin kuvailukenttiin liittyvät käsitteelliset monitulkintaisuudet ja semanttinen päällekkäisyys johtavat osaltaan epäyhtenäiseen ja epätarkkaan metadataan. Eli DC:n kuvailukenttien määrittelyt ja ohjeistus kenttien käytöstä eivät ole niin yksiselitteisiä ja ymmärrettäviä kuin standardia luotaessa oli tarkoitettu.

Useammassa tutkimuksessa on havaittu, että tietyt kuvailukentät aiheuttavat ongelmia. Tällaisia kenttiä ovat *Laji*, *Kuvailu*, *Formaatti*, *Lähde* ja *Suhde* (Park & Childress 2009, Godby ym. 2003). Ongelmia aiheutti muun muassa se, että kenttiä koskeva ohjeistus koettiin liian monitulkintaiseksi tai eri kenttiin tallennettava tieto koettiin päällekkäiseksi, jolloin ei oltu varmoja, mihin kenttään tieto pitäisi tallentaa. Esimerkiksi *Laji* ja *Formaatti* -kenttiä on tutkimuksessa havaittu käytettävän samantyyppisen tiedon tallentamiseen. *Tekijä*, *Muu tekijä* ja *Kustantaja* -kenttien osalta näihin kenttiin tallennettava tieto on koettu päällekkäiseksi tai ei ole hahmotettu, miten nämä kentät eroavat semanttisesti toisistaan.

Myös DC:n käyttöohjeet voivat olla osasyynä metadatan ongelmiin. Chutturin (2014) tutkimuksessa verrattiin erityyppisten aineistojen metadatan virheellisyyksiä, kun data luotiin erilaisten ohjeistusten avulla. Toisellä ryhmällä oli käytössään vain DC:n kuvailukenttien lyhyet määrittelyt. Toisella ryhmällä oli määrittelyjen lisäksi käytössään DC:n laajempi ohjeistus kuvailukenttien käyttöön (niin sanottu Dublin Core Best Practice Guidelines).

Tuloksena oli, että pelkkien kuvailukenttien määrittelyjen avulla luotu metadata sisälsi huomattavasti enemmän virheitä kuin data, joka luotiin käyttäen avuksi yksityiskohtaisempaa ohjetta. Tarkempi ohjeistus ei kuitenkaan estänyt virheellistä metadataa: yhtäkään täysin virheetöntä metadatatietuetta ei tutkimuksessa luotu. Kuvailtavalla aineistolla ei puolestaan ollut vaikutusta metadatan virheellisyyksiin. (Chuttur 2014.)

3.4 Yhteenvetoa tutkimuksista

Sekä manuaalisilla että automaattisilla menetelmillä laatua tutkittaessa on metadatasta löydetty puutteita. Julkaisuarkiston metadatan laadun ongelmat ovat hyvin samantyyppisiä Parkin ja Richardin (2011) sekä Kurtzin (2010) tutkimuksissa. Molemmissa tutkimuksissa aineistona oli otos koko kokoelmasta, ja niissä keskityttiin vain tiettyihin piirteisiin. Laajempi laatukriteeristö olisi tuonut varmasti esiin myös muunlaisia laadun ongelmia, joita muissa tutkimuksissa on havaittu. Pienen otoksen perusteella on myös mahdotonta sanoa, kuinka yleistettävissä tulokset ovat kaikkiin kokoelman metadatatietueisiin.

Buin ja Parkin (2006) tilastollinen analyysi keskittyi lähinnä kuvailukenttien käytön eli täydellisyyden laskemiseen. Täydellisyys tosin katsotaan merkittäväksi laatupiirteeksi: jos kuvailukenttiä on käytetty niukasti, ei metadata yksinkertaisesti voi kuvata kohdettaan kovin monipuolisesti. Buin ja Parkin käyttämällä tilastollisella menetelmällä olisi voinut varmasti tutkia metadatan laatua laajemminkin, mutta jatkotutkimusta ei ole ilmeisesti tehty. Kuvailukenttien käyttöön keskit-

tyi myös Wardin (2004) tutkimus, jossa havaittiin, että DC-formaatin kenttiä ei käytetä niin kattavasti kuin olisi mahdollista. Keskimäärin vain noin puolet kentistä oli käytössä.

Ochoan ja Duvalin (2009) kehittelemät metriikat mahdollistavat laadun tutkimisen automaattisesti, jolloin analyysissa voidaan ottaa huomioon kaikki metadatatietueet. Vaikka Ochoa ja Duval testaavat metriikoitaan käytännössä, jää kysymys tulosten merkityksellisyydestä hieman auki. Miten laskettuja laatuarvoja pitäisi oikeastaan tulkita? Milloin metadata voidaan arvioida huonolaatuiseksi, milloin laadultaan hyväksi? Reiche ja Höfig (2013) soveltavat omassa tutkimuksessaan Ochoan ja Duvalin metriikoita käytäntöön. Kumpikin tutkijapari tulee samaan tulokseen: metriikoiden avulla pystytään tunnistamaan metadatatietueiden joukosta sellaiset, joiden metadatan laadussa on ongelmia.

Toisaalta useassa tutkimuksessa on havaittu, että tietyt DC:n kuvailukentät ovat ongelmallisia (muun muassa Park & Childress 2009). Esimerkiksi tiettyjä kenttiä käytetään samantyyppisen tiedon tallentamiseen, jolloin vähintään toista näistä kentistä käytetään formaatin ohjeistuksen vastaisesti. Myös DC:n ohjeistuksessa on havaittu puutteita: metadata sisältää virheitä, vaikka ohjeistus olisikin yksityiskohtaista (Chuttur 2014).

Metadatan laatuun näyttäisi siis vaikuttavan osaltaan itse metadatastandardi ja sen käytön ohjeistuksen epäselvyydet ja monitulkintaisuudet. Tutkielmassani en etsi niinkään syitä siihen, miksi metadatan laadussa on puutteita, vaan haluan selvittää, millaista metadata on laadultaan ja millaisia laatuun liittyvät ongelmat ovat.

4 TUTKIMUSASETELMA- JA MENETELMÄT

Tässä työssä tutkitaan suomalaisten julkaisuarkistojen metadatan laatua automaattisia menetelmiä käyttäen. Tutkielman tutkimuskysymys on:

Millaista on julkaisuarkistojen opinnäytetöiden metadatan laatu automaattisilla menetelmillä mitattuna?

Alaluvussa 4.1 kuvataan tutkimusympäristöä eli selvitetään, mitä ja millaisia tutkimuksen kohteena olevat julkaisuarkistot ovat muun muassa aineistojen määrän suhteen. Alaluvussa 4.2 kuvataan aineiston keruumenetelmää ja itse tutkittavaa aineistoa. Alaluvussa 4.3 avataan käytettyä tutkimusmenetelmää ja viimeisessä alaluvussa sitä, miten metadatan laatua tässä työssä analysoitiin.

4.1 Julkaisuarkistot

Lynchin (2003) määritelmän mukaan julkaisuarkisto on palvelujen kokonaisuus, jonka avulla organisaatiossa luotuja digitaalisia aineistoja hallitaan ja jaetaan. Perusajatus on, että arkiston aineistot ovat avoimesti kaikkien saatavilla. Näitä aineistoja ovat muun muassa opinnäytetyöt, artikkelit, sarjajulkaiset, opetusmateriaalit ja raportit. Julkaisuarkistoja on erilaisilla organisaatioilla kuten yliopistoilla ja tutkimuslaitoksilla. Vastuu julkaisuarkiston toiminnasta voi olla organisaation eri osastoilla, kuten kirjastolla. Lynch korostaa, että julkaisuarkisto on pohjimmiltaan osoitus organisaation sitoutumisesta arkistoon tallennettujen aineistojen pitkäaikaisen säilymisen, jakamisen ja avoimen saatavuuden turvaamiseen. (Lynch 2003.)

Maailmanlaajuisesti avoimia julkaisuarkistoja on OpenDOAR-sivuston mukaan tällä hetkellä 3238 kappaletta, Suomessa 16. Sivustolla on tietyt kriteerit, joiden perusteella se hyväksyy julkaisuarkiston listalleen. Tärkeimpiä kriteereitä on, että julkaisuarkisto on avoimesti saavutettavissa, se sisältää akateemista tutkimusaineistoa ja tarjoaa aineistonsa kokotekstinä kaikkien saataville. OpenDOAR-sivustolla näkyvät ne arkistot, joita organisaatiot ovat sinne ehdottaneet, ja jotka on arvioinnin jälkeen hyväksytty. (Open DOAR 2016.)

Tässä työssä tutkimuksen kohteena ovat kaikki suomalaiset julkaisuarkistot, joihin on talletettu opinnäytetöitä. Tällä rajauksella esimerkiksi tutkimuslaitosten julkaisuarkistot jäävät tämän tutkimuksen ulkopuolelle. Julkaisuarkistoja on tässä työssä mukana kaikkiaan 14.

Tutkimani julkaisuarkistot ovat ainakin jossain määrin noudattaneet aineiston kuvailussa Dublin Coreen pohjautuvaa Kansallista metadataformaattia elektronisille opinnäytteille (J. Ilva, sähköpostiviesti, 4.4.2016). Formaatissa on pyritty tarkentamaan DC:n monikäsitteistä semantiikkaa, jotta se soveltuisi paremmin nimenomaan opinnäytteiden metadatan julkaisemiseen. Standardissa on yhteensä 24 kenttää, joista 6 pakollista ja 18 vapaaehtoista. Korkeakoulut voivat laajentaa kuvailukenttien joukkoa, mikäli tähän on tarvetta. (Ilva ym. 2006.)

Oulun yliopiston Jultika- ja Itä-Suomen yliopiston UEF Electronic Publications -julkaisuarkistoja lukuunottamatta kaikkien tutkimuksessa mukana olevien arkistojen taustalla on DSpace-ohjelmisto. Jultikan taustalla oleva ohjelmisto on puolestaan Fedora, UEF:n taustaohjelmisto jäi epäselväksi. DHanken-julkaisuarkiston aineistot ovat saatavissa Helsingin yliopiston Helda-arkiston kautta. Maanpuolustuskorkeakoulun, LUTPubin, Turun yliopiston ja Åbo akademin aineistot ovat saatavilla Kansalliskirjaston Doria-julkaisuarkiston (www.doria.fi) kautta, jossa ne muodostavat omat kokoelmansa. Muut julkaisuarkistot ovat korkeakoulujen omia. Theseus-julkaisuarkistoon tallennetaan nähdäkseni kaikkien Suomen ammattikorkeakoulujen opinnäytetyöt.

Metadatan luomisen ja tallentamisen käytännöistä kysyttiin joiltakin julkaisuarkistoilta sähköpostin välityksellä (liite 1), jotta saataisiin yleiskuva siitä, miten metadataa luodaan ja miten sen laatua valvotaan. Sähköposti lähetettiin korkeakoulujen kirjastoihin (ammattikorkeakoulujen osalta kahteen isoon yksikköön), koska kaikissa organisaatioissa julkaisuarkiston toiminta on ainakin joiltakin osin kirjastojen vastuulla. Kyselyyn vastasi 8 kirjastoa. Kysely osoitti, että käytännöt vaihtelevat julkaisuarkistojen välillä jonkin verran. Yleisin käytäntö tallentamisen osalta on, että työn tekijä tallentaa metadatan itse verkkolomakkeen kautta. Vain yhdessä arkistossa metadata luodaan kokonaan kirjastossa. Kaikissa kyselyyn vastanneissa arkistoissa metadata tarkistetaan kirjaston toimesta ainakin jossain määrin ennen työn julkaisua. Muun muassa kirjoitusvirheet pyritään karsimaan pois. Kaikissa julkaisuarkistoissa oli olemassa edes jonkinlaisia yhteisiä suosituksia tai toimintatapoja esimerkiksi käytettävien kuvailukenttien osalta. Huomionarvoista on, että metadatan luomisessa on joissain arkistoissa eroja sen mukaan, minkä tason opinnäytetyöstä on kyse. Esimerkiksi pro gradujen ja lisensiaatintöiden metadatan tallennuksen käytännöt eroavat väitöskirjojen metadatan tallennuksesta. Väitöskirjojen metadata tulee joissain arkistoissa kirjaston tietokannasta julkaisuarkistoon, jolloin metadatan on luotu kirjastossa, mutta gradujen metadatan luo työn tekijä itse.

Taulukossa 1 on esitetty tutkimuksen kohteena olevat julkaisuarkistot ja niiden aineistomäärät. Suluissa olevat lyhenteet joidenkin julkaisuarkistojen nimen kohdalla eivät ole virallisia, vaan niitä käytetään tässä työssä tilan säästämiseksi.

Organisaatio	Julkaisuarkiston nimi	Aineistoja kaikkiaan	Opinnäytetöitä
Aalto-yliopisto	Aaltodoc	18538	8620
Hanken Svenska Handelshögskolan	DHanken	2570	2019
Helsingin yliopisto	Helda	48963	27067
Oulun yliopisto	Jultika	7778	4417
Jyväskylän yliopisto	Jyx	41623	17027
Lapin yliopisto	Lauda	2771	2559
Lappeenrannan teknillinen yliopisto	LutPub	10206	9860
Maanpuolustus-korkeakoulu	Maanpuolustuskorkeakoulun julkaisuarkisto (MPKK)	2147	1783
Tampereen yliopisto	TamPub	25336	22965
ammattikorkeakoulut	Theseus	94524	92334
Tampereen teknillinen yliopisto	TUTDPub	9871	9082
Itä-Suomen yliopisto	UEF Electronic Publications (UEF)	-	6878
Turun yliopisto	Turun yliopiston julkaisuarkisto (UTU)	9268	4063
Åbo akademi	Åbo akademin julkaisuarkisto (Åbo)	844	733
Yhteensä			209407

Taulukko 1: Organisaatiot, niiden julkaisuarkistot ja tietuemäärät. Tietuemäärät ovat aineiston lataamishetkeltä. UEF:n arkistosta aineistojen kokonaismäärää ei saanut selville.

4.2 Aineiston keruu ja aineiston kuvaus

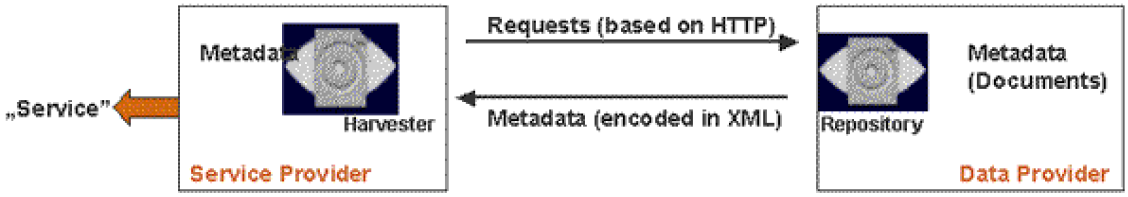
Tutkimuksen aineisto koostui kaikista niistä opinnäytetöiden metadatatietueista, joita julkaisuarkistoihin oli keruuhetkellä tallennettu. Rajasin tutkimusaineiston koskemaan arkistoissa olevia opinnäytetöitä. Opinnäytetyöt olivat kandidaatin- ja pro gradu -tutkielmia, ammattikorkeakoulujen opinnäytetöitä, väitöskirjoja, lisensiaatintöitä ja syventäviä töitä. Opinnäytetyöt olivat kaikissa arkistoissa määrältään suurin aineistoryhmä.

Keskityin analyysissani opinnäytetöihin, koska niiden metadata on varmuudella luotu organisaatiossa itsessään –niin sanotusti primääriluetteloitu– eikä ladattu jostain toisesta palvelusta. Metadatan saattoi olettaa siis olevan mahdollisimman hyvälaatuista, vaikka metadatan olisikin luonut

työn tekijä itse. Tutkimusaineiston rajaaminen opinnäytetöihin oli perusteltua myös siitä syystä, että joissakin julkaisuarkistoissa lähes kaikki aineisto oli pelkästään opinnäytetöitä. Aineisto eri arkistojen välillä oli tyypiltään myös melko epäyhtenäistä: joissakin arkistoissa on paljon erityyppisiä aineistoja, toisissa suurimmaksi osaksi vain opinnäytteitä. Kun aineisto rajattiin vain yhteen aineistotyyppiin, varmistettin, että metadatan laatukin oli yhtenäisempää.

Aineiston haravointi tapahtui OAI-PMH-standardin määrittelemällä tavalla. Open Archives Initiative (OAI) on organisaatio, joka kehittää ja edistää standardeja yhteentoimivuuden takaamiseksi. Tavoitteena on, että digitaalisten arkistojen sisältämät aineistot olisivat siirrettävissä järjestelmien välillä tehokkaasti.

Aineistojen metadatan haravoimiseksi on kehitetty OAI-PMH-standardi (The Open Archives Initiative Protocol for Metadata Harvesting). Standardi määrittää menetelmät ja protokollat, joilla metadata on saatavissa arkistoista. OAI-PMH on siis avoin rajapinta, jonka kautta metadatatiedostojen haravointi on mahdollista.



Kuva 2: OAI-PMH-standardin perustoiminnot

Tutkimusaineisto ladattiin marraskuussa 2015 (Aaltodoc, Helda, Lauda, Tampub, Theuseus, TutDPub, UEF ja UTU) ja lokakuussa 2016 (DHanken, Jultika, Jyx, LUTPub, Mpkk ja Åbo). Opinnäytteiden metadatan haravoiminen tapahtui seuraavasti.

Ensin poimittiin kohdekokoelmien tunnisteet selaimessa. DSpace-ohjelmistolla toteutetuissa julkaisuarkistoissa tunnisteet muodostavat kokoelman URL-osoitteen loppuosan. Esimerkiksi Jyväskylän yliopiston pro gradu -tutkielmia sisältävän kokoelman URL-osoite on https://jyx.jyu.fi/dspace/handle/123456789/390, jossa kokoelmatunnus on 123456789/390. Tässä muodossa kokoelmatunnusta ei kuitenkaan voi vielä käyttää aineistojen lataamisessa, vaan tunniste täytyy täsmätä OAI-PMH-rajapinnan kautta ladatun kokoelmalistauksen sisältämien tunnisteiden muotoon.

Rajapinta antoi kokoelmalistauksen komennolla *ListSets*. Näin saadusta listasta etsittiin rajapinnan käyttämät opinnäytekokoelmia vastaavat kokoelmatunnukset, jotka oli poimittu verkkosi-

vuilta (esimerkiksi edellä mainittua pro gradu -kokoelman URL-osoitteesta erotettavaa tunnistetta vastaava tunnus on hdl_123456789_390)

Tyypillisesti julkaisuarkiston metadatan voi ladata OAI-PMH-rajapinnan kautta useassa eri muodossa. Arkiston kaikki tuetut metadataformaatit listattiin komennolla *ListMetadataFormats*. Tässä vaiheessa valittiin metadataformaatiksi se Dublin Coreen perustuva formaatti, jossa oli silmämääräisesti arvioituna eniten kuvailukenttiä tarjolla. Valitut metadataformaatit olivat perus Dublin Core (Jultika, UEF), Qualified Dublin Core (Aaltodoc, Jyx, TutDPub), RDF (DHanken, Helda) ja Kansalliskirjaston metadataformaatti (Lauda, LUTPub, Mpkk, Tampub, Theseus, UTU, Åbo).

Julkaisuarkistojen rajapintoihin ei ollut toteutettu tukea kaikille käytetyille kuvailukentille. Tämä tarkoitti, että rajapinnan kautta ei välttämättä saanut ladattua kaikkea arkistoon tallennettua dataa, jolloin metadatatietueessa voi olla käytetty kuvailukenttiä, joita ei tässä työssä olevassa aineistossa ollut käytettävissä. Tämä seikka täytyi ottaa aineiston analyysissa huomioon jokaisen arkiston kohdalla erikseen.

Opinnäytetöiden metadata ladattiin *ListRecords*-komennolla. OAI-PMH-rajapinta antaa ladata kerrallaan vain rajatun verran tietueita. Siksi kuvailutietojen lataamisessa on käytettävä haravointiohjelmaksi kutsuttua sovellusta, jolla koko kokoelman kuvailutiedot voidaan ladata kerralla. Tässä tapauksessa käytettiin oaimi-ohjelmaa (https://github.com/miku/oaimi), joka on yksinkertainen komentorivisovellus. OAI-PMH-rajapinta antoi metadatatietueet XML-muodossa.

Alla on esimerkki XML-muotoisesta metadatatietueesta Lauda-julkaisuarkistossa (tietueesta poistettu englanninkielinen description-kenttä ja lyhennetty suomenkielisen description-kentän tekstiä). Laudan aineisto on ladattu Kansalliskirjaston metadataformaatissa. `Element`-elementti kertoo kuvailukentän nimen, `qualifier`-attribuutti kenttään mahdollisesti liittyvän tarkenteen, `value`-attribuutissa on kentän arvo.

```
<metadata>
        <identifier type="handle" value="10024/59509"/>
        <type name="item"/>
        <link href="http://lauda.ulapland.fi/handle/10024/59509"/>
        <field element="contributor" language="none" qualifier="aut-
        hor" schema="dc" value="Vaattovaara, Virpi"/>
        <field element="date" language="none" qualifier="accessioned"
        schema="dc" value="2010-09-15T10:11:40Z"/>
        <field element="date" language="none" qualifier="availablesc-
        hema="dc" value="2010-09-15T10:11:40Z"/>
        <field element="date" language="none" qualifier="issued" sche-
        ma="dc" value="2005"/>
        <field element="date" language="none" qualifier="dateaccep-
        ted" schema="dc" value="2005"/>
        <field element="identifier" language="none" qualifier="uri"
        schema="dc" va-
        lue="http://lauda.ulapland.fi/handle/10024/59509"/>
        <field element="identifier" language="none" qualifier="urn"
        schema="dc" value="URN:NBN:fi:ula-201011261027"/>
        <field element="description" language="en" qualifier="abst-
        ract" schema="dc" value="Tutkimuksessani kuvaan teorian ja
        käytännön kohtaamisia yliopisto-opiskelussa opiskelijoiden
        silmin. Tarkastelen verkko-opiskelijoiden..."/>
        <field element="language" language="none" qualifier="iso"
        schema="dc" value="fi"/>
        <field element="type" language="none" qualifier="ontasot"
        schema="dc" value="fi=Lisensiaatintyö|en=Licentiate Thesis|"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="narratiivisuus"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="vuorovaikutus"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="reflektiivisyys"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="oppimiskäsitykset"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="yliopistopedagogiikka"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="verkko-opetus"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="verkko-oppiminen"/>
        <field element="subject" language="none" qualifier="ysa" sche-
        ma="dc" value="verkko-opiskelu"/>
        <field element="contributor" language="none" schema="dc" va-
        lue="fi=Kasvatustieteiden tiedekunta|en=Faculty of
        Education|"/>
        <field element="rights" language="-" schema="dc" value="ope-
        nAccess"/>
        <field element="title" language="none" schema="dc" va-
        lue="Verkko-opiskelijoiden kokemuksia yliopisto-opiskelusta:
        kertomuksia teoriasta ja käytännöstä"/>
        <field element="type" language="none" schema="dc" value="li-
        centiateThesis"/>
        <file bundle="ORIGINAL" href="http://lauda.ulapland.fi/bitst-
        ream/10024/59509/1/4750.pdf" name="4750.pdf" sequence="1"
        size="416189" type="application/pdf"/>
</metadata>
```

Kuva 3: Esimerkki XML-muotoisesta metadatatietueesta.

4.3 Tutkimusmenetelmänä kuvaileva tilastoanalyysi

Tilastollisten menetelmien avulla pyritään löytämään empiirisistä, kokemusperäisistä ilmiöistä säännönmukaiset sekä toisaalta satunnaiset tekijät, arvioimaan ilmiöiden välisiä yhteyksiä sekä erottamaan ilmiöt toisistaan (Metsämuuronen 2009, 35).

Tilastollinen tutkimus on tutkimusotteeltaan kvantitatiivista eli määrällistä, ja sen avulla selvitetään lukumääriin ja prosenttiosuuksiin liittyviä kysymyksiä. Asioita kuvataan numeeristen suureiden avulla, ja tuloksia havainnollistetaan taulukoilla ja kuvioilla. Tilastollista tutkimusta voidaan luonnehtia kartoittavaksi: sen avulla saadaan yleensä kartoitettua olemassa oleva tilanne, mutta ei pystytä selvittämään asioiden syitä. Tilastollisessa tutkimuksessa tutkimuksen kohteena olevia tutkimusyksiköitä kutsutaan tilastoyksiköiksi. (Heikkilä 2008.)

Tilastoyksiköistä kerätään tietoja mittaamalla. Heikkilä (2008, 81) toteaa, että mittaaminen on tilastollisessa tutkimuksessa laajempi käsite kuin esimerkiksi fysikaalisten suureiden arvojen mittaaminen. "Tutkimuksessa mittaamista on kaikki, missä voidaan nähdä eroja ja antaa tutkimusyksiköille jonkinlaisia symboleja eroja luonnehtimaan", Heikkilä kirjoittaa (2008, 183).

Mittaaminen on tilastoyksiköiden ominaisuuksien määrittämistä, jossa tarkasteltavaan ominaisuuteen liitetään mittaluku tai -symboli. Mitattavia ominaisuuksia kutsutaan muuttujiksi, joita voivat olla esimerkiksi ikä, sukupuoli tai ansiotulot. Muuttujat luokitellaan kvantitatiivisiksi eli määrällisiksi tai kvalitatiivisiksi eli laadullisiksi. Tilastollisessa tutkimuksessa voidaan laadullinen ominaisuus mitata määrällisesti ja päinvastoin. (Heikkilä 2008, 14, 81.)

Mittaamisen yhteydessä pitäisi pohtia mittarin validiteettia ja reliabiliteettia. Käytetyn mittarin on mitattava sitä asiaa, mitä sillä halutaan mitata eli sen on oltava validi. Mittarin on myös mitattava aina ja kokonaisuudessaan samaa asiaa eli oltava johdonmukainen ja luotettava, reliaabeli. (Metsamuuronen 2009, 64–65.)

Mitta-asteikon käsitteellä kuvataan tilastollisten muuttujien mittaustason ilmaisukykyä. Mitta-asteikot jaetaan niiden mittaustason mukaan neljään ryhmään.

Luokitteluasteikon tasoisten muuttujien arvot voidaan jakaa vain luokkiin, mutta niitä ei voida asettaa järjestykseen (esimerkkinä sukupuoli). Järjestysasteikon tasoisia muuttujia voidaan laittaa mitattavan ominaisuuden mukaiseen luonnolliseen järjestykseen, mutta mittausten etäisyyttä toisistaan ei voida tarkasti mitata, koska arvot eivät välttämättä ole tasavälein (esimerkkinä mielipidemittaus). Luokittelu- tai järjestysasteikon tasoisten muuttujien arvoilla ei voi tehdä laskutoimituksia. (Heikkilä 2008, 81–82.)

Välimatka-asteikolla mittausarvojen etäisyys toisistaan tiedetään, mutta asteikolla ei ole yksiselitteistä nollakohtaa. Asteikolla pystytään mittaamaan yksittäisten luokkien tai havaintoarvojen eroja (esimerkkinä lämpötila). Suhdeasteikolla on absoluuttinen nollapiste, ja sillä voidaan laskea lukujen suhteita (esimerkiksi lasten lukumäärä, tuotteen hinta). (Heikkilä 2008, 81–82.)

Tilastollisen tutkimuksen analyysitavat voidaan jakaa tilastolliseen päättelyyn ja kuvailevaan tilastoanalyysiin. Tilastollisessa päättelyssä tavoitteena on luotettavien johtopäätösten tekeminen perusjoukosta otoksen perusteella. Päättelyn avulla pyritään arvioimaan, kuinka hyvin otoksesta saadut tulokset ovat yleistettävissä koskemaan koko perusjoukkoa. Kuvailevassa tilastoanalyysissa pyritään kuvailemaan ja tiivistämään jonkin määrällisen muuttujan jakaumaa tai useamman muuttujan yhteisvaihtelua. Pyrkimyksenä ei kuitenkaan ole tehdä tulosten pohjalta yleistyksiä laajempaan perusjoukkoon, kuten tilastollisessa päättelyssä. (Holopainen 2008, 165.)

Kuvailevassa tilastoanalyysissa muuttujan arvoissa oleva informaatio pelkistetään muuttujaa kuvaileviin tunnuslukuihin. Suurtenkin aineistojen tieto saadaan tunnuslukujen avulla tiiviseen muotoon, mutta osa informaatiosta häviää. Tunnusluvut jaetaan keskilukuihin ja hajontalukuihin. Keskilukuja ovat esimerkiksi keskiarvo, moodi ja mediaani. Hajontalukujen (esimerkiksi arvojen vaihteluväli, ala- ja yläneljännes, keskihajonta) avulla ilmaistaan, kuinka paljon mittaustulokset vaihtelevat keskiarvon ympärillä. Mitä pienempi hajonta on, sitä lähempänä havainnot ovat toisiaan tai keskimääräistä arvoa. Se, mitä tunnuslukuja käytetään aineiston kuvailuun, riippuu muuttujien mitta-asteikosta. (Heikkilä 2008, 82.)

Jos tutkimuksen kohteena on otoksen sijaan jokainen perusjoukon jäsen, puhutaan kokonaistutkimuksesta. Kokonaistutkimus kannattaa tehdä, mikäli perusjoukko on kooltaan pieni tai esimerkiksi siinä tapauksessa, että mitattava ominaisuus vaihtelee suuresti. (Heikkilä 2008, 33.)

Tutkielmani on kokonaistutkimus eli tutkimuksen kohteena ovat kaikki julkaisuarkistoihin tallennetut opinnäytetöiden metadatatietueet. Tutkimusote on kvantitatiivinen, ja analyysimenetelmänä käytetään kuvailevaa tilastoanalyysia. Kun mittaus kohdistetaan kaikkiin julkaisuarkistojen metadatatietueisiin, voidaan olettaa, että metadatan laadusta saadaan luotettava kuva. Koska ei ole olemassa raja-arvoja, joiden perusteella voitaisiin sanoa, milloin metadata on hyvälaatuista ja milloin ei, täytyy tutkimuksessa tyytyä kuvaamaan, mikä tilanne metadatan laadun osalta on mittaushetkellä kussakin julkaisuarkistossa.

Tutkielman tilastoyksiköitä ovat julkaisuarkistoissa olevat opinnäytetöiden metadatatietueet. Metadatan laadun mittaamiseen käytetään Ochoan ja Duvalin (2009) kehittämiä metriikoita. Nämä metriikat ovat siis mittareita, joilla laatua mitataan. Ochoan ja Duvalin metriikat on valittu, koska niitä on testattu aiemmissa tutkimuksissa myös suuriin aineistomääriin, ja niiden avulla on

saatu eroteltua huonolaatuiset metadatatietueet. Ne soveltuvat myös erilaisten aineistojen ja järjestelmien metadatan laadun analysointiin. Mittarit on operationalisoitu metadatan laatu -käsitteen määrittelyn perusteella.

Eri ominaisuuksia mittaavat metriikat tuottavat jokaiselle metadatatietueelle oman mittalukunsa, jota tässä tutkielmassa kutsutaan laatuarvoksi. Tutkielmassa käytetty mitta-asteikko on metadatan laatuarvon osalta tasoltaan suhdeasteikko. Mittareiden tuottamilla arvoilla on siis absoluuttinen nollapiste, niihin voidaan kohdistaa laskutoimituksia ja niitä voidaan suhteuttaa keskenään. Suhdeasteikollisia arvoja voidaan myös kuvata kaikilla tunnusluvuilla.

4.4 Metadatan laadun mittaaminen

Metadatan laatua mitattiin täydellisyyden ja painotetun täydellisyyden metriikoilla. Näillä mittareilla saatiin tietoa siitä, kuinka paljon kuvailukenttiä on metadatatietueissa käytetty eli kuinka kattavaa metadata on.

Nämä metriikat valittiin, koska niihin tarvittava data on saatavilla julkaisuarkistojen metadatatietueista. Kyseisten metriikoiden valintaa tuki myös se, että täydellisyys on tarkkuuden ja johdonmukaisuuden ohella Parkin (2009) mukaan metadatan laadun piirteitä, joiden sisällöstä ja tarkoituksesta ollaan eri tutkijoiden kesken yksimielisimpiä. Täydellisyyden ja painotetun täydellisyyden mittaaminen antaa jo suhteellisen kattavan kuvan metadatan laadusta. Jos kuvailukenttiä on käytetty vähän, ei metadata voi antaa tallenteesta kovinkaan kattavaa tai tarkkaa kuvaa eikä näin ollen olla laadukasta.

Käytettyjen metriikoiden perusajatus on, että tutkimuksen kohteena olevaa metadataa verrataan johonkin metadatastandardiin. Täytyy olla siis olemassa jokin pinta, johon julkaisuarkistojen metadataa peilataan. Tässä työssä tuo peilauspinta oli Kansallinen metadataformaatti opinnäytetöille. Vaikka formaattia oli useimmassa arkistossa noudatettu vain jossain määrin, oli se paras valinta verrattavaksi standardiksi erityisesti siitä syystä, että formaatti on kehitetty nimenomaan opinnäytetöiden metadataa varten. Voidaan siis olettaa, että formaatin suosituksia noudattava metadata on laadukasta ja antaa kuvailemastaan opinnäytetyöstä kattavan kuvan. Formaattia käytettiin eräänlaisena kehikkona, joka määritti, millaista tietoa metadatan pitäisi ihannetapauksessa kuvaamastaan tallenteesta antaa.

Täydellisyys

Täydellisyydellä mitattiin, kuinka paljon kuvailukenttiä oli tietueissa käytetty suhteessa standardin suosittamaan kenttien kokonaismäärään. Kaavassa *P(i)* on 1 mikäli kenttää on käytetty ja sii-

nä on jokin arvo, 0 mikäli kenttää ei ole tai siinä on tyhjäarvo. *N* on standardin suosittelemien kenttien kokonaismäärä. Täydellisyyttä mitattiin kaavalla

$$Qcomp = \frac{\sum_{i=1}^{N} P(i)}{N} \quad (1)$$

Samaa kenttää oli tietueessa voitu toistaa, mutta mittauksessa otettiin huomioon kentän ensimmäinen esiintymä. Täydellisyys sai laatuarvon välillä 0 ja 1. Eli jos metadatatietueessa oli käytetty 14 kuvailukenttää ja standardin suosittama kenttien kokonaismäärä oli 18, sai tietue laatuarvoksi 14/18 = 0.78.

Painotettu täydellisyys

Kaikkia metadatatietueessa esiintyviä kenttiä ei voida pitää yhtä olennaisina tai tärkeinä esimerkiksi tiedonhaun kannalta. Voidaan olettaa esimerkiksi, että tallenteen nimi on useissa tapauksissa arkiston käyttäjille olennaisempi tieto kuin tallenteen koko. Painotetun täydellisyyden mittari otti huomioon nämä erot kuvailukenttien tärkeydessä.

Kullekin kuvailukentälle päätettiin painokerroin *αi*, joka oli isompi, jos kyseessä oli tiedonhaun näkökulmasta keskeinen kenttä. Kaavassa *P(i)* on 1, jos kenttää on käytetty, muuten 0. Painotettua täydellisyyttä mitattiin kaavalla

$$Qwcomp = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i} \quad (2)$$

Mittaamista varten täytyi siis ensin päättää kunkin kuvailukentän painokerroin. Painokertoimina Ochoa ja Duval (2009) suosittelevat käyttämään mitä tahansa positiivista arvoa, joka kuvaa kunkin kuvailukentän tärkeyttä tietyn tehtävän tai kontekstin näkökulmasta. Ochoa ja Duval esittävät, että painokerroin voisi määräytyä sen perusteella, mihin kuvailukenttiin tiedonhaut ovat kyseisessä järjestelmässä kohdistuneet.

Koska tällaista dataa ei ollut julkaisuarkistojen osalta käytettävissä, kenttien painokertoimen määrittämiseen käytettiin tiedonhaun tutkimuksessa aiheesta saatuja tuloksia. Kirkland (2013) on tutkinut kirjaston tietokantaan tehtyjä tiedonhakuja ja verrannut, millaiset haut johtavat aineiston lainaamiseen eli ovat tässä mielessä onnistuneita. Tutkimuksessa selvisi, että tiedonhaussa *Nimeke*, *Tekijä*, *Aihe* ja *Kuvailu*-kentät ovat käyttäjille tässä mielessä hyödyllisimpiä (Kirkland, 2013). Voidaan hyvällä syyllä olettaa, että myös julkaisuarkistoista tietoa haettaessa näitä samoja

kenttiä käytetään useimmiten, joten niitä voidaan nimittää tiedonhaun kannalta merkittäviksi kentiksi.

Painotettua täydellisyyttä laskettaessa niin sanottuja merkittäviä kenttiä olivat nämä samat neljä (*Nimeke*, *Tekijä*, *Aihe* ja *Kuvailu*). Näiden kenttien painokerroin oli 1. Kaikkien muiden kuvailukenttien painokerroin oli puolestaan 0.5. Esimerkiksi, jos metadatatietueessa oli käytetty kenttiä *Tekijä*, *Nimeke* (painokerroin 1) sekä viittä muuta kenttää (painokerroin 0.5), ja kenttien kokonaismäärä oli 18, sai tietue painotetun täydellisyyden laatuarvoksi 4.5/11 = 0.41. Tässä mittarissa kuvailukenttien määrä suhteutettiin summaan, joka saadaan, kun kukin kenttä kerrotaan ennaltamäärätyllä painokertoimella. Painotettu täydellisyys sai laatuarvon 0 ja 1 väliltä.

Kansallisen metadataformaatin kuvailukentistä jätettiin huomiotta 6 kenttää (opinnäytteen vaihtoehtoinen nimeke, opinnäytteen rinnakkaisen ilmiasun tiedot, opinnäytteen rinnakkaisen ilmiasun identifiointitunnus, sarjan nimi, missä opinnäyte on ilmestynyt, sarjan ISSN-tunnus, missä opinnäyte on ilmestynyt, opinnäytteeseen liittyvät osat). Nämä ovat tietoja, joita ei kaikista opinnäytteistä ole voitu tallentaa, joten ne rajattiin analyysin ulkopuolelle. Tämän jälkeen formaatin määrittämiksi kenttien kokonaismääräksi tuli 18 kuvailukenttää.

Metriikoiden suorittamiseen käytettiin XQuery-kieltä, joka on XML-muotoisen datan analysointiin tarkoitettu kyselykieli. Metriikat muutettin XQuery-kyselyiksi, joiden avulla jokaiselle tietueelle saatiin mitattua laatuarvo.

Jokaisesta arkistosta tutkittiin ensin XQuery-kyselyn avulla, mitä kuvailukenttiä ja niihin mahdollisesti liittyviä kenttätarkenteita kussakin arkistossa oli käytetty. Tämän jälkeen julkaisuarkistoissa käytettyjen kuvailukenttien ja Kansallisen metadataformaatin suosittelemien kuvailukenttien välille tehtiin vastaavuudet eli selvitettiin, mihin kuvailukenttään julkaisuarkistoissa on mikäkin formaatin suosittama tieto tallennettu. Mikään julkaisuarkistoista ei ollut käyttänyt Kansallista metadataformaattia täysin suosituksen mukaisesti. Se, mitä kenttää kukin julkaisuarkisto oli käyttänyt minkäkin tiedon tallentamiseen, täytyi tarkistaa jokaisen arkiston kohdalla erikseen. Tieto saattoi olla tallennettuna metadatatietueessa eri nimiseen kenttään kuin formaatissa suositellaan. Tieto saattoi puuttua kokonaan eli sitä ei ollut tallennettu ollenkaan. Kenttä saattoi myös puuttua tutkimusaineistosta rajapinnan toteutuksesta johtuen. Nämä kaikki seikat oli otettava huomioon kuvailukenttien vastaavuuksia päätettäessä.

Liitteessä 2 on kaikkien julkaisuarkistojen käyttämät kuvailukentät ja niihin mahdollisesti liittyvät tarkennekentät. Taulukossa on käytetty seuraavia merkintätapoja puuttuvien kenttien kohdalla :

- *NA*: kenttää ei ole julkaisuarkistossa käytetty ollenkaan. Näissä tapauksissa kenttä otettiin mukaan laatuarvoja laskettaessa eli huomioitiin kenttien kokonaismäärässä. Näissä tapauksissa tarkistettiin ensin, onko tieto tallennettu johonkin muuhun kenttään, joka voitaisiin mittauksessa huomioida.

- *-1*: kenttää on julkaisuarkistossa käytetty, mutta rajapinnan toteutuksen takia kenttää ei saanut datan lataamisvaiheessa ulos. Näissä tapauksissa kenttää ei ole huomioitu kenttien kokonaismäärässä (eli *-1*:llä merkityt kentät on vähennetty kenttien kokonaismäärästä).

Mittarit muutettiin XQuery-kyselyiksi, joilla laatuarvot laskettiin. Kyselyt ovat saatavissa GitHub-lähdekoodiverkkopalvelussa (https://github.com/lauraannukka/etd-metadata-quality.git).

Täydellisyyden ja painotetun täydellisyyden mittareilla saaduille laatuarvoille laskettiin seuraavat tunnusluvut: pienin arvo, suurin arvo, vaihteluväli, mediaani, aritmeettinen keskiarvo, alaneljännes, yläneljännes ja kvartiiliväli.

Pienin ja suurin arvo sekä vaihteluväli eli näiden arvojen erotus antoivat kuvan laatuarvojen vaihtelun ääripäistä arkistojen sisällä. Keskiluvuilla aritmeettinen keskiarvo (arvojen summa jaettuna arvojen lukumäärällä) ja mediaani (suuruusjärjestykseen järjestettyjen arvojen keskimäinen tai kahden keskimmäisen keskiarvo) mitattiin arvojen jakauman keskimääräisyyttä. Mediaanin erityinen hyöty keskilukuna on, että siihen eivät vaikuta muista muuttujan arvoista huomattavasti poikkeavat suuret tai pienet arvot. Muista arvoista selvästi poikkeavat arvot vaikuttavat voimakkaasti keskiarvoon varsinkin, jos tutkittavien yksiköiden määrä on pieni. Jos keskiarvo ja mediaani ovat lähellä toisiaan, viittaa se jakauman symmetrisyyteen. Jos keskiarvo on mediaania suurempi, viittaa se oikealle vinoon jakaumaan. Jos keskiarvo on mediaania pienempi, niin tämä viittaa vasemmalle vinoon jakaumaan. (Heikkilä 2008, 83–84.) Keskilukujen perusteella metadatan laadusta saatiin yleiskuva eli millä tasolla laatu keskimäärin kussakin arkistossa oli. Mitä lähempänä keskiluvut olivat 1:stä, sitä laadukkaampaa metadata oli koko arkiston tasolla.

Laatuarvojen hajontaa kuvattiin ala- ja yläneljänneksen arvoilla sekä kvartiilivälin pituudella. Korkeintaan alaneljänneksen suuruisia laatuarvoja on neljäsosa (25 %) kaikista havainnoista, ja vähintään yläneljänneksen suuruisia havaintoja on vastaavasti neljäsosa. Ala- ja yläneljänneksen väliin jää puolet kaikista laatuarvoista. Kvartiiliväli on arvojen jakauman väli alaneljännänneksestä yläneljännekseen, ja kvartiilivälin pituus ilmoittaa näiden arvojen erotuksen. Kvartiiliväliin eivät vaikuta poikkeavan suuret tai pienet arvot (niin sanotut ääriarvot), joten se on tässä mielessä hyödyllinen hajontaluku. (Heikkilä 2008, 85–86.)

Tunnuslukujen avulla kuvattiin laatuarvojen jakaumaa ja jakauman muotoa. Tunnusluvut antoivat tietoa siitä, millä tasolla kunkin julkaisuarkiston metadatan laatu oli. Laatuarvojen jakauman lisäksi selvitettiin, millainen suhde metadatan saamilla laatuarvoilla oli tiettyjen muuttujien kanssa. Mielenkiinnon kohteena oli, onko metadatan laadulla yhteys tietueen tallennusvuoteen tai opinnäytteen tasoon.

Koska kyseessä oli kokonaistutkimus, voidaan esittää epäilys siitä, onko muuttujien välisten yhteyksien (tässä tapauksessa tallennusvuoden ja opinnäytteen tason vaikutus laatuarvoihin) testaaminen tilastollisin menetelmin mielekästä. Heikkilän (2008, 190) mukaan tilastollisen testauksen menetelmin tutkitaan yleensä, voiko otoksesta saatuja tuloksia yleistää koko perusjoukkoon. Kun on tutkittu koko perusjoukko, saadut tulokset koskevat koko tutkittavaa joukkoa, joten tilastollinen testaus ei olisi tässä mielessä enää tarpeellista (Heikkilä, 190). Kokonaistutkimuksella saatujen tulosten tilastollisen testaamisen puolesta puhuu kuitenkin se, että perusjoukossa havaituista eroista voidaan testaamalla selvittää ovatko ne satunnaisia vai systemaattisia (Pitkänen 1994, 13; tässä Heikkilä 2008, 190).

Vaikka kyseessä olikin kokonaistutkimus, testattiin muuttujien välisten riippuvuuksien merkitsevyyttä tilastollisin menetelmin. Tilastollisen testauksen vaiheisiin kuuluvat hypoteesin asettaminen, otoksen poimiminen (mitä tässä tutkimuksessa ei siis tehty), tilastollisen testin valinta, testin suorittaminen, tulosten tulkinta ja johtopäätösten tekeminen (Heikkilä, 191). Tilastollista testiä valittaessa täytyi tehdä valinta parametrisen ja ei-parametrisen testimenetelmän välillä. Tämän tutkielman tutkimusaineiston kohdalla ei voitu olla varmoja, että kaikki reunaehdot parametrisen menetelmän käytölle täyttyivät (esimerkiksi ei voitu olla varmoja, että muuttujien arvot olisivat normaalisti jakautuneet) (Heikkila 2008, 193). Tästä syystä käytettiin ei-parametrista (niin kutsuttu robusti, vakaa) testausmenetelmää.

5 JULKAISUARKISTOJEN METADATAN LAATU

5.1 Metadatan laatu täydellisyyden näkökulmasta

Täydellisyys-metriikka mittaa, kuinka paljon kuvailukenttiä julkaisuarkistoissa on käytetty suhteessa Kansallisen metadataformaatin suosittamaan kuvailukenttien kokonaismäärään. Kenttien kokonaismäärä tässä analyysissa on korkeintaan 18. Luku vaihtelee 15 ja 18 välillä, kun huomioidaan haravointirajapinnasta johtuva kenttien puuttuminen joissakin arkistoissa (tästä kerrottu tarkemmin luvussa 4.4).

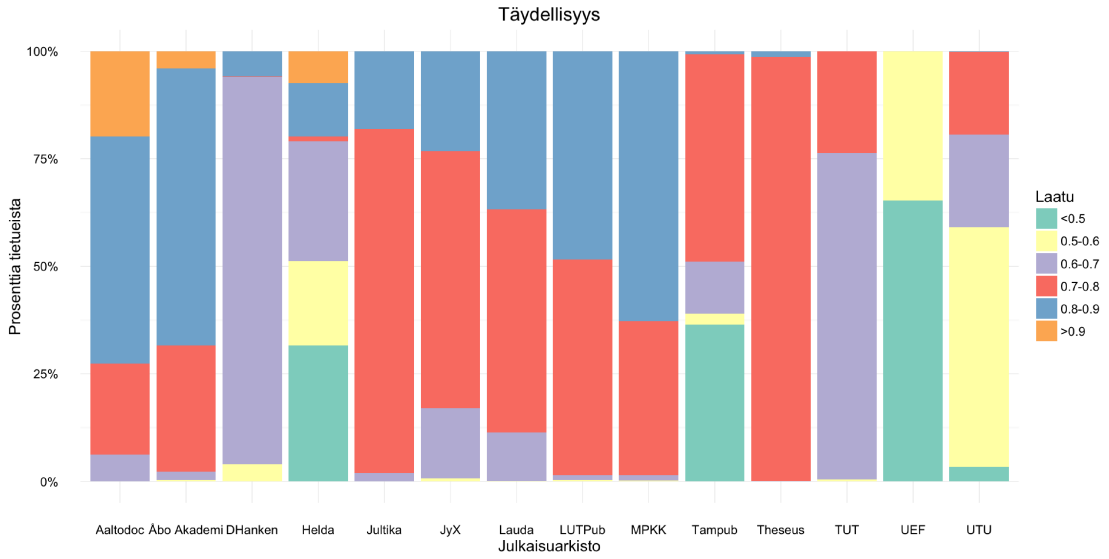
Mittauksissa saadut metadatan täydellisyyden laatuarvot osoittavat, että sekä julkaisuarkistojen sisällä että julkaisuarkistojen välillä metadatan laatu vaihtelee paikoitellen paljon. Laatuarvojen jakauma on tiivistetty laatuarvojen keskimääräisyyttä ja hajontaa kuvaaviin tunnuslukuihin (taulukko 2).

Arkisto	Min	Max	Vaihtelu-väli	Mediaani	Keskiarvo	Ala-neljännes	Ylä-neljännes	Kvartiili-väli
Aaltodoc	0,62	0,94	0,32	0,81	0,83	0,75	0,88	0,13
DHanken	0,5	0,88	0,38	0,69	0,69	0,69	0,69	0
Helda	0,25	0,94	0,69	0,56	0,62	0,44	0,69	0,25
Jultika	0,62	0,81	0,19	0,75	0,76	0,75	0,75	0
Jyx	0,35	0,88	0,53	0,76	0,74	0,71	0,76	0,05
Lauda	0,56	0,89	0,33	0,78	0,78	0,72	0,83	0,11
Lutpub	0,53	0,88	0,35	0,76	0,77	0,71	0,82	0,11
Mpkk	0,56	0,89	0,33	0,83	0,81	0,78	0,83	0,05
Tampub	0,44	0,83	0,39	0,67	0,61	0,44	0,72	0,28
Theseus	0,61	0,89	0,28	0,78	0,78	0,78	0,78	0
TutDPub	0,56	0,78	0,22	0,67	0,67	0,61	0,67	0,06
UEF	0,44	0,56	0,12	0,5	0,52	0,5	0,56	0,06
UTU	0,41	0,88	0,47	0,53	0,59	0,53	0,65	0,12
Åbo	0,59	0,94	0,35	0,82	0,81	0,76	0,82	0,06

Taulukko 2: Laatuarvojen jakaumaa kuvaavat tunnusluvut täydellisyys-mittarilla mitattuna.

Täydellisyyden laatuarvon tunnuslukujen perusteella parhaimman laatuista metadata on Aaltodocissa, Åbo:ssa ja Maanpuolustuskorkeakoulun (Mpkk) arkistossa. Näissä arkistoissa sekä mediaani että keskiarvo ovat kaikissa yli 0.8. Kolmen arkiston muita korkeampi metadatan laatu näkyy myös laatuarvojen jakaumaa kuvaavissa ositetuissa pylväissä (kuva 4). Laatuarvot on jaettu kuuteen luokkaan, joita pylväissä kuvataan eri väreillä. Kunkin luokan laatuarvojen lukumäärää

kuvataan suhteellisina prosenttiosuuksina. Mitä korkeampi laatuarvoluokan osuus on, sitä enemmän kyseiseen luokkaan kuuluvia arvoja arkistossa on.



Kuva 4: Laatuarvojen jakauma julkaisarkistossa täydellisyyden näkökulmasta.

Aaltodocin täydellisyyden laatuarvot painottuvat hieman enemmän jakauman yläpäähän eli korkeita laatuarvoja (jopa yli 0.9:n) saaneita tietueita on suhteellisen paljon. Jakauman painottumisesta keskimääräistä korkeampiin laatuarvoihin kertoo myös Aaltodocin keskiarvon ja mediaanin suhde: keskiarvo on 0.02 yksikköä suurempi kuin mediaani eli jakauman voidaan sanoa olevan vino oikealle. Åbossa ja Mpkk:ssa keskilukujen suhde on toisinpäin eli mediaani on keskiarvoa suurempi, mikä kertoo siitä, että arvojen jakauma on hieman vino vasemmalle. Keskimääräistä matalampia laatuarvoja on siis arvojen joukossa suhteellisen paljon. Tämä näkyy myös jakaumaa kuvaavista pylväistä: 0.9 ja sen yli olevia arvoja on Åbossa selvästi vähemmän kuin Aaltodocissa ja Mpkk:ssa ei ollenkaan. Kaikissa kolmessa arkistossa suurin osa laatuarvoista on luokkaa 0.8–0.9.

Tunnuslukujen perusteella näissä kolmessa arkistossa kuvailukenttiä on käytetty tasaisen paljon: Mpkk:ssa keskimäärin 15 kenttää per tietue, kun mittauksessa kenttien kokonaismäärä sen kohdalla on 18. Aaltodocissa ja Åbossa taas keskimäärin 13 kuvailukenttää laskennassa huomioon otetuista 16 kentästä on käytössä. Voidaankin olettaa metadatan näissä arkistoissa antavan tietoa tallenteesta suurimmassa osassa tapauksista monipuolisesti ja kattavasti.

Täytyy muistaa, että kuvailukenttien käytön keskimääräisiin lukuihin vaikuttaa se, mikä kuvailukenttien kokonaismäärä on täydellisyys-mittarissa kunkin arkiston kohdalla ollut. Kenttien kokonaismäärä on matalampi, mikäli datan lataamisvaiheessa rajapinta ei antanut kaikkia arkistos-

sa käytössä olevia kenttiä ulos. Näin ollen keskimääräisiä kuvailukenttien käyttömääriä ei kannata käyttää ainakaan arkistojen vertailuun.

Arkistoja, joissa keskiluvut sijoittuvat välille 0.72–0.78, on kaikkiaan viisi: Jultika, Jyx, Lauda, LUTPub ja Theseus. Vaikka Jyxiä ja Laudaa lukuunottamatta näiden arkistojen laatuarvojen jakauma on painottunut keskimääräistä korkeampiin laatuarvoihin (eli keskiarvo on suurempi kuin mediaani), eivät keskiluvut nouse yli 0.8, koska korkeimman ryhmän laatuarvoja ei ole näissä arkistoissa ollenkaan, ja toisaltaa keskimääräisiä arvoja alemman luokan (0.6–0.7) arvoja on jonkin verran. Tämä näkyy myös ositetuissa pylväissä: kaikissa viidessä arkistossa suurin osa laatuarvoista on luokkaa 0.7–0.8. Jyxissä taas keskiarvo on mediaania pienempi eli keskimääräistä matalampien laatuarvojen osuus on korostunut, ja tämä tekee arvojen jakaumasta vinon vasemmalle. Laudassa keskiarvo ja mediaani ovat saman suuruisia eli arvojen jakauma on symmetrinen. Keskimäärin kuvailukenttiä näissä arkistoissa on käytetty 12–14.

Kolmantena ryhmänä voidaan erottaa 0.6–0.7 keskiluvut saaneet arkistot: DHanken, Helda, Tampub ja TutDPub. Kuvailukenttien käytön kannalta tämän luokan keskiluvut tarkoittavat keskimäärin 10–12 kenttää per tietue. Vaikka korkein laatuarvo saattaa näissä arkistoissa olla jopa yli 0.9:n (kuten Heldassa) tai päästä lähelle sitä (kuten DHankenissa ja TutDPubissa), vaikuttavat arkistosta riippuen keskimääräistä matalampien (jopa alle 0.5:n) laatuarvojen paikoin suurikin osuus keskilukuihin alentavasti. Helda on arvojen jakauman suhteen selvästi epäsymmetrinen eli keskimääräistä korkeammat laatuarvot nostavat keskiarvon huomattavasti mediaania suuremmaksi. Heldassa suurinta laatuarvojen luokkaa on kuvasta hankala erottaa, mutta laatuarvojen frekvenssien perusteella se on luokka 0.6–0.7. Tampubissa jakauma on päinvastainen eli mediaania matalammat laatuarvot vaikuttavat keskiarvoon laskevasti ja tekevät jakaumasta vinon vasemmalle. Tampubin suurilukuisin arvojen yksittäinen luokka on kuitenkin 0.7–0.8, mutta tätä luokkaa matalampia arvoja on kaikkiaan hieman enemmän ja siksi arkiston keskiluvut eivät nouse tätä korkeammiksi. DHankenissa ja TutDPubissa jakauma on keskittynyt selkeästi mediaanin ympärille eli suurin osa laatuarvoista on 0.6–0.7.

UEF:ssa ja UTU:ssa keskiluvut jäävät alle 0.6:n. UEF:ssa täydellisyyden laatuarvo on suurimmassa osassa tietueista 0.5 tai sen alle, UTU:ssa puolestaan luokkaa 0.5–0.6. UEF:n keskilukujen perusteella laatuarvojen jakauma painottuu jonkin verran keskimääräistä korkeampiin arvoihin, mutta UTU:ssa samansuuntainen jakauman painottuminen on hyvin selvää. Keskiarvo on UTU:ssa mediaania korkeampi peräti 0.06 yksikköä. Tällä perusteella keskimääräistä korkeampia laatuarvoja on suhteellisen paljon, mutta yli 0.8 laatuarvon saaneita tietueita on kuitenkin kokonaisuuteen nähden hyvin vähän. Muuten keskiarvokin olisi korkeampi. Näissä arkistoissa kenttiä on käytetty keskimäärin 8–10 per tietue.

Keskilukujen lisäksi metadatan täydellisyyden laatuarvojen jakaumaa voidaan tarkastella arvojen hajonnan suhteen. Tunnusluvuista arvojen hajonnasta kertovat alaneljännes, yläneljännes ja kvartiiliväli. Laatuarvoista puolet sijoittuvat ala- ja yläneljänneksen väliin. Myös pienin ja suurin arvo (taulukossa 2 ilmoitettu sarakkeissa *Min* ja *Max*) sekä näiden erotus eli vaihteluväli kertovat osaltaan hajonnasta eli millaisella skaalalla arvot vaihtelevat.

Hajontalukujen perusteella laatuarvojen hajonta on suurinta Heldassa ja Tampubissa. Niissä sekä arvojen vaihteluväli että kvartiiliväli ovat suuria. Tämä näkyy hyvin myös ositetuissa pylväissä: Heldassa ovat kaikki laatuarvojen luokat matalimmasta korkeimpaan selvästi näkyvissä, Tampubissa samoin lukuunottamatta korkeinta (yli 0.9:n) luokkaa. Metadatan laatu vaihtelee näissä arkistoissa kuvailukenttien käytön näkökulmasta hyvin paljon tietueiden välillä. Molemmissa arkistoissa on paljon tietueita, joissa kuvailukenttiä voidaan sanoa käytetyn kattavasti, mutta paljon on myös tietueita, joissa kuvailukenttiä on käytetty hyvin vähän. Esimerkiksi Heldassa on tietueita, jotka ovat saaneet vain 0.25:n laatuarvon eli 16 huomioon otetusta kentästä on käytetty vain 4:ää. Tampubin laatuarvojen hajonnassa huomionarvoista on, että matalin arvo ja alaneljännes on sama eli peräti neljäsosa kaikista laatuarvoista on suuruudeltaan vain 0.44. Kuvailukenttien käytön kannalta tämä tarkoittaa Tampubissa, että näin matalan laadun tietueissa on käytössä vain 8 mittauksessa huomioon otetuista 18 kuvailukentästä.

Toisaalta taas Jyxin vaihteluväli on toiseksi suurin (0.53), mutta kvartiiliväli 0.05. Pienimmän arvon (0.35) saaneita tietueita on siis Jyxissä hyvin vähän, koska alaneljännes on niinkin korkea kuin 0.71. Pienin ja suurin arvo ei siis kerro kaikissa tapauksissa arvojen todellista hajontaa. Aaltodocissa, Laudassa, LUTPubissa ja UTU:ssa kvartiiliväli on yli 0.1:n eli niissä arvojen hajontaa on myös havaittavissa. Tämä näkyy myös ositetuissa pylväissä eli laatuarvoja on näissä arkistoissa useimmasta arvojen luokasta. Hajonta ei ole yhtä selkeää Jyxissä, Mpkk:ssa, TutDPubissa, UEF:ssa ja Åbossa (kvartiiliväli niissä on alle 0.1). Keskimääräistä matalammat ja korkemmat arvot ovat näissä melko lähellä toisiaan eli kuten pylväskuviostakin näkee, eri laatuarvojen luokkia on näissä vain muutamia.

Arvojen jakauma on hajonnaltaan hyvin pientä eli kvartiiliväli on 0 kolmessa arkistossa (DHanken, Jultika, Theseus). Näissä arkistoissa suurin osa laatuarvoista on sama kuin mediaani eli laatuarvot ovat keskittyneet arvojakauman puoliväliin. Tämä näkyy myös pylväsvisualisoinnissa: kaikissa kolmessa arkistossa jakaumaa hallitsee selvästi yhden värinen pylväs. Mediaania pienemmät ja suuremmat arvot ovat siis näissä arkistoissa lukumääriltään pieniä. Kun laatuarvojen hajonta on suurimmassa osassa metadatatietueista olematonta, tarkoittaa tämä, että metadatan laatu on näissä arkistoissa tasalaatuista.

Metadatan laatu, vaikka se täydellisyyden näkökulmasta olisikin melko matalaa, voi kuitenkin olla tasalaatuista kautta arkiston. Tällainen tilanne on esimerkiksi DHankenissa, jossa keskiluvut jäävät alle 0.7, mutta laatuarvojen hajonta on hyvin vähäistä. Tasalaatuinen metadata viittaisi siihen, että arkistoissa on noudatettu tiettyä linjaa johdonmukaisesti kuvailukenttien käytössä. Taas hajonnaltaan suurissa arkistoissa (erityisesti Helda ja Tampub), kuvailukenttien käyttö vaihtelee jostain syystä metadatatietueiden välillä hyvin paljon.

5.2 Metadatan laatu painotetun täydellisyyden näkökulmasta

Painotettu täydellisyys -metriikka ottaa laatuarvoa mitattaessa huomioon, kuinka merkittävä käytetty kuvailukenttä on tiedonhaun kannalta. Tällä perusteella kuvailukentät saavat mittauksen yhteydessä painoarvon 0.5 tai 1. Kiinnostavaa on katsoa, miten julkaisuarvojen saamat laatuarvot muuttuvat täydellisyys-mittarilla saatuihin arvoihin. Kuinka paljon tiedonhaun kannalta merkittävien kenttien painottaminen vaikuttaa metadatan laatuun?

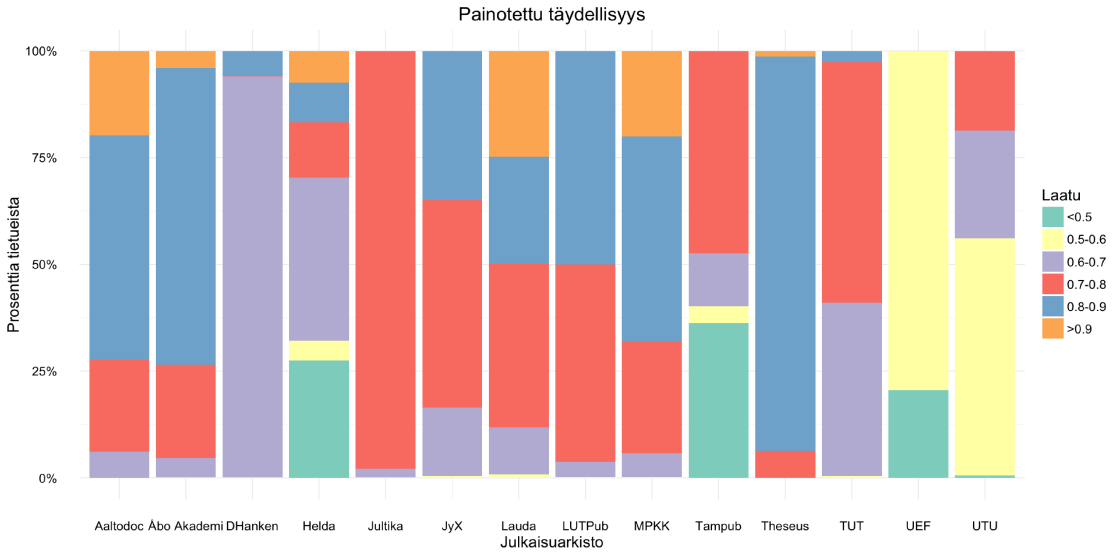
Tunnusluvut osoittavat, että useimpien arkistojen osalta metadatan laatu paranee, kun käytetyille kuvailukentille annetaan painoarvo sen mukaan, kuinka merkittävästä kentästä on tiedonhaun kannalta kyse. Kun tunnusluvut ovat kauttaaltaan korkeampia painotetun täydellisyyden osalta, voidaan sanoa, että arkistossa on käytetty tasaisen kattavasti suuremman painoarvon kenttiä (*Tekijä*, *Nimeke*, *Aihe*, *Kuvailu*). Painottu täydellisyys -mittarilla saatujen laatuarvojen tunnusluvut on esitetty taulukossa 3.

Arkisto	Min	Max	Vaihtelu-väli	Mediaani	Keski-arvo	Ala-neljännes	Ylä-neljännes	Kvartiili-väli
Aaltodoc	0,6	0,95	0,35	0,85	0,86	0,8	0,9	0,1
DHanken	0,55	0,9	0,35	0,7	0,71	0,7	0,7	0
Helda	0,3	0,95	0,65	0,65	0,67	0,5	0,75	0,25
Jultika	0,55	0,77	0,22	0,73	0,74	0,73	0,73	0
Jyx	0,38	0,86	0,48	0,76	0,76	0,71	0,81	0,1
Lauda	0,59	0,91	0,32	0,77	0,81	0,77	0,86	0,09
Lutpub	0,57	0,9	0,33	0,81	0,81	0,76	0,86	0,1
MPKK	0,59	0,91	0,32	0,86	0,83	0,77	0,86	0,09
Tampub	0,45	0,77	0,32	0,64	0,61	0,45	0,73	0,28
Theseus	0,64	0,91	0,27	0,82	0,82	0,82	0,82	0
TutDPub	0,55	0,82	0,27	0,73	0,7	0,64	0,73	0,09
UEF	0,45	0,6	0,15	0,55	0,56	0,55	0,6	0,05
UTU	0,43	0,9	0,47	0,57	0,59	0,52	0,67	0,15
Åbo	0,57	0,95	0,38	0,86	0,82	0,76	0,86	0,1

Taulukko 3: Laatuarvojen jakaumaa kuvaavat tunnusluvut painotettu täydellisyys -mittarilla mitattuna.

Molemmat keskiluvut nousevat 0.03–0.9 yksikköä kuudessa arkistossa (Aaltodoc, Helda, LUTPub, Theseus, TutDPub ja UEF). Koska laatuarvot voivat vaihdella rajoitetun marginaalin sisällä (0–1) ja koska arvojen kokonaismäärä on melko suuri (tietueita arkistoissa 733–92330), voidaan pientäkin muutosta keskiluvuissa pitää merkittävänä. Hieman maltillisempaa keskilukujen nousu on DHankenissa, Jyxissä, Mpkk:ssa, UTU:ssa ja Åbossa.

Täydellisyyden ja painotetun täydellisyyden metriikoilla mitattujen laatuarvojen ero näkyy myös ositetuissa pylväissä (kuva 5).



Kuva 5: Laatuarvojen jakauma julkaisarkistoissa painotetun täydellisyyden näkökulmasta.

Selkein kohennus keskilukujen perusteella laadussa tapahtuu Heldassa: mediaani nousee 0.09, keskiarvo 0.05 yksikköä. Myös ylä- ja alaneljännes kasvavat verrattuna täydellisyyden vastaaviin lukuihin huomattavan paljon. Arvojen jakauma painottuu Heldassa edelleen keskimääräistä korkeampiin arvoihin, mutta keskilukujen välinen ero ei ole niin suuri kuin täydellisyydellä mitattuna. LUTPubissa keskiluvut nousevat myös selvästi (mediaani 0.05 ja keskiarvo 0.04 yksikköä). Myös frekvenssiltään suurin arvojen luokka on LUTPubissa painotetussa täydellisyydessä pykälää korkeampi (0.8–0.9). Arvojen jakauma myös tasaantuu täydellisyyden laatuarvoihin verrattuna eli suurin osa arvoista on lähellä mediaania.

Theuseuksessa molemmat keskiluvut nousevat 0.04 yksikköä. Laadun kohenemisen näkee selvästi pylväsvisualisoinnissa: Theseuksen hallitseva laatuarvojen luokka painotetussa täydellisyydessä on luokkaa korkeampi eli 0.8–0.9, ja jopa yli 0.9 arvoja on jonkin verran. Samoin TutDpubissa lukumäärältään suurin arvojen luokka nousee yhden pykälän luokasta 0.6–0.7 luokkaan 0.7–0.8. Kahden mittaustavan välillä on eroja myös jakauman muodossa. Kun TutDPubissa täydellisyyden laatuarvot keskittyivät mediaanin ympärille, painotetussa täydellisyydessä arvojen ja-

kauma painottuu keskimääräistä matalampiin arvoihin (keskiarvo on pienempi kuin mediaani). UEF:ssa painotettu täydellisyys nousee samoin yhden luokan ylöspäin (alle 0.5 arvoista luokkaan 0.5–0.6). Aaltodocissa arvojakauman muutos ei ilmene kovinkaan selvästi pylväskuviossa, mutta keskilukujen nousun perusteella metadata on laadukkaampaa painotetulla täydellisyydellä mitattuna. Voidaankin sanoa, että suuremman painoarvon kenttiä on näissä arkistoissa käytetty tasaisen kattavasti koko arkiston tasolla.

Mpkk:ssa ja Laudassa keskilukujen kasvu on maltillisempaa. Molemmissa arkistoissa ero laatuarvojen jakaumien eroissa näkyy pylväskuviossa: yli 0.9 laatuarvoja saavia tietueita on kummassakin arkistossa huomattava määrä toisin kuin täydellisyyden mittarilla mitattuna. Jakauman muoto pysyy näissä arkistoissa tosin samana: Mpkk:ssa keskimääräistä matalampia arvoja on paljon, Laudassa taas jakauma on vino oikealle eli se painottuu keskimääräistä korkeampiin laatuarvoihin.

UTU:ssa ja Åbossa vain toinen keskiluvuista nousee painotetulla täydellisyydellä mitattaessa. UTU:ssa mediaani kasvaa, mutta keskiarvo ei muutu, Åbossa tilanne on toisinpäin. UTU:ssa jakauma on edelleen vino oikealle, mutta ei niin selvästi kuin ensimmäisellä mittarilla mitattuna. Åbossa taas painotetun täydellisyyden mittarilla saatujen laatuarvojen jakauma keskittyy selvästi keskimääräistä matalampiin arvoihin. Yli mediaanin olevien arvojen esiintyminen on siis Åbossa vähäistä.

Mielenkiintoisia tapauksia täydellisyyden ja painotetun täydellisyyden eroja analysoitaessa ovat Jultika ja Tampub. Näissä painotetun täydellisyyden laatuarvon keskiarvo on 0.02 ja 0.03 yksikköä pienempi kuin täydellisyyden laatuarvot eli metadata onkin näissä laadultaan heikompaa, kun tiettyjä kenttiä painotetaan. Tämä viittaisi siihen, että merkittäviä kenttiä on käytetty epätasaisesti koko arkiston tasolla. Mediaani on molemmilla mittareilla Tampubissa tosin sama eli jakauman keskikohta pysyy samana.

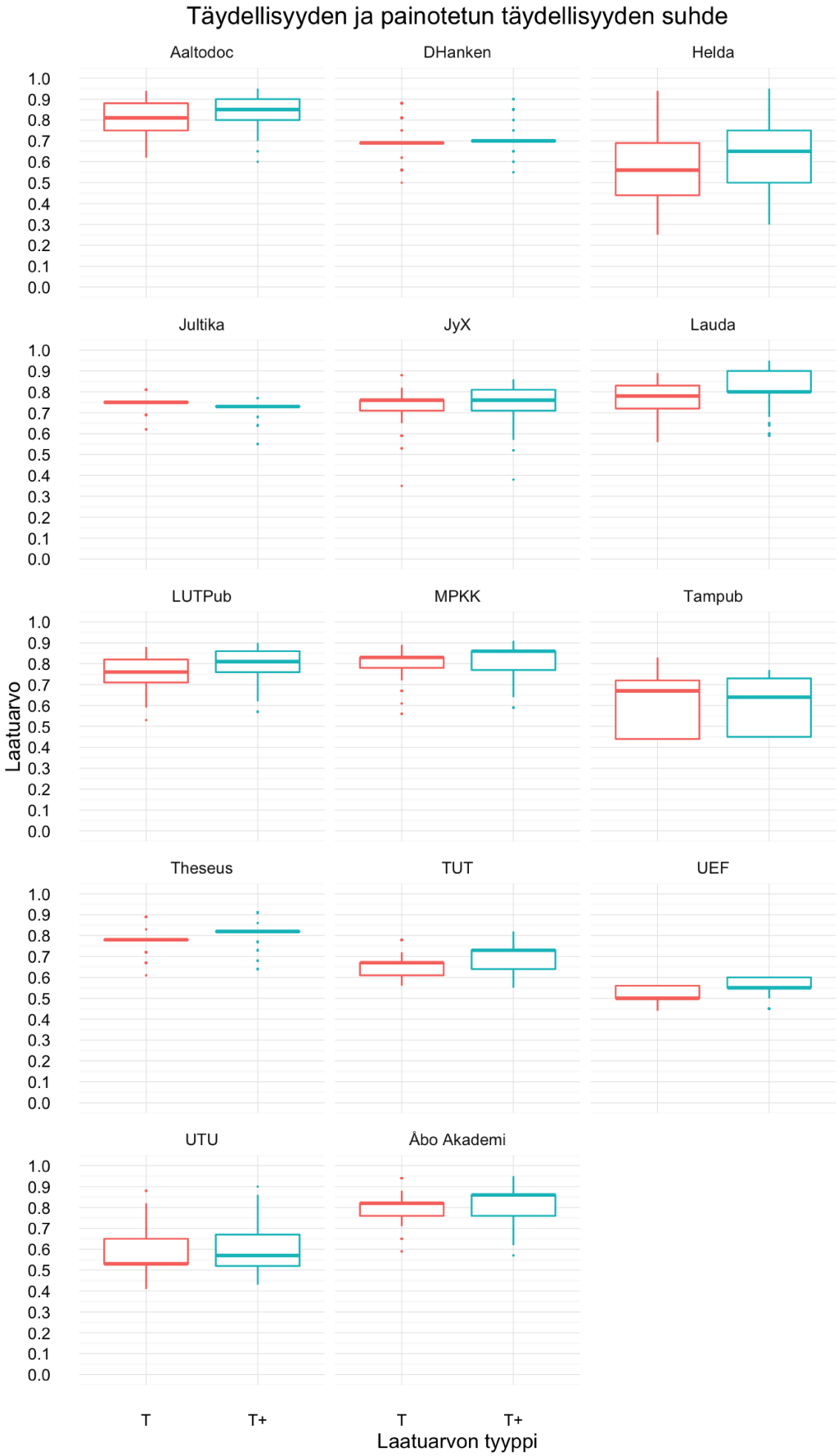
Painotetun täydellisyyden laatuarvojen hajonta kasvaa toisissa arkistoissa, toisissa se pienenee verrattuna täydellisyyden laatuarvoihin. Kvartiiliväli kasvaa viidessä arkistossa (Jyx, Mpkk, TutDPub, UTU, Åbo). Hajonnan kasvuun vaikuttaa näissä arkistoissa se, että yläneljännes on painotetulla täydellisyydellä mitattuna korkeampi kuin täydellisyydellä mitattuna. Hajontaa kasvattaa siis erityisesti keskimääräistä korkeampien laatuarvojen lukumäärän kasvu.

Arkistoissa, joissa hajonta on painotetulla täydellisyydellä pienempää kuin täydellisyydellä mitattuna, hajonnan muutokseen puolestaan vaikuttaa alaneljänneksen kasvu. Näissä arkistoissa (Aaltodoc, Lauda ja UEF) painotetun täydellisyyden laatuarvot kasvavat nimenomaan keskimääräistä pienempien arvojen nousun takia.

Ylä- tai alaneljänneksen kasvun vaikutus näkyy hyvin ositettuja pylväitä vertailemalla (kuvat 4 ja 5). Esimerkiksi Laudassa, jossa kvartiiliväli pienenee eniten, luokan 0.6–0.7 laatuarvojen osuus on selvästi matalampi painotetulla täydellisyydellä mitattuna. Taas muun muassa Mpkk:ssa, jossa kvartiiliväli kasvaa, yli 0.9 laatuarvojen osuus on huomattava painotetussa täydellisyydessä.

DHankenissa, Heldassa, Jultikassa, Tampubissa ja Theseuksessa hajonta ei muutu kahden eri mittarin laatuarvoissa. Keskimääräistä matalampien ja korkeampien laatuarvojen osuudet kasvavat siis yhtä paljon Jultikaa lukuunottamatta. Jultikassa ala- ja yläneljännes laskevat peräti 0.02 yksikköä, mikä kertoo siitä, että painotetulla täydellisyydellä mitattuna metadata on laadultaan heikompaa kuin täydellisyydellä mitattuna. Tämä ero näkyi myös Jultikan keskiluvuissa.

Täydellisyyden ja painotetun täydellisyyden mittareilla saatujen laatuarvojen keskimääräisyyden ja hajonnan eroavaisuuksia havainnollistavat laatikko-jana-kuviot (kuva 6). Laatuarvojen jakaumien mediaani on laatikkojen keskiviiva, laatikoiden alareuna kuvaa jakauman alaneljännestä ja yläreuna yläneljännestä. Kuvasta näkyy hyvin, miten arvojen jakaumat eroavat tai pysyvät ennallaan kahden mittarin välillä eri arkistoissa niin keskilukujen kuin hajonnankin näkökulmasta. Esimerkiksi Heldan kohdalla kuvasta näkee, miten hajonta pysyy samana (eli laatikot ovat samanlevyiset), mutta laatuarvojen taso nousee selvästi painotetussa täydellisyydessä.



Kuva 6: Täydellisyys (T) ja painotettu täydellisyys (T+) -mittareilla saatujen laatuarvojen jakaumat rinnakkain.

Millaista konkreettisesti on täydellisyyden ja painotetun täydellisyyden mittareilla mitattuna huonolaatuinen metadata? Täydellisyyden näkökulmasta kaikista arkistoista pienimmän laatuarvon (0.25) saanut tietue löytyy Helda-arkistosta (kuva 7).

```
<Publication about="oai:helda.helsinki.fi:10138/145492">
  <creator>Tanhua, Marja-Liisa</creator>
  <date>2014-12-19T07:30:42Z</date>
  <date>2014-12-19</date>
  <identifier>http://hdl.handle.net/10138/145492</identifier>
  <description>Ks.</description>
  <description>Kontio, Kirsi</description>
  <source>Opinnäyteviitteet</source>
  <source>Opinnäytteet</source>
  <source>Suomenkielen laitos</source>
</Publication>
```

Kuva 7: Esimerkki matalan laatuarvon saaneesta metadatatietueesta.

Kyseessä on tietue Heldan Opinnäyteviitteet-kokoelmasta, joka on osa Opinnäytteet-pääkokoelmaa. Suurin puute metadatan laadun kannalta on, että tietueessa ei ole opinnäytteen nimeä, mitä voidaan pitää hyvin merkittävänä laadullisena puutteena. *Kuvailu*-kenttää (description) ei ole käytetty tarkoitetulla tavalla. Kokonaisuudessaan metadata ei anna kuvailemastaan kohteesta oikeastaan mitään tietoa. Voidaankin kysyä, mitä tarkoitusta tai käyttöä varten näin niukkoja kuvailutietoja julkaisuarkistossa säilytetään?

Korkein laatuarvo painotetulla täydellisyydellä mitattuna on 0.95. Kuvassa 8 esimerkkinä tietue Aaltodocista. (Tilan säästämiseksi muun muassa description-kentän tekstiä lyhennetty.)

Mittauksessa huomioon otetuista 16 kentästä on tietuessa käytetty 15:tä. Painotetussa täydellisyydessä merkittävien kenttien osuus otetaan huomioon. Tietueessa ovat kaikki neljä korkeamman painokertoimen kenttää. Oikeudet-kenttä on ainoa, joka mittauksessa mukana olleista kuvailukentistä tietueesta puuttuu. Kuten esimerkkikuva näyttää, hyvälaatuinen metadata antaa tietoa opinnäytteestä kattavasti: metadata kertoo sekä opinnäytteen sisällöstä että tekijätiedoista, mutta myös auttavat tallenteen eri versioiden tunnistamisessa.

```
<dcterms:qualifieddc xmlns:dcterms="http://purl.org/dc/terms/"
  <dc:title xml:lang="en">Uniformly quasiregular mappings on elliptic
       riemannian manifolds</dc:title>
  <dc:creator>Kangaslampi, Riikka</dc:creator>
  <dc:contributor type="advisor" xml:lang="en">Mathematics</dc:contri-
       butor>
  <dc:contributor type="department" xml:lang="fi">Matematiikan ja sys-
       teemianalyysin laitos</dc:contributor>
  <dc:contributor xml:lang="fi">Aalto-yliopisto</dc:contributor>
  <dc:contributor xml:lang="en">Aalto University</dc:contributor>
  <dc:subject type="keyword" xml:lang="en">uniformly quasiregular map-
       pings</dc:subject>
  <dc:subject type="keyword" xml:lang="en">riemannian mani-
       folds</dc:subject>
  <dc:subject type="keyword" xml:lang="en">elliptic manifolds</dc:sub-
       ject>
  <dc:subject type="keyword" xml:lang="en">Zalcman's lemma</dc:sub-
       ject>
  <dc:subject type="keyword" xml:lang="en">Julia set</dc:subject>
  <dc:subject type="keyword" xml:lang="en">Lattès mappings</dc:sub-
       ject>
  <dcterms:abstract xml:lang="en">In this thesis we study uniformly
       quasiregular (abbreviated uqr) mappings on compact riemannian
       manifolds..</dcterms:abstract>
  <dcterms:available>2012-08-20T07:36:25Z</dcterms:available>
  <dcterms:issued>2008</dcterms:issued>
  <dc:type xml:lang="fi">G4 Monografiaväitöskirja</dc:type>
  <dc:type type="dcmitype" xml:lang="en">text</dc:type>
  <dc:type type="ontasot" xml:lang="fi">Väitöskirja (monogra-
       fia)</dc:type>
  <dc:type type="ontasot" xml:lang="en">Doctoral dissertation (monog-
       raph)</dc:type>
  <dc:identifier type="isbn" xml:lang=" ">978-951-22-9425-1</dc:iden-
       tifier>
  <dc:identifier type="isbn" xml:lang="#8195;">951-41-0989-9 (prin-
       ted)</dc:identifier>
  <dc:identifier
       type="uri">https://aaltodoc.aalto.fi/handle/123456789/4497</dc
       :identifier>
  <dc:identifier type="urn" xml:lang=" ">URN:ISBN:978-951-22-9425-
       1</dc:identifier>
  <dc:language type="iso" xml:lang="en">en</dc:language>
  <dc:publisher xml:lang="en">Teknillinen korkeakoulu</dc:publisher>
  <kk:permaddress type="urn">http://www.urn.fi/URN:ISBN:978-951-22-
       9425-1</kk:permaddress>
  <kk:file bundle="ORIGINAL"
       href="https://aaltodoc.aalto.fi:443/bitstream/123456789/4497/1
       /isbn9789512294251.pdf" length="947609"
       name="isbn9789512294251.pdf" sequence="1"
       type="application/pdf"/>
</dcterms:qualifieddc>
```

Kuva 8: Esimerkki korkean laatuarvon saaneesta tietueesta.

Selvittämällä, kuinka paljon ja mitä metadataformaatin suosittamia kuvailukenttiä eri arkistoista puuttuu, saadaan selville, minkä kenttien puuttuminen vaikuttaa täydellisyyden ja painotetun täydellisyyden laatuarvoihin eniten. Taulukossa 4 ovat puuttuvien kenttien prosenttiosuudet kussakin arkistossa eli niiden tietueiden osuus, joista kyseinen kuvailukenttä puuttuu suhteessa arkiston kaikkien metadatatietueiden määrään. Taulukossa viiva (-) tarkoittaa, että rajapinta ei ole antanut kuvailukenttää ulos. Tilan säästämiseksi kenttien nimiä on taulukkoon lyhennetty.

Arkisto	Tekijä	Nimeke	Työn taso	Kieli	Hyv.aika	Tekopaikka	Työn laji	Julk.aika	Id.tunnus
Aaltodoc	0	0	0,48	0	0	0,01	32,4	0,01	0
DHanken	0,05	0	3,9	0,05	0	0,05	4	0	94
Helda	0,007	0,09	0,15	25,3	0	3,9	66	-	60,8
Jultika	0	0	0	0	0	-	100	100	0
Jyx	0,006	0	0,05	0,01	0,006	0,01	18,9	0	0
Lauda	0	0	0	0,08	0	0	100	0	53,3
Lutpub	0	0	0,19	0,52	0	0,03	100	2,5	40
MPKK	0,17	0	0	1,79	0	0,06	100	0	0,73
Tampub	0	0	0	36,8	0	0	100	0	48,4
Theseus	0,002	0,001	0	0,009	0	0	100	0	98,7
TUTDPub	0,02	0	0,21	0,44	0	0,07	100	9,6	0
UEF	0	0	-	100	100	-	0	0	0
UTU	0,02	0	50,9	0,44	0	-	100	0	10,9
Åbo	0	0	0,82	1,5	0	6,9	99,3	0	6

Arkisto	Kuvailu	Aihe	Tieteenala	Oikeudet	Julkaisija	Henkilöt	Koko	Tiedostomuot	URL-osoite
Aaltodoc	7,6	0,15	-	99,9	62,3	42,2	-	28,2	0
DHanken	90,1	0	-	4	0,1	-	-	97	0,05
Helda	36,7	16,9	-	71,4	61,5	78,6	100	90,6	0
Jultika	1,13	0,95	100	0	0	81,8	-	0	0
Jyx	61,6	3,5	35,5	10,5	89,6	91,6	25,3	-	0
Lauda	6,7	16,2	100	0	0,4	100	70,7	43,9	0
Lutpub	10,4	4,2	-	99,9	95,3	1	1,2	0,43	0
MPKK	30,5	7,6	0	0	49,6	100	1	43,7	0
Tampub	28,4	0,009	0,02	100	85,2	100	44,2	51,3	0
Theseus	2,89	0,02	0,003	3,4	0,06	100	98,7	0,04	0
TUTDPub	0,83	61,1	100	62,9	100	95,7	3,7	67,6	0
UEF	100	0	100	0	0	100	100	45,5	100
UTU	52,2	96,6	100	99,6	74,1	100	99,8	7,1	0,07
Åbo	12,4	34,2	-	1,4	10,2	61,3	82,4	5,2	2,3

Taulukko 4: Puuttuvien kuvailukenttien osuudet suhteessa arkistojen metadatatietueiden kokonaismäärään.

Taulukon luvut osoittavat, että kenttien puuttuminen on hyvin vaihtelevaa: kenttä, jota on yhdessä arkistossa käytetty lähes kaikissa tietueissa, saattaa puuttua toisesta arkistosta kokonaan. Esimerkiksi *Opinnäytteeseen liittyvät henkilöt* -kenttä (taulukossa nimellä *Henkilöt*) puuttuu kuudesta arkistosta kokonaan, kuudessa arkistossa yli puolesta kaikista tietueista, mutta LUTPubissa vain prosentista tietueista. Samanlainen epäsymmetriä on havaittavissa *Työn laji* -kentän osalta. Kun kahdeksasta arkistosta kenttä puuttuu kokonaan, on kenttä UEF:ssä kaikissa tietueissa ja DHankenistakin se puuttuu vain 4 prosentista tietueista. Nämä kaksi kenttää puuttuvat julkaisu-

arkistoista kaikkein useimmin eli kokonaisuudessa niiden puuttumisella on suurin vaikutus täydellisyys- ja painotettu täydellisyys -mittareilla saatuihin laatuarvoihin. Mutta, kuten jo todettiin, joissain arkistoissa näiden kenttien puuttuminen on vähäistä tai ne esiintyvät jopa kaikissa tietueissa.

Jos kenttien puuttumista tarkastellaan Kansallisen metadataformaatin suositusten perusteella, tilanne näyttää metadatan laadun kannalta hyvältä. Formaatti määrittää pakollisiksi kuvailukentiksi (eli kentät, jotka pitäisi olla kaikissa tietueissa) *Nimekkeen*, *Tekijän*, *Työn tason*, *Kielen*, *Hyväksymisajan* ja *Tekopaikan*. Näistä pakollisista kentistä *Kieli*-kenttä puuttuu useimmin: UEF:ssa sitä ei ole käytetty ollenkaan, Heldasta se puuttuu neljäsosasta ja Tampubissa reilusta kolmanneksesta kaikista tietueista. Muiden pakollisten kenttien puuttuminen on arkistoissa hyvin vähäistä lukuunottamatta UTU:a, josta *Työn taso* -kenttä puuttuu noin puolesta kaikista tietueista. Tätä voidaan pitää vakavana puutteena, koska työn taso (onko kyse esimerkiksi väitöskirjasta vai pro gradusta) on opinnäytetöiden kohdalla oleellinen tieto.

Toisaalta, jos puuttuvia kenttiä katsotaan tiedonhaun kannalta eikä formaatin suositusten näkökulmasta, on tilanne toinen. Tiedonhaun kannalta merkittävistä neljästä kentästä *Aihe*- ja *Kuvailu*-kentät puuttuvat eniten. *Tekijä*- ja *Nimeke*-kenttien puuttuminen on hyvin vähäistä. Huomattavia *Aihe*-kentän puuttumisen osalta ovat TutDPub ja UTU. TutDPubissa kyseinen kenttä puuttuu 61,1 prosentista ja UTU:ssa peräti 96,6 prosentista tietueista. Lisäksi UTU:ssa yli puolesta tietueista puuttuu *Kuvailu*-kenttä. Huomattavaa tämän kentän puuttuminen on myös Jyxissä ja DHankenissa.

Molemmat kentät antavat tietoa opinnäytteen sisällöstä: *Kuvailu*-kenttään tallennetaan työn tiivistelmä, *Aihe*-kenttään kuvailevia asiasanoja. Jos molemmat sisältöä kuvaavat kentät puuttuvat, ei opinnäytteen sisällöstä saa metadatan perusteella oikeastaan mitään muuta tietoa kuin sen, mitä työn nimi antaa. Ollakseen laadukasta metadatan pitäisi kuvata opinnäytettä monipuolisesti eli antaa tietoa niin sisällöstä ja tekijöistä kuin auttaa työn tunnistamisessa ja eri versioiden erottamisessa toisistaan. Toisaalta sisältöä kuvaavien kenttien puuttumisella on myös vaikutusta tiedonhaussa. Kun opinnäytettä haetaan esimerkiksi asiasanojen perusteella, jää suurin osa tallenteista löytymättä, jos asiasanat puuttuvat lähes kaikista tietueista.

Täydellisyyden ja painotetun täydellisyyden laatuarvoja vertailtaessa Jultikan ja Tampubin kohdalla laatuarvojen taso laski jälkimmäisellä mittarilla, vaikka kaikissa muissa arkistoissa taso nousi jopa huomattavasti. Tämän oletettiin kertovan siitä, että tiedonhaun kannalta merkittäviä kenttiä olisi käytetty koko arkiston tasolla epätasaisesti. Puuttuvien kenttien osuuksia tarkastelemalla tämä ei kuitenkaan näyttäisi pitävän paikkaansa. Molemmissa arkistoissa *Tekijä*- ja *Nimeke*-kentät löytyvät kaikista tietueista. Jultikassa myös *Aihe*- ja *Kuvailu*-kenttien puuttuminen on erit-

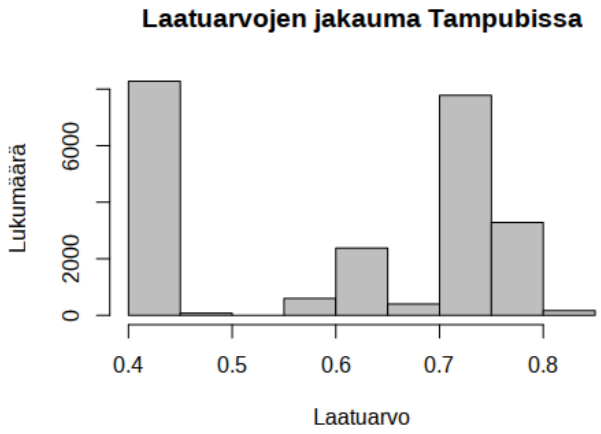
täin vähäistä. Tampubissa *Kuvailu*-kenttä puuttuu noin kolmannesta kaikista tietueista eli sen puuttuminen voisi vaikuttaa painotetun täydellisyyden hieman alempaan laatuarvojen tasoon. Jultikassa puuttuvat kentät eivät anna siis suoraan vastausta kahden mittarin laatuarvojen eroon.

Taulukon luvut osoittavat, että kaikkia formaatin suosittamia kuvailukenttiä on julkaisuarkistoissa käytetty, mutta kenttien käyttö on hyvin vaihtelevaa arkistojen välillä. Taulukon luvuista ei siis voida yksiselitteisesti päätellä, että mikään formaatin määrittämistä kentistä olisi niin sanotusti alikäytetty. Vaikka kenttien *Opinnäytteeseen liittyvät henkilöt* ja *Työn laji* puuttuminen onkin taajaa monessa arkistossa, on niitä toisessa arkistossa käytetty kaikissa tai lähes kaikissa tietueissa.

5.3 Metadatan laatu suhteessa aikaan ja opinnäytteen tasoon

Ovatko täydellisyyden ja painotetun täydellisyyden mittareilla saadut laatuarvot yhteydessä muihin metadatatietueen ominaisuuksiin vai onko kyse sattumasta? Tätä selvitetään tutkimalla, kuinka voimakas yhteys vallitsee kahden mittarin antamien laatuarvojen ja metadatatietueen tallennusvuoden ja toisaalta laatuarvojen ja opinnäytetyön tason välillä. Tallennusvuoden ja laatuarvon välisen yhteyden voimakkuuden selvittäminen voi kertoa osaltaan julkaisuarkistojen toiminnan vakiintumisesta metadatan luomisen osalta. Voidaan olettaa, että kun julkaisuarkisto on otettu käyttöön, ei metadatan suhteen ole ollut vielä vakiintunutta linjaa (esimerkiksi mitä kuvailukenttiä käytetään tai missä muodossa arvot tallennetaan). Tämän olettaisi heijastuvan metadatan laatuun eli laatu paranisi ajan myötä, kun julkaisuarkiston toimintakin on vakiintunut. Opinnäytteen tason ja laatuarvon välisen yhteyden mittaaminen on puolestaan kiinnostavaa, koska monessa julkaisuarkistossa eri tasoisten (esimerkiksi pro gradujen ja väitöskirjojen) opinnäytteiden metadatan luomisessa on erilaisia käytäntöjä. Nämä eroavaisuudet käytännöissä voivat osaltaan olla yhteydessä siihen, kuinka laadukasta metadata kenttien käytön osalta on.

Muuttujien välisiä riippuvuuksia voidaan mitata ja saatuja tuloksia testata erilaisin tilastollisin menetelmin. Siihen, mikä testaustapa valitaan, vaikuttaa tarkasteltavien muuttujien mitta-asteikko. Koska laatuarvo ja metadatatietueen tallennusvuosi ovat suhdeasteikollisia muuttujia, voidaan näiden välistä korrelaatiota mitata laskemalla korrelaatiokerroin, joka ilmaisee kahden muuttujan välisen riippuvuuden voimakkuuden. Laatuarvon ja tallennusvuoden välinen riippuvuus ilmaistaan tässä Spearmanin järjestyskorrelaatiokertoimen avulla. Kyseinen menetelmä on valittu, koska se ei oleta muuttujien arvojen olevan normaalisti jakautuneita. Esimerkiksi Tampubissa (kuva 9) täydellisyyden laatuarvot eivät ole normaalisti jakautuneet (eli jakauma ei noudata niin sanottua Gaussin kellokäyrää), vaan keskimääräistä pienempien arvojen osuus on huomattavan suuri.



Kuva 9: Esimerkki ei-normaalista arvojen jakaumasta.

Spearmanin järjestyskorrelaatiokertoimen arvot vaihtelevat -1:n ja +1:n välillä. Mitä lähempänä arvo on nollaa, sitä pienempää muuttujien välinen riippuvuus on. Korrelaatiokertoimen etumerkki osoittaa muuttujien välisen riippuvuuden suunnan eli pieneekö vai suureneeko toisen muuttujan arvo toisen kasvaessa. (Heikkilä 2008, 203–204.)

Korrelaatiokertoimet osoittavat (taulukko 5), että metadatatietueen luomisvuoden ja täydellisyyden laatuarvon välillä ei suurimmassa osassa arkistoista voida sanoa olevan yhteyttä eli metadatan laatu ei ole riippuvainen tietueen tallennusvuodesta. Se, kuinka kaukana 0:sta korrelaatiokertoimen on oltava, jotta voidaan todeta korrelaation olevan merkittävää, riippuu tutkittavien yksiköiden määrästä. Mitä enemmän tutkittavia yksiköitä on, sitä pienempi korrelaatiokertoimen on oltava, jotta korrelaatiota voidaan pitää merkittävänä. Yleisenä raja-arvona Heikkilä (2008) esittää korrelaatiokerrointa +/-0.3. Jos luku on alle tämän, ei korrelaatiolla voida sanoa olevan käytännön merkitystä.

Raja-arvolla +/-0.3 kuudessa arkistossa täydellisyyden osalta tietueen tallennusvuodella voidaan epäillä olevan vaikutusta laatuarvoon. Neljässä arkistossa (Jyx, Lauda, LUTPub, TutDPub) korrelaatio on positiivinen eli vaikuttaisi siltä, että näissä arkistoissa, mitä uudemmasta metadatatietueesta on kyse, sitä parempaa metadata on laadultaan. Mielenkiintoista on, että kahdessa arkistossa - Aaltodocissa ja UTU:ssa - korrelaatio on negatiivinen eli laatuarvo laskee, mitä uudemmasta aineistosta on kyse. UTU:n kohdalla (korrelaatiokertoimet -0.72 ja -0.746) negatiivinen korrelaatio on jopa melko voimakasta.

Arkisto	Korrelaatiokerroin täydellisyys	Korrelaatiokerroin p.täydellisyys
Aaltodoc	**-0,386**	**-0,389**
DHanken	-0,26	-0,261
Helda	-0,237	-0,203
Jultika	-0,002	-0,001
Jyx	**0,565**	**0,55**
Lauda	**0,65**	**0,547**
Lutpub	**0,484**	**0,432**
MPKK	-0,259	-0,228
Tampub	0,145	0,087
Theseus	0,14	0,143
TutDPub	**0,573**	0,224
UEF	-0,064	0,129
UTU	**-0,72**	**-0,746**
Åbo	0,21	-0,118

Taulukko 5: Spearmanin järjestyskorrelaatiokertoimet täydellisyys- ja painotettu täydellisyys -metriikoilla saatujen laatuarvojen osalta. Lihavoituna arvot, jotka ylittävät rajaarvon +/-0.3.

Kun verrataan täydellisyyden laatuarvoja painotetun täydellisyyden laatuarvoihin ja näiden korrelaatiokertoimia, tulokset osoittavat, että vaihtelua luvuissa ei juuri ole lukuunottamatta yhtä poikkeusta. Mielenkiintoinen ero on tästä näkökulmasta TutDPubissa, jossa korrelaatiokerroin on yli 0.3 yksikköä pienempi painotetussa täydellisyydessä. Metadata on siis painotetun täydellisyyden näkökulmasta tasalaatuisempaa eri ikäisillä tietueilla kuin täydellisyyden osalta. Tästä kertovat myös TutDPubin laatuarvot, jotka painotetun täydellisyyden osalta kertovat laadukkaammasta metadatasta.

Tallennusvuoden ja laatuarvon välinen korrelaatio tai oikeastaan sen puuttuminen näkyy kenties parhaiten kuvan avulla. Kuvassa 9 on esitetty tallennusvuosi suhteessa laatuarvoon ryhmiteltynä opinnäytteen tason mukaan. UEF puuttuu tästä kuvasta, koska rajapinnan kautta ei saatu opinnäytteen tason ilmoittamaa kuvailukenttää. Kuvaan on otettu vuodet 2006–2016, koska tätä varhaisemmalta ajalta metadatatietueita on arkistoissa hyvin vähän. Opinnäyteryhmien jaottelu on selvitetty liitteessä 3.

Kuten korrelaatiokerrointenkin perusteella saattoi todeta, Laudassa sekä ylempien tutkintojen että jatkotutkintojen opinnäytteiden metadatan laatu paranee hieman mitä uudemmasta aineistosta on kyse. Selkeintä laadun paraneminen on vuosina 2012 ja 2013, mutta myös tämän jälkeen laatuarvot ovat hienoisessa kasvussa. Myös Jyxin korrelaatiokertoimet kielivät laadun kasvusta ajan myötä. Laatuarvot ovat kasvaneet erityisesti ylempi korkeakoulututkinto -ryhmässä, jossain

määrin alemman tutkinnon töissä. Vähäisintä laadun muutokset korrelaatiokerrointen mukaan suhteessa aikaan ovat Jultikassa ja Theseuksessa. Tämä näkyy myös kuvassa, jossa laatuarvot pysyvät kaikissa opinnäyteryhmissä samalla tasolla. Erityisesti jälkimmäisessä kahden eri ryhmän opinnäytteiden metadatan laatu pysyy samalla tasolla vuodesta toiseen.

Aaltodocissa negatiiviseen korrelaatiokertoimeen näyttäisi vaikuttavan nimenomaan ylempi korkeakoulututkinto -ryhmään kuuluvien töiden laatuarvojen mataloituminen uudempien tietueiden osalta. TutDPubissa puolestaan laatuarvot ovat hienoisesti nousseet, mutta aivan viime vuosina kääntyneet laskuun. Heldan laatuarvojen hajonta näkyy myös suhteessa aikaan ja opinnäytteen tasoon: selkeää suuntaa laatuarvoissa ei ole havaittavissa missään opinnäytetyöryhmässä, vaan arvot heittelevät eri vuosina paikoin hyvin paljon. Tampubissa vuoden 2013 aikana laatu on molemmissa opinnäyteryhmissä huonontunut selvästi, mutta noussut taas seuraavana vuonna. UTU:n korrelaatiokertoimet olivat melko lähellä -1:tä. Kuvan perusteella laatu on laskenut ajan myötä nimenomaan Jatkotutkinto-ryhmän metadatatietueissa erityisesti vuosien 2008 ja 2009 välillä.

Mitenkään vahvana laatuarvon ja tallennusvuoden välistä yhteyttä ei ainakaan visualisoinnin perusteella voida pitää. Laatuarvo ei siis useimmissa arkistoissa ole riippuvainen metadatatietueen tallennusvuodesta. Joissakin arkistoissa korrelaatiokerroin ylittää raja-arvon +/-0.3, mutta olisi liian uskaliasta sanoa, että muuttujien välinen korrelaatio olisi näissäkään kovin vahvaa tai selvää ainakaan kaikkien opinnäyteryhmien osalta. Mielenkiintoisinta informaatioita laatuarvojen ja tallenusvuoden suhteesta antaakin laatikko-jana-kuviot, jotka näyttävät, miten julkaisuarkistoissa on saattanut tapahtua metadatan laadussa hyppäyksiä suuntaan tai toiseen yhdenkin vuoden aikana. Täytyy myös ottaa huomioon, että jokin muu muuttuja (tai useampikin muuttuja) voi olla yhteydessä metadatan laatuun. Korrelaatiokertoimia laskettaessa mitattiin vain tallennusvuoden ja laatuarvojen välisen suhteen voimakkuutta. Kuvan 10 perusteella näyttäisi esimerkiksi siltä, että opinnäytteen tasolla voisi olla vahvempi vaikutus laatuun kuin ajallisilla tekijöillä.

Kuva 10: Täydellisyyden laatuarvojen vaihtelut suhteessa aikaa opinnäytteen tasoon mukaan ryhmiteltynä.

Opinnäytteen tason ja laatuarvon välistä yhteyttä ei voida mitata korrelaatiokertoimen avulla, koska opinnäytteen taso on mitta-asteikoltaan järjestysasteikollinen. Kyseeseen tulevat tässä tapauksessa ryhmien välisiä keskilukuja vertaavat testit.

Ennen testin suorittamista opinnäytteen tasot on yhtenäistettävä mielekkäisiin ryhmiin. Opinnäytetöille luonnollinen ryhmäjako noudattaa korkeakoulussa suoritettavien tutkintojen tasoja. Opinnäytteet on jaettu tässä kolmeen ryhmään: alempi korkeakoulututkinto/muut tutkielmat, ylempi korkeakoulututkinto, jatkotutkinto (liitteessä 3 ovat julkaisuarkistoissa käytetyt opinnäytteiden tasot ja niitä vastaavat yhtenäistetyt tasot). Kaikissa arkistoissa ei ollut tallennettuna Alempi korkeakoulututkinto -ryhmään kuuluvia opinnäytteitä, ja Theseuksessa puolestaan ei ollut Jatkotutkinto-luokan töitä. UEF:n osalta tätä analyysia ei voitu tehdä, koska rajapinnan kautta ei saatu opinnäytteen tason ilmoittamaa kuvailukenttää.

Opinnäytetyön tason ja laatuarvon välistä suhdetta havainnollistavat ryhmäkeskiarvot. Mikäli eri ryhmien välillä keskiarvot eroavat toisistaan riittävästi, voidaan olettaa, että se, minkä tasoisesta opinnäytteestä on kyse, vaikuttaa metadatan laatuun. Testin nollahypoteesi on, että ryhmien keskiarvot eivät eroa toisistaan, ja näin ollen opinnäytteen tasolla ei ole vaikutusta metadatan laatuun. Taulukossa 6 ovat eri opinnäytetyöryhmien ryhmäkeskiarvot täydellisyyden ja painotetun täydellisyyden osalta sekä kunkin ryhmän koko (eli metadatatietueiden määrä kyseisessä ryhmässä). Ryhmäkeskiarvoja ja arvojen jakaumaa tarkastelemalla vaikuttaisi siltä, että eri ryhmien välillä keskiarvot poikkeavat toisistaan useammassa arkistossa. Yleinen suuntaus keskiarvoissa on, että mitä "korkeamman" tason työstä on kyse, sitä parempaa metadata on laadultaan molemmilla mittareilla mitattuna. Näin ei kuitenkaan ole kaikissa arkistoissa: Jyxissä, Laudassa ja Åbossa Ylempi korkeakoulututkinto -luokan keskiarvo on korkeampi kuin Jatkotutkinto-luokan. Mpkk:ssa, LUTPubissa ja UTU:ssa puolestaan Alempi ja Ylempi tutkinto -luokat saavat korkeamman keskiarvon kuin Jatkotutkinto-luokan työt. Ryhmien keskiarvojen väliset suhteet myös muuttuvat mittareiden välillä muutamissa tapauksissa.

Ryhmien välisissä keskiarvoissa on aina jonkin verran eroavaisuuksia. Keskiarvojen eroja täytyy testata tilastollisen menetelmän avulla, jotta voidaan päättää, ovatko erot tilastollisesti merkitseviä vai johtuvatko ne sattumasta. Koska vertailtavia keskiarvoja on enemmän kuin 2 (useassa arkistossa on kolmen eri ryhmän opinnäytetöitä), käytetään niin sanottuja *k*-otoksen vertailuja. Tässä tapauksessa on valittu testaustavaksi parametriton vaihtoehto muun muassa siksi, että ryhmien varianssit eivät ole yhtä suuria eivätkä laatuarvot kaikissa ryhmissä ole normaalisti jakautuneet. Kyseeseen tulee Kruskall-Wallisin testi, joka kertoo onko ryhmien keskiarvojen tai mediaanien välillä tilastollisesti merkittävää eroa vai johtuvatko erot ryhmien välillä sattumasta. Testi

soveltuu useamman kuin kahden ryhmän vertailuun ja ryhmille, jotka eivät ole keskenään yhtä suuria. (Metsämuuronen 2009, 1051–1052.)

Kruskall-Wallisin testin tulosten mukaan suurimmassa osassa julkaisuarkistoista ryhmät eroavat toisistaan tilastollisesti merkitsevästi (eli p-arvo < 0.05). Tämä viittaisi siihen, että opinnäytteen tasolla on vaikutusta metadatan laatuun. Niiden arkistojen osalta, joissa vertailtavia ryhmiä on 3, testi ei kerro, mitkä ryhmät poikkeavat toisistaan tilastollisesti merkitsevästi. Tämä selviää vasta niin sanotussa post hoc -testauksessa. Testausmenetelmäksi tilanteeseen, jossa ryhmien koot eroavat toisistaan, sopii Dunnin testi. (Metsämuuronen 2009, 1056–1057.)

Testin mukaan suurimmassa osassa arkistoista kaikkien ryhmien välinen keskiarvojen ero on tilastollisesti merkitsevä 5 %:n merkitsevyystasolla. Mitä pienempi p-arvo on, sitä enemmän todisteet puhuvat nollahypoteesia vastaan. Koska lähes kaikkien ryhmien osalta p-arvo oli huomattavasti merkitsevyystasoa 0.05 pienempi, vaikuttaisi siltä, että ryhmien keskiarvojen erot eivät johdu sattumasta. Tämän perusteella voidaankin sanoa, että sillä, minkä tasoinen opinnäyte on kyseessä, on vaikutusta metadatan laatuun.

Opinnäyteryhmät, joissa ero *ei* ollut tilastollisesti merkitsevä täydellisyys- ja painotettu täydellisyys -mittarilla mitattujen laatuarvojen osalta, on korostettu taulukkoon 6 keltaisella värillä. Kuten ryhmäkeskiarvoja vertaamalla näkee, on ero näiden ryhmien välillä hyvin pientä. Huomattava on, että vaikka Theseuksessa kahden ryhmän keskiarvojen välinen ero näyttäisi olevan molemmilla mittareilla mitattuna lähes olematon, testin perusteella ero on kuitenkin tilastollisesti merkitsevä. Tähän vaikuttaa paitsi se, että laatuarvojen vaihteluväli on pieni (0–1), myös havaintoyksiköiden määrä. Koska Theseuksessa testauksessa laskettavia yksiköitä on paljon (92330 laatuarvoa), ovat hyvin pienetkin erot testissä merkitseviä.

Aaltodoc

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,799	0,835	6027
Jatko	0,902	0,917	2551

DHanken

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,69	0,7	1817
Jatko	0,842	0,871	122

Helda

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,44	0,5	3891
Ylempi	0,582	0,645	15959
Jatko	0,797	0,809	7154

Jultika

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,74	0,715	391
Ylempi	0,749	0,729	2005
Jatko	0,774	0,745	2021

Jyx

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,714	0,714	1520
Ylempi	0,747	0,76	13493
Jatko	0,718	0,737	2007

Lauda

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,782	0,811	2178
Jatko	0,739	0,762	180

Lutpub

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,782	0,798	2291
Ylempi	0,758	0,804	6941
Jatko	0,845	0,872	609

MPKK

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,815	0,829	1019
Ylempi	0,814	0,834	603
Jatko	0,793	0,806	161

Tampub

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,599	0,604	20408
Jatko	0,708	0,678	2557

Theseus

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	0,778	0,817	85520
Ylempi	0,778	0,818	6810
Jatko			

TutDPub

	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,662	0,708	8399
Jatko	0,721	0,752	664

UTU

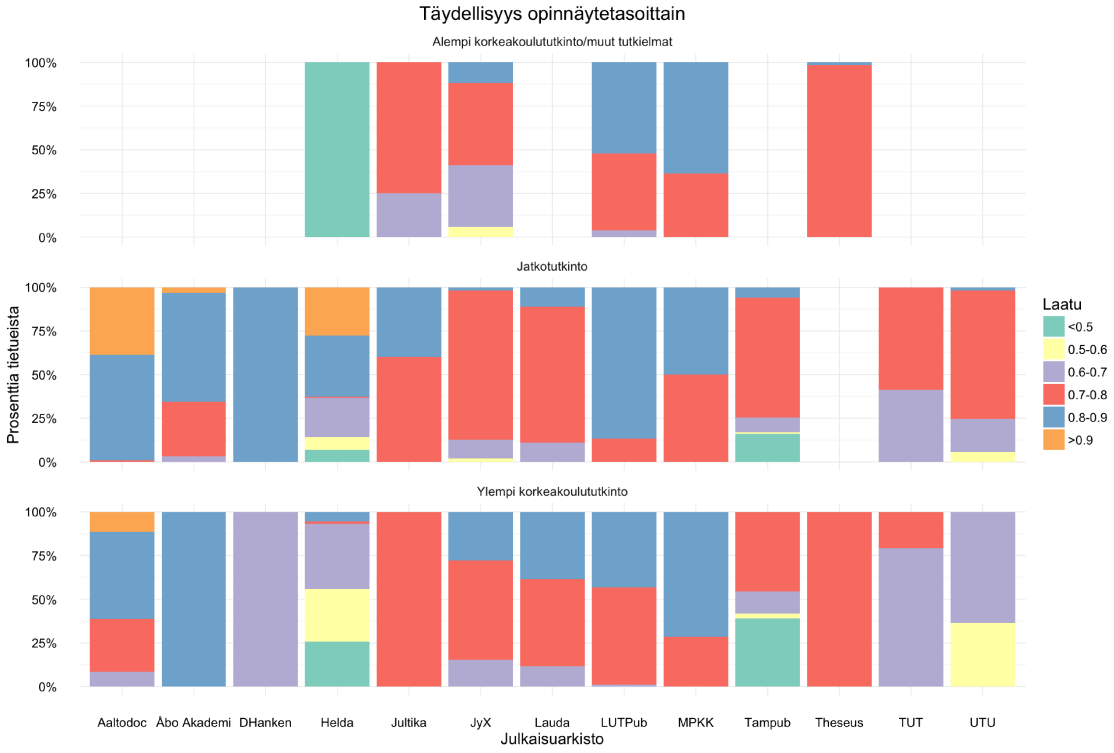
	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi	*	*	11
Ylempi	0,618	0,639	969
Jatko	0,697	0,689	1042

Åbo

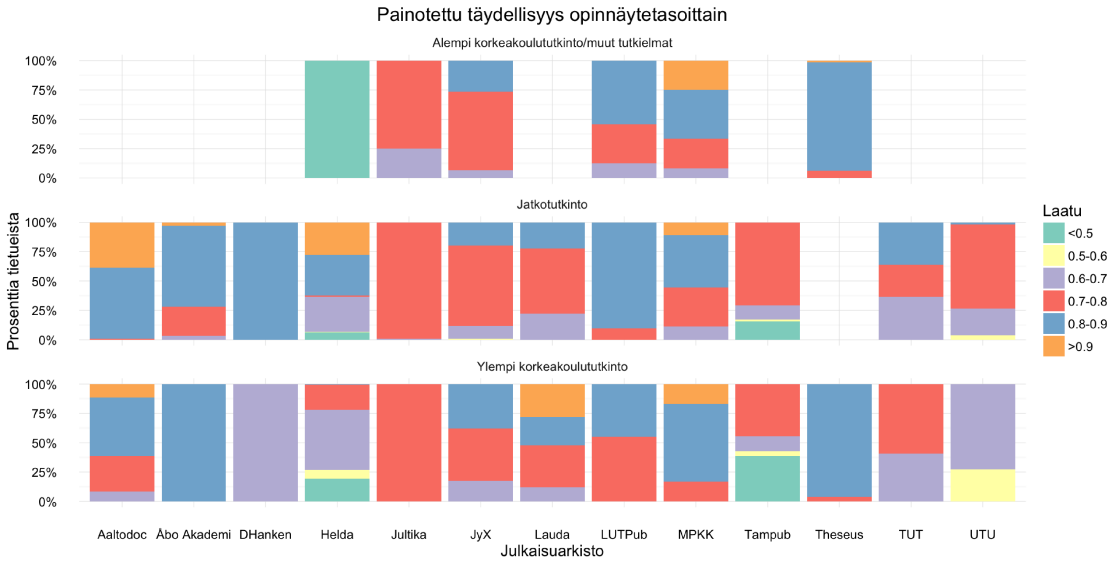
	Keskiarvo täydellisyys	Keskiarvo p.täydellisyys	Yksiköitä
Alempi			
Ylempi	0,818	0,848	89
Jatko	0,806	0,822	636

*Taulukko 6: Yhtenäistettyjen opinnäytetyöryhmien ryhmäkeskiarvot ja havaintoyksiköiden määrät. Merkintä * tarkoittaa, että opinnäyteryhmä jätetty testistä pois, koska yksiköiden määrä on niin pieni muihin ryhmiin verrattuna.*

Kun täydellisyyden- ja painotetun täydellisyyden -mittareilla saatuja laatuarvoja katsotaan pylväsvisualisoinneista (kuvat 11 ja 12), nähdään, missä arvoluokissa metadatan laatu eri ryhmissä on ja miten arvojen jakaumat muuttuvat eri mittaustapojen välillä.



Kuva 11: Täydellisyys opinnäytetasoittain.



Kuva 12: Painotettu täydellisyys opinnäytetasoittain.

Kuvien perusteella opinnäytteen taso näyttäisi selittävän, miksi Jultikassa laatuarvot olivat painotetun täydellisyyden osalta matalampia kuin täydellisyydellä mitattuna. Jatkotutkinto-ryhmään kuuluvien töiden metadatan täydellisyyden laatuarvoissa on huomattava osuus luokkaa 0.8–0.9. Painotetussa täydellisyydessä Jatkotutkinto-ryhmässä kaikki tietueet ovat laatuarvoltaan luokkaa matalampaa eli 0.7–0.8. Koska korkeamman painoarvon kenttien (*Tekijä*, *Nimeke*, *Aihe*, *Kuvailu*) puuttuminen todettiin Jultikassa hyvin vähäiseksi (luku 5.2), täytyy eron syntyä jonkin muun kentän puuttumisesta. Koska tämän työn tarkoituksena ei ole selvittää tarkempia syitä siihen, miksi metadata on laadultaan sellaista kuin se on, jää tämä seikka Jultikan osalta tässä työssä arvoitukseksi.

Korkeamman asteen opinnäytteiden metadata on monessa arkistossa parhaimman laatuista. Selkein ero eri opinnäyteryhmien välillä on kuvasta havaittavissa esimerkiksi DHankenissa, jossa Jatkotutkinnot ovat laadultaan luokkaa 0.8–0.9 ja Ylemmän korkeakoulututkinnon työt 0.6–0.7. Myös Aaltodocissa Jatkotutkinnot-ryhmän metadatan korkeampi laatu näkyy kuvasta: yli 0.9 laatuarvoja on Jatkotutkinto-ryhmässä huomattavasti enemmän kuin Ylempi korkeakoulututkinto -ryhmässä.

6 JOHTOPÄÄTÖKSET

Metadatan laatua analysoivissa tutkimuksissa pyritään mittaamaan manuaalisin tai automaattisin menetelmin, millaista järjestelmän tallenteita kuvaileva data tiettyihin laatuvaatimuksiin verrattuna on. Tässä tutkielmassa selvitettiin, millaista on laadultaan suomalaisten julkaisuarkistojen opinnäytetöiden metadata.

Koska metadatan laatu ei ole yksiselitteinen ilmiö, vaan se on aina sidoksissa muun muassa käyttöympäristöönsä ja aikaan, ei metadatan laatua määritellä yksiselitteisesti, vaan useamman eri ominaisuuden kautta (muun muassa johdonmukaisuus, täydellisyys, tarkkuus). Metadatan laatua on siis tarkasteltava useammasta näkökulmasta, jotta kaikki laatuun liittyvät piirteet tulevat esiin. Tässä tutkielmassa laatua tutkittiin täydellisyyden ja painotetun täydellisyyden näkökulmasta eli huomio kiinnitettiin metadatakenttien käyttöön. Mittareina toimi Ochoan ja Duvallin (2009) kehittämät metriikat metadatan laadun mittaamiseen. Käytetyissä metriikoissa tutkittavan metadatan laatua verrataan tiettyyn metadatastandardiin. Tässä työssä standardiksi valittiin Kansallinen metadataformaatti opinnäytetöille. Vaikka mikään julkaisuarkisto ei ollut noudattanut formaattia täydellisesti, otettiin se peilauspinnaksi, jota vasten metadatan laatua mitattiin. Formaatin suositukset edustivat eräänlaista ihannetapausta laadukkaasta metadatasta.

Laatua tutkittiin automaattisin menetelmin, jolloin analyysin kohteeksi voitiin ottaa kaikki 14 julkaisuarkistossa datan lataushetkellä olleet opinnäytteiden metadatatietueet. Näin ollen voitiin olettaa, että mittaustulokset antoivat laadusta luotettavan kuvan. Vaikka tutkimusaineistoa ladattaessa kaikkia julkaisuarkistoissa käytettyjä kuvailukenttiä ei saatu mittaukseen mukaan, voidaan tuloksia pitää luotettavina, koska tällaisia kenttiä oli arkistoa kohden vähän ja kenttien puuttuminen otettiin huomioon laatua mitattaessa. Täydellisyys ja painotettu täydellisyys -mittarit muutettiin XQuery-kyselyiksi, joilla jokaiselle metadatatietueelle saatiin mitattua molemmilla mittareilla laatuarvo. Laatuarvo vaihteli 0:n ja 1:n välillä. Tutkielman analyysimenetelmä oli kuvaileva tilastoanalyysi. Tavoitteena oli kuvata metadatan laatua tietyllä hetkellä, ei etsiä syitä siihen, miksi metadata on laadultaan sellaista kuin se on. Vaikka kyseessä oli kokonaistutkimus, testattiin lisäksi, onko metadatan laatu yhteydessä tietueen tallennusvuoteen tai siihen, minkä tasoisesta opinnäytteestä on kyse.

Tutkimus vahvisti aiemmissa metadatan laatua mittaavissa tutkimuksissa saatuja tuloksia. Tässäkin analyysissa metadatan laadussa havaittiin paikoitellen suuria puutteita, kun metadatan laatua mitattiin täydellisyyden ja painotetun täydellisyyden näkökulmasta. Kuvailukenttien käyttö vaihteli sekä arkistojen sisällä että arkistojen välillä paikoin hyvinkin paljon. Samansuuntaisia tulok-

sia on saatu myös esimerkiksi Parkin ja Richardin (2011), Kurtzin (2010) sekä Buin ja Parkin (2006) tutkimuksissa.

Yksittäiset laatuarvot vaihtelivat kaikki julkaisuarkistot huomioiden täydellisyys-mittarilla 0.25–0.94. Kenttien käytön näkökulmasta tämä tarkoitti, että arkistoissa oli metadatatietueita, joissa oli käytetty vain 4 kuvailukenttää ja toisaalta tietueita, joissa kenttiä oli 16, kun kenttien kokonaismäärä oli korkeintaan 18. Täydellisyys-mittarilla mitattuna kolmessa arkistossa metadata oli keskimäärin luokkaa 0.8–0.9, mitä voidaan pitää hyvänä tuloksena. Kahdessa arkistossa metadata oli laatuarvoiltaan luokkaa 0.5–0.6, mikä oli matalin tulos. Missään arkistossa ei saavutettu korkeinta mahdollista eli laatuarvoa 1.

Kun metadatan laadun mittauksessa otettiin huomioon kenttien tärkeys tiedonhaun kannalta, nousivat laatuarvot lähes kaikissa arkistoissa. Joissakin arkistoissa laadun parannusta voidaan pitää merkittävänä eli laatuarvojen taso kaikkiaan nousi luokkaa korkeammaksi, toisissa parannus oli maltillisempaa. Korkeinta laatuarvoa ei kuitenkaan tälläkään mittarilla mitattaessa mikään arkisto saavuttanut. Arvot vaihtelivat painotetun täydellisyyden mittarilla mitattuna kaikki arkistot huomioiden 0.3–0.95. Tästä voidaan päätellä, että monessa arkistossa tiedonhaun kannalta merkittäviä kenttiä (*Nimeke*, *Tekijä*, *Kuvailu*, *Aihe*) on käytetty tasaisesti kautta koko arkiston. Toisaalta kahdessa arkistossa laatuarvojen taso laski painotetulla täydellisyydellä mitattuna. Ainoana syynä tason laskuun ei kuitenkaan voitu pitää merkittävien kenttien puuttumista. Tämän laatuarvojen tason heikentyminen syiden etsiminen jätettiin tulevien tutkimusten tehtäväksi.

Keskiluvut antavat laatuarvojen tasosta helposti omaksuttavaa tietoa. Keskiarvon ja mediaanin perusteella laatuarvojen jakaumasta pääsee jo hyvin perille. Sellaisen arkiston metadatan, jossa laatuarvot ovat keskimäärin luokkaa 0.8–0.9, voi varauksetta sanoa kuvailutietojen täydellisyyden osalta olevan laadukkaampaa kuin arkiston, jossa täydellisyyden laatuarvot ovat keskimäärin luokkaa 0.5–0.6. Mittarit mittasivat sitä, mitä pitikin eli ne kykenivät erottamaan huonolaatuista ja hyvälaatuista metadataa sisältävät tietueet toisistaan, kuten esimerkkitietueet osoittivat.

Keskilukujen lisäksi arvojen jakaumaa on kuitenkin hyvä tarkastella myös siitä näkökulmasta, millä välillä arvojen jakaumat vaihtelevat. Laatuarvojen jakaumaa kuvattiin myös hajonnasta kertovilla tunnusluvuilla. Arkistotasolla laatuarvojen hajonnan erot tulivat hyvin esiin. Esimerkiksi Heldassa, jossa hajonta oli tunnuslukujen perusteella kaikista suurinta, olivat kaikki laatuarvojen luokat matalimmasta korkeimpaan edustettuina. Aivan päinvastainen oli tilanne muun muassa Theseuksessa: laatuarvojen kuudesta luokasta esiin nousi vain kaksi, joista toiseen kuuluivat yli 90 prosenttia kaikista laatuarvoista. Laatuarvojen hajonta oli siis Theseuksessa erittäin pientä. Eri mittareiden välillä hajonta muuttui joidenkin arkistojen kohdalla: toisissa hajonta kasvoi, toisissa pieneni.

Hajonnan tarkastelu osoitti myös, että vaikka julkaisuarkiston metadata on keskimäärin tasoltaan suhteellisen matalaa (esimerkiksi luokkaa 0.5–0.6), voi laatuarvojen jakauma olla hyvin tasainen eli hajontaa arvojen välillä on hyvin vähän. Laadun tasaisuus (eli pieni hajonta) kautta koko arkiston kertoo osaltaan siitä, että julkaisuarkistoissa on noudatettu tiettyä linjaa metadataa luotaessa ja tallennettaessa. Tämä linja voi tarkoittaa yhdessä sovittuja käytäntöjä esimerkiksi käytettävien kuvailukenttien suhteen tai tarkkaa laadunvalvontaa ennen metadatan julkaisua tai sitä, että metadatan tallennusvaiheessa jotkin kuvailukentät on pakko täyttää ennen kuin työn saa julkaisuarkistoon tallennettua. Koska julkaisuarkistoille tehty kysely oli niin pienimuotoinen ja yleisellä tasolla pysyttelevä, ei sen vastausten perusteella voida päätellä, onko julkaisuarkistojen käytännöillä metadatan luomisen ja tallentamisen osalta vaikutusta metadatan laatuun.

Kun tutkittiin mittauksessa saatujen laatuarvojen suhdetta metadatatietueen tallennusvuoteen ja toisaalta opinnäytteen tasoon, olivat tulokset niidenkin osalta hajaantuneet eri arkistojen välillä. Sen, milloin metadatatietue on tallennettu, voidaan epäillä olevan yhteydessä metadatan laatuun ainakin muutamissa arkistoissa. Joissakin tapauksissa riippuvuus on positiivista, mutta ehkä hieman yllättäen myös negatiivista. Tämä rikkoi ainakin työn tekijän ennakko-oletuksia, siitä, että metadatan laatu paranisi ajan myötä, kun julkaisuarkistojen käytännöt metadatan luomisessa vakiintuvat. Näin ei siis ainakaan kaikissa tapauksissa ole. Kiinnostava ilmiö, joka laatuarvojen ja tallennusvuoden korrelaatioita tutkittaessa tuli ilmi, oli se, että metadatan laatu saattoi laskea yhtenä vuonna, mutta nousta taas seuraavana. Tämän osoitti hyvin visualisointi (kuva 10), jossa opinnäytteen tason mukaan ryhmiteltyjä laatuarvoja tarkasteltiin suhteessa aikaan.

Opinnäytteen tason ja metadatan laadun välistä yhteyttä voidaan kuvailla melko vahvaksi. Opinnäyteryhmien keskilukujen erojen tilastollista merkitsevyyttä mittaamalla saatiin tulokseksi, että eri arkistoissa lähes kaikki opinnäyteryhmät eroavat tilastollisesti merkittävästi toisistaan eli eroja ei voida pitää sattuman aiheuttamina. Kuvailukenttien käytössä on siis eroja eri tasoisten opinnäytteiden välillä, mikä vaikuttaa metadatan laatuun. Vain kolmessa arkistossa tiettyjen opinnäyteryhmien ero ei ollut tilastollisesti merkitsevä. Erot eri opinnäyteryhmien välillä osoittivat useamman arkiston kohdalla, että mitä korkeamman tason opinnäytteestä oli kyse, sitä laadukkaampaa metadata oli eli esimerkiksi väitöskirjojen metadatassa kuvailukenttiä on käytetty yleisesti ottaen enemmän kuin pro gradu -tutkielmien. Tähän voi olla syynä esimerkiksi se, että väitöskirjoista on tallennettu metadatatietueeseen tietoa, jota alemman tason töistä ei ole tallennettu.

Kuten kirjallisuuskatsauksessa todettiin, metadatastandardilla voi olla vaikutusta metadatan laatuun. Koska suomalaiset julkaisuarkistot eivät ainakaan laajassa mitassa käytä yhteistä metadatastandardia, näkyy tämä useissa tapauksissa hyvin suurina metadatan laadun vaihteluina paitsi ar-

kistojen sisällä myös arkistojen välillä. Kansallinen metadataformaatti opinnäytetöille voisi olla hyvä suositus niin metadatakenttien käytölle kuin kenttiin tallennettaville arvoillekin. Yleinen suositus on, että metadataa luotaessa noudatettaisiin yhteistä standardia, jolloin laadun vaihtelut olisivat todennäköisesti vähäisempiä ja erot arkistojen sisällä ja arkistojen välillä pienempiä. Jos julkaisuarkistoissa käytettyjen kuvailukenttien ja Kansallisen metadataformaatin määrittämien kuvailukenttien välille ei olisi tässä työssä tehty ennen laadun mittaamista vastaavuuksia jokaisen arkiston kohdalla erikseen, olisivat laatuarvot olleet huomattavasti matalammat. Arkistoissa oli käytetty melko paljon niin sanottuja omia kenttiä, joita ei muissa arkistoissa ole käytetty saati standardissa määritelty. Nämä omat kentät voivat aiheuttaa ongelmia ainakin siinä tapauksessa, kun metadataa haravoidaan yhteisarkistoihin.

Vaikka tietyt formaatin suosittamat kuvailukentät puuttuivat useammasta arkistosta, ei mitään kenttää voi määritellä yksiselitteisesti alikäytetyksi. Voidaan kuitenkin kysyä, onko sellaiset kuvailukentät, joita on käytetty hyvin vähäisessä määrin, tarpeellista olla mukana metadatastandardissa? Toisaalta kenttien käyttö ja käyttämättömyys vaihteli arkistojen välillä paikoin hyvinkin paljon. Kenttien käyttö ja käyttämättömyys osoittavat myös, miten eri julkaisuarkistot painottavat eri asioita metadataa luodessaan. Esimerkiksi *Oikeudet*-kenttä saattoi puuttua 99,9 prosentista tietueita (kuten Aaltodocissa ja LUTPubissa) tai löytyä kaikista tietueista, kuten Maanpuoluskorkeakoulun julkaisuarkistossa. Maanpuolustuskorkeakoulun voidaan olettaa tuottavan opinnäytteitä, joiden kohdalla julkaisuoikeuksien ilmaiseminen on tärkeää esimerkiksi tutkielman salassapidettävän luonteen vuoksi. Eri organisaatioiden toiminnan painotukset heijastuvat näin siis myös metadatan käyttöön. Tämä onkin hyvä esimerkki siitä, kuinka metadata ja sen laatu on sidoksissa järjestelmän käyttöympäristöön.

Tutkimuskirjallisuudessa esitetyt määritelmät metadatan laadusta pitivät paikkansa tämän tutkielman tulosten perusteella. Keskeistä määrittelyissä on, että metadatan laatu on moniulotteista. Tässä tutkielmassa laatua tutkittiin täydellisyyden ja painotetun täydellisyyden näkökulmasta eli selvitettiin kuvailukenttien käytön kattavuutta. Koska eri laatupiirteet eivät välttämättä ole sidoksissa toisiinsa, olisivat julkaisuarkistot voineet saada aivan erilaisia laatuarvoja, jos laadun mittauksessa olisi keskitytty esimerkiksi arvojen johdonmukaisuuteen eli siihen, kuinka yhdenmukaisia kenttiin tallennetut arvot ovat metadatatietueissa. Johdonmukaisuuden mittaaminen automaattisin menetelmin olisikin yksi mielenkiintoinen jatkotutkimuksen aihe metadatan laadun osalta. Tässä tutkielmassa käytetyt XQuery-kyselyt taipuvat paitsi arvojen johdonmukaisuuden myös useiden muiden metadatan laatuun vaikuttavien ominaisuuksien selvittämiseen. Koska metadatan laatu on aina sidoksissa metadatan käyttöympäristöön ja käyttäjiin, olisi mielenkiintoista tulevaisuudessa myös selvittää esimerkiksi, miten aineistoja haetaan, mihin kuvailukenttiin tiedonhaut arkistoissa kohdistuvat ja millainen vaikutus metadatalla ja sen laadulla on tiedonha-

kuun. Metadatan laadun määrittelyissä korostetaan myös, että yhden järjestelmän metadatan laatu ei kerro välttämättä mitään toisen järjestelmän metadatan laadusta. Tämäkin määrittely osoittautui paikkansa pitäväksi tämän tutkielman tulosten perusteella. Metadatan laadussa on suuria vaihteluita julkaisuarkistojen välillä eikä yhden järjestelmän metadatan laatua voi yleistää koskemaan kaikki muita arkistoja.

Koska tieteen avoimuus ja aineistojen näkyvyys ja vapaa saatavuus ovat tulevaisuudessa mitä todennäköisimmin entistä olennaisempia aspekteja, ei julkaisuarkistojen merkitys tule ainakaan vähenemään. Julkaisuarkisto on keskeinen väylä tarjota avoimesti kaikkien saataville korkeakouluissa tuotettuja julkaisuja ja tutkimuksia. Siksi olisikin tärkeää kiinnittää huomiota myös siihen, miten kattavasti, yhdenmukaisesti ja johdonmukaisesti arkistoon tallennetut aineistot on kuvailtu. Esimerkiksi aineistojen löytämisessä, tunnistamisessa ja hallinnoinnissa metadatalla on keskeinen rooli. Jos metadata on laadultaan heikkoa, asettuu julkaisuarkistojen tarjoama hyöty kyseenalaiseksi. Tämä tutkielma antaa julkaisuarkistojen käyttöön tietoa siitä, millaista arkistojen opinnäytetöiden metadata on laadultaan ja toivottavasti myös kannustaa kiinnittämään entistä enemmän huomiota metadatan laatuun ja ehkä myös parantamaan jo olemassa olevien aineistojen metadataa.

LÄHTEET

Avoin tiede ja tutkimus. (2016). Avoimuuden käsikirja tutkijoille. Noudettu 11. marraskuuta 2016, osoitteesta http://avointiede.fi/www-kasikirja

Barton, J., Currier, S., & Hey, J. M. N. (2003). Building Quality Assurance into Metadata Creation: An Analysis Based on the Learning Objects and e-Prints Communities of Practice. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 39–48*. Noudettu 10. lokakuuta 2016 osoitteesta http://dcpapers.dublincore.org/pubs/article/viewFile/732/728

Besiki Stvilia, Les Gasser, M. B. T., & Smith, L. C. (2007). A Framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733. http://doi.org/10.1002/asi.20652

Bruce, T. R., & Hillmann, D. I. (2004). The Continuum Of Metadata Quality: Defining, Expressing, Exploiting. Teoksessa Hillmann, D. I. & Westbrooks, E. L. (Toim.), *Metadata in Practice* (s. 238–256). Chicago: ALA Editions.

Bui, Y., & Park, J.-R. (2006). An Assessment of Metadata Quality: A case study of the National Science Digital Library Metadata Repository. *Proceedings of CAISACSI 2006, Information Science Revisited: Approaches to Innovation*. Noudettu 10. marraskuuta 2016 osoitteesta http://idea.library.drexel.edu/handle/1860/1600

Chuttur, M. Y. (2014). Investigating the Effect of Definitions and Best Practice Guidelines on Errors in Dublin Core Metadata Records. *Journal of Information Science*, *40*(1), 28–37. http://doi.org/10.1177/0165551513507405

Confederation of Open Access Repositories. (2011). *The Case for Interoperability for Open Access Repositories*. Noudettu 14. marraskuuta 2016 osoitteesta https://www.coar-repositories.org/files/A-Case-for-Interoperability-Final-Version.pdf

Currier, S., Barton, J., O 'Beirne, R., & Ryan, B. (2004). Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process. *Research in Learning Technology*, *12*(12), 1741–1629. http://doi.org/10.1080/0968776042000211494

Dublin Core Metadata Initiative.(2016). *DCMI Specifications*. Noudettu 14. marraskuuta 2016 osoitteesta http://dublincore.org/specifications/

Gavrilis, D., Makri, D.-N., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C., & Constantopoulos, P. (2015). Measuring Quality in Metadata Repositories. Teoksessa S. Kapidakis, C. Mazurek, & M. Werla (Toim.), *Research and Advanced Technology for Digital Libraries* (Vsk. 9316, ss. 56–67). Springer: Berlin. http://doi.org/10.1007/978-3-319-24592-8_5

Godby, C. J., Smith, D., & Childress, E. (2003). Two Paths to Interoperable Metadata. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 19–27.* Noudettu 11. marraskuuta 2016 osoitteesta http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf

Greenberg, J. (2005). Understanding Metadata. *Cataloging & Classification Quarterly*, *40*(3–4), 17–36. http://doi.org/10.1300/J104v40n03

Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Proceedings of the International Conference on Dublin Core and Metadata Applications,* 38–45. Noudettu 13. marraskuuta 2016 osoitteesta http://dcpapers.dublincore.org/pubs/article/view/647/643

Heikkilä, T. (2008). *Tilastollinen tutkimus*. Helsinki: Business Edita.

Hider, P. (2012). *Information Resource Description : Creating and Managing Metadata*. London: Facet Publishing.

Hillmann, D. I. (2008). Metadata Quality: From Evaluation to Augmentation. *Cataloging & Classification Quarterly*, *46*(1), 65–80. http://doi.org/10.1080/01639370802183008

Holopainen, M. (2008). *Tilastolliset menetelmät*. (5. uud. p.). Porvoo: WSOY Oppimateriaalit.

Hughes, B. (2005). Metadata Quality Evaluation: Experience from the Open Language Archives Community. Teoksessa Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, & E. Lim (Toim.), *Digital Libraries: International Collaboration and Cross-Fertilization* (Vsk. 3334, s. 135–148). Springer: Berlin. http://doi.org/10.1007/978-3-540-30544-6_34

Ilva, J., Lager, L., Saijos, J., & Stenvall, J. (2006). Kansallinen metadataformaatti opinnäytetöille. Noudettu 11. marraskuuta 2016 osoitteesta http://www.doria.fi/bitstream/handle/10024/88787/metadata1_0.pdf?sequence=1

Kirkland, L. N. (2013). The Relationship of Metadata to Item Circulation. *Cataloging & Classification Quarterly*, *51*(5), 510–531. http://doi.org/10.1080/01639374.2012.762963

Kurtz, M. (2010). Dublin Core, DSpace, and a Brief Analysis of Three University Repositories. *Information Technology and Libraries*, *29*(March), 40–47. http://doi.org/http://dx.doi.org/10.6017/ital.v29i1.3157

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Libraries and the Academy, 3 (2), 327–336.* https://doi.org/10.1353/pla.2003.0039

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., & Manitsaris, A. (2008). A Conceptual Framework for Metadata Quality Assessment. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 104–113. Noudettu osoitteesta http://portal.acm.org/citation.cfm?id=1503418.1503429

Metsämuuronen, J. (2009). *Tutkimuksen tekemisen perusteet ihmistieteissä : tutkijalaitos* (4. laitos.). Helsinki: International Methelp.

Moulaison, H. L., Rathbun-Grubb, S., Abbas, J., Greenberg, J., La Barre, K., Rodríguez, E. M., Šauper, A. (2012). Emerging Trends in Metadata Research. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. http://doi.org/10.1002/meet.14504901174

Ochoa, X. (2014). Metadata Quality. Teoksessa Sicilia Miguel-Angel (Toim.), *Handbook of Metadata, Semantics and Ontologies* (s. 63–88). Singapore: World Scientific.

Ochoa, X., & Duval, E. (2009). Automatic Evaluation of Metadata Quality in Digital Repositories. *International Journal on Digital Libraries*, *10(2–3),* 67–91. http://doi.org/10.1007/s00799-009-0054-4

Open DOAR - The Directory of Open Access Repositories. (2016). Noudettu 10. marraskuuta 2016 osoitteesta http://www.opendoar.org/

Park, E. G., & Richard, M. (2011). Metadata Assessment in e-Theses And Dissertations Of Canadian Institutional Repositories . *The Electronic Library Iss*, *29*(1), 394–407. http://dx.doi.org/10.1108/02640471111141124

Park, J.-R. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly, 47 (3–4),* 213–228. http://doi.org/10.1080/01639370902737240

Park, J.-R., & Childress, E. (2009). Dublin Core Metadata Semantics: An Analysis of the Perspectives of Information Professionals. *Journal of Information Science*, *35*(6), 727–739. http://doi.org/10.1177/0165551509337871

Reiche, K. J., & Höfig, E. (2013). Implementation of Metadata Quality Metrics and Application on Public Government Data. *2013 Proceedings of the 37th Annual Computer Software and Applications Conference Workshops*, 236–241. http://doi.org/10.1109/COMPSACW.2013.32

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is ”Quality” Metadata ”Shareable” Metadata? The Implications of Local Metadata Practices for Federated Collections. *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries,* s. 223–237. Noudettu 10. lokakuuta 2016 osoitteesta http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/pdf/shreeves05.pdf

Sicilia, M.-A. (2014). Metadata Research: Making Digital Resources Useful Again. Teoksessa Sicilia Miguel-Angel (Toim.), *Handbook of Metadata, Semantics and Ontologies* (s. 1–8). Singapore: World Scientific.

Stenvall, J. (2002). *Dublin Core -formaatin käyttöopas*. Noudettu 9. marraskuuta 2016 osoitteesta https://www.kiwi.fi/display/DublinCore/Tervetuloa?preview=/45780340/46564816/dc_opas.pdf

Tani, A., Candela, L., & Castelli, D. (2013). Dealing With Metadata Quality: The Legacy of Digital Library Efforts. *Information Processing & Management*, *49*(6), 1194–1205. http://doi.org/10.1016/j.ipm.2013.05.003

Van Hooland, S., & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums : How to Clean, Link and Publish Your Metadata*. London: Facet Publishing.

Ward, J. (2004). Unqualified Dublin Core Usage in OAI-PMH Data Providers. *OCLC Systems & Services: International Digital Library Perspectives*, *20*(1), 40–47. http://dx.doi.org/10.1108/10650750410527322

Hei,
teen informaatiotutkimuksen gradua Tampereen yliopistossa. Tutkielmani aihe on suomalaisten julkaisuarkistojen opinnäytetöiden metadatan laatu. Tarvitsisin nyt hieman taustatietoa tutkielmaani varten.

Kysymykseni koskevat opinnäytetöiden kuvailutietojen (metadatan) luomista ja tallennusta.

1. Millainen on opinnäytetöiden metadatan luomis- ja tallennusprosessi? (esimerkiksi kuka metadatan luo: opinnäytetyön tekijä itse vai joku muu? Tapahtuuko metadatan luominen ja tallennus verkkolomakkeen kautta vai jollakin muulla tavalla; miten?)
2. Jos metadatan luo opinnäytetyön tekijä itse, tarkistetaanko metadata ennen kuin opinnäytetyö julkaistaan arkistossa? Millaisiin seikkoihin tarkistuksessa kiinnitetään huomiota?
3. Onko organisaation sisällä ohjeistusta julkaisuarkiston metadatan suhteen? (esimerkiksi mitä kuvailukenttiä käytetään? missä muodossa tallennettavat arvot ovat?)

Voitte vastata sähköpostilla lähettämällä vastauksen tähän viestiin.
Kiitos paljon tiedoista!

Kuvailukenttä/Arkisto	Tekijän nimi	Nimeke	Taso	Kieli	Hyväksymisaika	Tekopaikka	Laji	Julkistamisaika	Identifiointitunnus
Aaltodoc	creator	title	type type=ontasot	language	available	contributor	type @type=dcmitype	issued	identifier type=urn
DHanken*	creator	title	type	language	date	contributor	type contains(text| Text)	-1 identifier (contains URN)	identifier (contains URN)
Helda*	creator	title	type	language	date	contributor	type (contains text | Text)	-1 -1	identifier (contains URN)
Jultika	creator	title	type	language	date	-1	NA	NA	identifier contains (urn)
Jyx	contributor @type=author	title	type @type=ontasot	language	date @type=available	contributor @type=yliopisto or laitos or tiedekunta	type @typedcmitype	date @type=issued	dentifier @type=urn
Lauda	element=contributor qualifier=author	element=title	element=type qualifier=ontas ot	element=language	element=date qualifier=available	element=contributor not(@qualifier)	NA	element=date qualifier=issued	element=identifier qualifier=urn
Lutpub	element=contributor, qualifier=author	element=title	element=type	element=language	element=date qualifier=available	element=contributor not(@qualifier)	NA	element=date qualifier=issued	element=identifier qualifier=urn
MPKK	element=contributor qualifier=author	element=title	element=type qualifier=okmta so	element=language	element=date qualifier=available	element=contributor not(@qualifier)	NA	element=date qualifier=issued	element=identifier qualifier=urn
Tampub	element=contributor qualifier=author	element=title	element=type qualifier=ontas ot	element=language	element=date qualifier=available	element=administrativ eunit OR faculty OR department	NA	element=date qualifier=issued	element=identifier qualifier=urn
Theseus	element=contributor, qualifier=author	element=title	element=type qualifier=ontas ot	element=language	element=date qualifier=available	element=organization	NA	element=date qualifier=issued	element=identifier qualifier=urn
TutDpub	creator	title	type	language	available	contributor @type=yliopisto	NA	issued	identifier type=urn
UEF	creator	title	-1	NA	NA	-1	type	date.created	identifier
UTU	element=contributor, qualifier=author	element=title	element=type qualifier=ontas ot	element=language	element=date qualifier=available	-1	NA	element=date qualifier=issued	element=identifier qualifier=urn
Åbo	element=contributor qualifier=author	element=title	element=type	element=language	element=date qualifier=available	element=contributor not(@qualifier)	element=type qualifier=dcmitype	element=date qualifier=issued	element=identifier qualifier=urn
*Kenttätarkentelta ei saanut rajapinnan kautta						-1			

Kuvailukenttä/ Arkisto	Kuvailu	Asiasanat	Tieteenala	Oikeudet	Julkaisija	Henkilöt	Koko	Tiedostomuoto	URL-osoite
Aaltodoc	abstract	subject	-1	rights	publisher	contributor @type=supervisor or contributor type=advisor	-1	file	identifier type=uri
DHanken*	description	subject	-1	rights	publisher	NA	NA	format	Identifier (contains “http”)
Helda*	description	subject	-1	rights	publisher	Ths or opn	NA	format	Identifier (contains “http”)
Jultika	description	subject	NA	rights	publisher	contributor	-1	format	identifier contains(http)
Jyx	element=description qualifier=abstract	subject @type=ysa or other or kota	contributor @type=oppiaine or tieteenala	rights	publisher	contributor @type=advisor	element=format qualifier=extent	-1	identifier @type=uri
Lauda	element=description qualifier=abstract	element=subject	element=programme	element=rights	element=publisher	NA	element=format qualifier=extent	file	element=identifier qualifier=uri
Lutpub	element=description qualifier=abstract	element=subject	-1	element=rights	element=publisher	element=ths OR rev	element=format qualifier=extent	file	element=identifier qualifier=uri
MPKK	element=description qualifier=abstract	element=subject qualifier= ysa or yso or puho	element=subject qualifier=tieteenala or oppiaine	element=rights	element=publisher	NA	element=format qualifier=extent	file	element=identifier qualifier=uri
Tampub	element=description	element=subject	element=subject qualifier=study or degreeprogramme	NA	element=publisher	element=contributor	element=format qualifier=extent	element=file	element=identifier qualifier=uri
Theseus	element=description qualifier=abstract	element=subject	element=programme	element=rights	element=publisher	NA	element=format qualifier=size or extent	element=file	element=identifier qualifier=uri
TutDpub	abstract	subject	NA	rights	NA	contributor @type=opn or ths	relation @type=isformatof	file	identifier type=uri
UEF	NA	subject	NA	rights	publisher	NA	NA	identifier not(contains(campus_ use))	NA
UTU	element=description qualifier=abstract	element=subject	element=contributor	element=rights	element=publisher	NA	element=format qualifier=extent	element=file	element=identifier qualifier=uri
Åbo	element=description	element=subject	-1	element=rights	element=publisher	element=ths or opn	element=format qualifier=extent	file	element=identifier qualifier=uri

Arkisto / Yhtenäistetty ontaso	Aaltodoc	DHanken	Helda	Julkika	Jyx	Lauda	Lutpub	Mpkk	Tampub	Theseus	Tut	UTU	Åbo
Alempi korkeakoulututkinto/muut tutkielmat			Laud. sivuainetutkielma	bachelor thesis, other	bachelor's thesis		bachelor's thesis, kandityö,	Ammattikorkeakoulututkinto kandidaatintutkinto		amk, amk-opinnäyte, bachelor degree, yh-examen, polytechnic thesis, ope-amk, erikoistumisopinnot		syventävien opintojen työ, sivuaineen tutkielma	
Ylempi korkeakoulututkinto	Diplomityö, gradu	Masters thesis, thesis	Pro gradu, thesis, tutkielma, opinnäyte, muu, master's thesis, tutkielma (eläinlääketieteen lisensiaatti)	master thesis	master thesis, pro gradu	pro gradu	pro gradu, pro gradu -tutkielma, diplomityö, pro gradu thesis, master's thesis, progradu,	Pro gradu diplomityö ylempi amk-opinnäytetyö	pro gradu, syventävä työ	yamk, Högre Yh-examen, master's thesis	Diplomityö	diplomityö, pro gradu, master thesis	Avhandling pro gradu, Diplomarbete, Master´s thesis
Jatkotutkinto	Lisensiaatintyö,Väitöskirja	Doctoral thesis	Lisensiaatintyö,Väitöskirja	doctoral thesis	lisensiaatintyö, väitöskirja	lisensiaatintyö, väitöskirja	lisensiaatintyö, licentiate thesis, väitöskirja	Lisensiaatintyö	lisensiaatintyö, väitöskirja		lisensiaatintyö, väitöskirja	lisensiaatintyö, väitöskirja	Doctoral dissertation, Doctoral dissertation (article based), Doctoral dissertation (monograph)