Tampereen yliopisto

Jussi Pajari

"TUTKIMUSAINEISTOJEN METATIEDOT"

"Metatietojen laatu data- ja metatietoarkistoissa"

Informaatioteknologian ja viestinnän tiedekunta

Pro gradu -tutkielma

[Joulukuu] [2019]

TIIVISTELMÄ

Jussi Pajari: Tutkimusaineistojen metatiedot. Metatietojen laatu data- ja metatietoarkistoissa.
Pro gradu -tutkielma, 79 s.
Tampereen yliopisto
Viestintätieteiden tiedekunta, informaatiotutkimus ja interaktiivinen media
Joulukuu 2019

Tiedepoliittinen avoin tiede ja tutkimus -liike on muuttanut tiedepolitiikkaa ja tieteen käytäntöjä kohti läpinäkyvyyttä ja avoimuutta. Tiedepolitiikan muutos on vaikuttanut niihin tahoihin, jotka tuottavat, rahoittavat tai julkaisevat tiedettä ja tutkimusta. Uudet tieteelliset käytännöt edellyttävät tutkijoilta ja tutkimusryhmiltä yhä avoimempaa tutkimusta. Tutkimusaineistojen dokumentoiminen, avoin jakaminen ja uudelleenkäyttö on saanut paljon huomiota kansainvälisesti ja kansallisesti. Tutkimusaineistojen avoimuus, jakaminen ja uudelleenkäyttö edellyttää hyvin dokumentoitua tutkimusaineistoa. Dokumentoinnilla tarkoitetaan tutkimusaineistojen metatietojen kirjaamista. Suomessa tutkimusaineistojen metatietojen laatua ei ole aikaisemmin kartoitettu.

Tutkimusmenetelmä on tilastollinen tutkimus. Tutkimuksen näyte koostui kolme data- ja metatietoarkistoa, joiden sisältö ladattiin kokonaisuudessaan ja analysoitiin automaattisin menetelmin. Analysoinnissa käytettiin täydellisyyden ja painotetun täydellisyyden metriikkaa sekä tutkimusta varten luotua vertailumittaria, jonka avulla metriikkojen tulokset voitiin operationalisoida. Tuloksia tarkastellaan keskiarvon, keskihajonnan ja mediaanin avulla sekä luokiteltiin laatuarvojen perusteella.

Tutkimuksen tulokset ovat positiiviset. Tutkimusaineistojen metatietojen laatu oli keskimäärin hyvä suuresta hajonnasta huolimatta. Tutkittujen arkistojen keskiarvo laadun suhteen oli molemmilla metriikoilla mitattuna 80 %. Tutkimuksen avulla saatiin selville myös ongelmakohdat metatietojen tuottamisessa. Ongelmallisia metatietokenttiä ovat tutkijan tunniste, organisaatiotiedot ja asiasanat.

Avainsanat: metadata, tutkimusaineisto, laatu, laadunhallinta, tietoarkistot, tietovarannot, metriikka, avoin tieto

SISÄLLYSLUETTELO

1 JOHDANTO

Tutkielmassani tutkitaan data- ja metatietoarkistoihin tallennettujen tutkimusaineistojen metatietojen laatua. Tutkielman aihe juontuu Suomessakin runsaasti näkyvyyttä saaneen tiedepoliittisen avoin tiede ja tutkimus (*Open Science and Research*) - liikkeen edesauttamasta kansainvälisen ja kansallisen tiedepolitiikan muutoksesta. Käytännössä avoin tiede ja tutkimus on sitä, että tutkimuksen teossa tuotetut tutkimusjulkaisut, tutkimusaineistot ja tutkimusmenetelmät ovat kenen tahansa saatavilla maksutta.

Tiedepolitiikan muutos kohti avoimempaa toimintakulttuuria vaikuttaa ensisijaisesti niihin tahoihin, jotka Suomessa joko tuottavat, rahoittavat tai julkaisevat tiedettä ja tutkimusta. Tämä tarkoittaa tutkimuslaitoksia, korkeakouluja ja rahoittajia sekä tiedelehtiä ja tiedekustantamoita. (Opetus- ja kulttuuriministeriö 2016.)

Tiedepolitiikka kytkeytyy korkeakoulupolitiikkaan, minkä johdosta avoin tiede ja tutkimus vaikuttaa Suomen korkeakouluissa toimiviin tieteellisiin korkeakoulukirjastoihin. Tiede- ja korkeakoulupolitiikan muutos on lisännyt yliopistoissa sekä ammattikorkeakouluissa tutkijoiden tarvetta tutkimuspalveluille. (Tenopir ym. 2017; Vilar & Zabukovec 2019.)

Avoimen tieteen ja tutkimuksen palvelut ovat verrattain uusia tieteellisten kirjastojen tarjoamia palveluita, jotka ovat nousemassa merkittäviksi kirjastopalveluiksi kansallisesti sekä kansainvälisesti. Keskeisiä uusia tutkimuspalveluja ovat julkaisu- ja rinnakkaistallennuspalvelut, aineistonhallintaan liittyvät palvelut sekä näkyvyyttä ja vaikuttavuutta edistävät palvelut kuten metriikka. (Corrall, Kennan & Afzal 2013; Cox & Pinfield 2014; Ala-Kyyny 2016; Ogungbeni ym. 2016; Rantasaari & Kanerva 2017; Tenopir 2017.)

Tutkielman kannalta keskeinen ongelma muodostuu siitä, että rahoittajien vaatiessa aineistonhallintaa ja aineistonhallintasuunnitelmia jää tutkimusaineiston kuvailu eli metatietojen tuottaminen tutkimusryhmän tai yksittäisen tutkijan vastuulle. Metatietojen tuottaminen tutkimusaineistoista on haastavaa, vaikka kuvailun tekisi tutkimusaineiston suunnittelija, kerääjä ja tuottaja.

Tutkimusaineiston metatietojen tuottaminen koettiin Ala-Kyynyn, Roinilan ja Korhosen (2017) haastattelututkimuksessa ongelmalliseksi. Tutkimuksen haastatteluista kävi myös ilmi, että metatietotietueet ovat hyvin puutteellisia eikä niiden tuottamiseen, parantamiseen tai laadunvalvontaan ole. Tutkimuksissa metatietojen tuottaminen koettiin myös hankalaksi (ks. Mayernik 2011; Tenopir 2012; Tenopir 2017.) Lisäksi Fecherin, Friesiken ja Hebingin (2014) systemaattisessa kirjallisuuskatsauksessa nousivat esiin niin ikään ongelmat metatietojen tuottamisesta sekä puutteellisesta metatietostandardien käytöstä.

Ala-Kyynyn, Roinilan ja Korhosen (2017) haastattelututkimuksen, Borgin ja Kuulan (2007) selvityksen sekä muun edellä esitellyn tutkimusnäytön perusteella on aiheellista tutkia systemaattisesti tutkimusorganisaatioiden tuottamien tutkimusaineistojen metatietojen laatua. Tutkielman tulosten perusteella saadaan kartoittava kuva tutkimusaineistojen metatietojen laadusta. Lisäksi tulosten perusteella voidaan löytää mahdolliset yksittäiset ongelmakohdat metatietojen tuottamisessa ja kehittää kirjastojen tutkimusdatapalveluja kansainvälistä ja kansallista tiedepolitiikkaa tukevaksi.

Ensin tarkastellaan tämän tutkimuksen kannalta keskeisiä käsitteitä luvussa 2. Tämän työn kannalta keskeisiä käsitteitä ovat avoin tiede ja tutkimus, tieteen avoimuus, aineistonhallinta ja FAIR-periaatteet. Lisäksi luvussa 2. tehdään katsaus kirjaston tutkimuspalveluihin. Luvussa 3 käydään läpi metatiedon määritelmä, mitä metatiedon laatu ja aiempaa tutkimusta metatietojen laadun mittauksen näkökulmasta. Luvussa 4 esitellään aiempi tutkimus ja luvussa 5 tutkimusasetelma. Luvussa esitellään 6 tulokset ja luvussa 7 johtopäätökset.

2 AVOIN TIEDE

Tässä luvussa tutustutaan avoimen tieteen ja tutkimuksen käsitteeseen. Ensimmäisessä alaluvussa tutustutaan avoimeen tieteeseen, sen eri osa-alueisiin ja näiden tuottamiin hyötyihin. Toisessa alaluvussa tarkastellaan avoimen tieteen vaikutuksista tieteellisiin kirjastoihin ja kirjaston tarjoamiin avoimen tieteen tutkimuspalveluihin.

2.1 Avoin tiede ja tutkimus

Avoin tiede on kansainvälisesti nostetta saanut tiedepoliittinen liike, jonka tarkoitus on tehdä tutkimuksesta mahdollisimman avointa. Käytännössä tämä tarkoittaa sitä, että tutkimuksen teossa tuotetut julkaisut, tutkimusaineistot ja menetelmät ovat ilmaiseksi kaikkien saatavilla. Suomessa avoimen tieteen kansallisen koordinaation voi karkeasti jakaa avoimeen julkaisemiseen, FAIR-periaatteilla tuotettuun tutkimusaineistoon ja avoimuuden kulttuuriin. (UniFI 2018.)

Avoin tiede ja tutkimus on saavuttanut viime vuosien aikana keskeisen aseman kansainvälisissä ja kansallisissa tiedepoliittisissa linjauksissa. Avoimuus on noussut merkittäväksi tavaksi edistää tiedettä, tutkimusta ja tieteellisen tiedon laatua. Tieteen avoimuudella vahvistetaan myös tieteellisen tiedon luotettavuutta, näkyvyyttä ja vaikuttavuutta. (Royal Society 2012, 41-42; Council of the European Union 2016; Farnham ym. 2017; Munafo 2017 ym.)

Suomi on sitoutunut kansainväliseen avoimen tieteen liikkeeseen. Vuonna 2014 Opetus- ja kulttuuriministeriö (OKM) perusti strategia- ja asiantuntijaryhmän sekä käynnisti Avoin tiede ja tutkimus -hankkeen (ATT-hanke)[1], jonka tavoite oli tehdä Suomesta johtava maa tieteen ja tutkimuksen avoimuudessa (Opetus- ja kulttuuriministeriö 2014.) Hanke päättyi vuonna 2017, jonka jälkeen avointa tiedettä on Suomessa koordinoinut Tieteellisten seurain valtuuskunta (TSV).

[1] ATT-hankkeen loppuraportti on luettavissa http://urn.fi/URN:ISBN:978-952-263-560-0

ATT-hankkeen johdosta Suomessa on alettu soveltaa avoimia toimintamalleja tutkimusprosesseissa. Erityisesti korkeakoulut, tutkimuslaitokset ja rahoittajat ovat kehittäneet toimintamallejaan avoimen tieteen ja tutkimuksen -periaatteiden mukaisesti OKM:n ohjauksessa. OKM edellyttää Suomessa toimivien tutkimusorganisaatioiden sitoutumista avoimen tieteen ja tutkimuksen linjauksiin. (Avoin tiede 2019.)

Avoin tiede ja tutkimus on saanut viime vuosina mukaan useita merkittäviä tutkimusrahoittajia (National Science Foundation 2011; Rockey 2012). Kansallisista rahoittajista Suomen Akatemia edellyttää avointa julkaisemista, aineistonhallintasuunnitelmaa (ks. luku 2.1.2) ja tutkimusaineiston avaamista (Suomen Akatemia 2019b). Business Finland (ent. Tekes) edellyttää avointa julkaisemista (Business Finland 2018). Kansainvälisistä rahoittajista merkittävin on ollut EU-rahoitteinen Horizon2020, joka vaatii täyttä avoimuutta kattaen julkaisut, tutkimusaineistot ja menetelmät sekä yksityiskohtaiset suunnitelmat näiden toteuttamiseksi (H2020 Programme 2017).

Tutkimusrahoittajat ovat tehneet myös yhteistyötä avoimuuden edistämiseksi. Vuonna 2018 cOALition S -yhteenliittymä, joka koostui 11 merkittävästä eurooppalaisesta tutkimusrahoittajasta julkaisi ”Plan S”-suunnitelman, jonka tavoite on saada kaikki julkisella rahoituksella toteutetut tutkimusjjulkaisut vapaasti saatavaksi vuodesta 2020 alkaen (Laine 2018a.) Plan S -suunnitelmaa on päivitetty vuonna 2019 ja uudessa suunnitelmassa asetettiin tavoitteeksi saada tutkimusjulkaisut avoimeksi vuodesta 2021 (cOALition S 2019).

2.1.1 Tieteen avoimuus

Avoimuus on perinteisesti ollut tieteen ja tutkimuksen keskeisiä periaatteita (Munthe & Welin 1996). Avoimuus edistää tieteen luotettavuutta, näkyvyyttä, löydettävyyttä ja toistettavuutta, joiden pohjalta tiedettä voidaan kutsua itseään korjaavaksi menetelmäksi. Tutkimusaineistojen, tulosten ja menetelmien avoin jakaminen kehittää tieteen avoimuuden toimintamalleja ja tarjoaa välineitä tieteen laadunhallintaan. Keskeinen syy avoimuuteen on tieteen, tieteellisen tiedon ja tutkimuksen laadun jatkuva kehittäminen. (Royal Society 2012, 41-42; Farnham ym. 2017; Munafo 2017 ym.)

Informaatioteknologian ja tietoteknisten apuvälineiden kehitys on vaikuttanut positiivisesti akateemisten prosessien tuottavuuteen ja tästä syystä tieteelliset käytännöt ovat olleet murroksessa. Akateemisilla prosesseilla tarkoitetaan tutkimustyön suunnittelua, tutkimustyötä, tieteellistä julkaisemista ja tieteellistä viestintää sekä tiedekentällä tehtävää tutkijoiden välistä yhteistyötä. Kehitys on mahdollistanut akateemisten toimijoiden yhteistyön entistä tiiviimmäksi ja paikasta riippumattomaksi (Ding 2010.) Negatiivisena puolena kasvavassa tieto- ja julkaisumäärässä on tieteessä ja tutkimuksessa alati kasvava laatuongelma (Steen 2011; Royal Society 2012, 41-43; Sarewitz 2016; Smaldino & McElreath 2016). Lisäksi tutkimuksien toistettavuuteen liittyy suuria luotettavuusongelmia (Buck 2015; Baker 2016; Lowndes ym. 2017). Esimerkiksi 70 % tutkijoista on epäonnistunut toisten tutkimustulosten toistamisessa ja 52 % piti asiaa tieteellisenä kriisinä (Baker 2016).

Avoin tiede ja tutkimus ja sen periaatteiden soveltaminen tieteellisiin prosesseihin voisi olla osittainen ratkaisu tieteen laatuongelmiin (Royal Society 2012; Nosek ym. 2015; Munafó ym. 2017). Myös avoimen lähdekoodin ja työkalujen hyödyntäminen voi auttaa parempaan tieteeseen ja esimerkiksi ilmastonmuutoksen vastaisessa taistelussa avoimuus ja omien tieteellisten tuotoksien jakaminen on erityisen tärkeää (Lowndes ym. 2017). On myös huomioitava, että pelkkä tutkimusaineistojen jakaminen ei riitä vaan avoimen tieteen tuomien hyötyjen täysimittainen hyödyntäminen vaatii myös kulttuurin kehittymistä tutkimusaineiston uudelleenkäytössä (Pasquetto, Randless & Borgman 2017).

Avoimen tieteen hyödyt tutkijoille

Avoin julkaiseminen lisää tutkijoiden saamia viittauksia sekä artikkeleiden saamia lukumääriä. Tämä käsite tunnetaan nimellä ”Open Access Citation Advantage (OACA)” (Swan 2010; Piwowar ym. 2018). Avointen julkaisujen kasvanutta viittausten määrää on tutkittu paljon. Kattavasti sitä on tutkinut esimerkiksi Archambault ym. (2014), Ottaviani (2016) ja Piwowar ym. (2018). Kaikissa tutkimuksissa ei kuitenkaan ole saatu tälle ilmiölle tukea ja julkaisujen viittausmetriikan mittaamisen käytetyt menetelmät ovat saaneet kritiikkiä (Davis 2011).

Avoimiin julkaisuihin kohdistuneiden viittausten lisäksi tutkijat saavat enemmän viittauksia ja näkyvyyttä avoimena jaettuun tutkimusaineistoon. Julkaisuihin ja tutkimusaineistoihin kohdistuvat viittaukset voivat myös vahvistaa tutkijoiden välistä kommunikaatiota ja tutkimusyhteistyötä. Lisäksi laadukkaasti dokumentoitu ja avoimella lisenssillä [2] jaettu tutkimusaineisto edistää tieteen laatua kehittäen tieteen keskeisiä periaatteita kuten toistettavuus, luotettavuus ja läpinäkyvyys. Tutkimusaineiston viittauksien määrä voi myös korreloida laadun kanssa ja auttaa tutkijaa meritoitumisessa. Muut tutkijat voivat hyödyntää tutkimusaineistoa joko toistamalla tutkimuksen tai tekemällä sen uudesta näkökulmasta eri tutkimuskysymysten vastaten (Waaijers, L & van der Graaf, M. 2011; Borg 2013; Amorim et al. 2016; McKiernan ym. 2016; Koltay 2016; Laine 2018b). Lisäksi löytyy näyttöä siitä, että tutkimusjulkaisunsa ohessa tutkimusaineistonsa julkaiseva saa enemmän viittauksia (Piwowar, Day & Fridsma 2007; Piwowar & Chapman 2010).

Suomessa tutkimusaineiston aineistoviittausten suhteen ollaan pitkällä, sillä Suomessa on luotu tiekartta tutkimusaineistojen viittauskäytänteistä (ks. Laine & Nykyri 2018). Tutkimusaineistoihin viitatessa keskeiseksi tekijäksi nousee ensisijaisesti tutkijoiden panos tutkimusaineistojen kuvailussa (Amorim et al. 2016; Laine & Nykyri 2018). Tutkijat eivät välttämättä tunne aineistonhallinnan kaikkia lainalaisuuksia, mutta pystyvät kuitenkin tarvittaessa tuottamaan muodollisia (formaaleja) kuvauksia ja geneeristä kuvailutietoa eli metatietoa omasta tutkimusaineistostaan (Amorim et al. 2016).

Käytännössä tutkimusaineistoon viittaaminen tapahtuu samoilla tutkimuksellisilla periaatteilla kuin julkaisuihinkin viittaaminen: ”Tunnustuksen antaminen toisten tekemälle tutkimustyölle on yksi tutkijan ammattietiikan tärkeimpiä periaatteita.” (Laine & Nykyri 2018). Lisääntyneen viittausmäärän johdosta tutkijat saavat enemmän näkyvyyttä sosiaalisessa mediassa. Sosiaalisen median vaikuttavuus on tutkijan näkyvyyden kannalta

[2] Avoimilla lisensseillä voidaan antaa käyttäjälle enemmän oikeuksia, suljetuilla taas rajata ks. https://creativecommons.fi/lisenssit/

tärkeää, koska se lisää yhteiskunnallista näkyvyyttä. (McKiernan ym. 2016; Tennant 2017b.)

Tutkimusorganisaatioiden lisäksi valtiot, julkinen sektori ja yksityisen sektorin toimijat ovat alkaneet avata omia aineistojaan kaikille avoimeksi (Welle Donker & van Loenen 2016). Erityisesti yksityisen sektorin kiinnostuminen avoimista tutkimusaineistoista ei ole yllättävää, sillä avointa dataa käyttävät yritykset kasvavat muita yrityksiä nopeammin erityisesti tietointensiivillä sektoreilla (Koski ym. 2017, 69-70).

2.1.2 Aineistonhallinta osana tutkimusprosessia

Tietoarkiston[3] aineistonhallinnan käsikirjan mukaan aineistonhallinnalla tarkoitetaan tutkimusaineiston kokonaisvaltaista hallintaa, joka kattaa datan ja siihen liittyvän metatiedon. Aineistonhallinnalla on tarkoitus varmistaa, että tutkimusaineisto säilyy käyttökuntoisena tutkimuksen tai hankkeen ajan sekä sen jälkeen. Laadukkaasti toteutettu tutkimusaineiston hallinta on osa hyvää tieteellistä käytäntöä. Aineistonhallinnan keskiössä ovat laadukkaat metatiedot, jotka mahdollistavat tutkimusaineiston löytymisen, tulkinnan ja uudelleenkäytön. (Tietoarkiston aineistonhallinnan käsikirja 2017.)

Aineistonhallintaan sisältyy myös tutkimusaineiston organisoinnin suunnittelua alkaen aineiston keräämisestä ja päättyen sen jakamiseen tai arkistointiin. Aineistonhallinnan tarkoitus on edesauttaa tutkimusaineiston luotettavuutta ja tutkimustulosten verifiointia (Whyte & Tedds 2011). Tätä ajattelutapaa voidaan kutsua elinkaariajatteluksi[4], jossa

[3] ”Tietoarkisto on tutkijoita kansainvälisesti palvelevan eurooppalaisten tietoarkistojen yhteenliittymän CESSDAn (Consortium of European Social Science Data Archives) Suomen kansallinen palveluntuottaja.” (ks. https://www.fsd.uta.fi/fi/tietoarkisto/#tietoarkisto-organisaationa). Vuonna 2017 Tietoarkisto sai Core Trust Seal -sertifikaatin, joka kertoo luotettavuudesta aineistojen pitkäaikaissäilytyksen ja jatkokäytön mahdollistajana (ks. https://www.fsd.uta.fi/fi/tietoarkisto/asiakirjat/vuke2017.pdf)

[4] Tutkimusaineiston elinkaariajattelu on yleistynyt avoimen tieteen ja tutkimuksen yleistymisen myötä. Elinkaariajattelun tiivistää Tietoarkiston aineistonhallinnan käsikirja: ”Tutkimusaineiston elinkaari on tavallisesti pidempi kuin aineiston tuottaneen tutkimushankkeen elinkaari. Hanke päättyy, kun rahoitus loppuu, mutta aineistoa voi ja kannattaa hyödyntää myöhemminkin.”
ks. https://www.fsd.uta.fi/aineistonhallinta/fi/miksi-aineistonhallintaa-ja-jatkokayttoa.html#elinkaari

tutkimusaineiston hallintaan liittyy useita aktiviteetteja ja prosesseja kuten aineiston tuottaminen, tietoturvasta ja tietosuojasta huolehtiminen, säilyttäminen, jakaminen ja uudelleenkäyttö sekä lopuksi tallentaminen ja julkaiseminen data- tai metatietoarkistossa[5] (Borg & Kuula 2007, 12; Cox & Penfield 2014). Näin ollen aineistonhallinnalla voidaan varmistaa myös tutkimusaineiston pitkäaikaissäilytys (PAS)[6] (Wooijers & van der Graaf 2011).

Aineistonhallinta ei ole yksinkertainen asia eikä se toteudu ilman tutkimusorganisaation tukea vaan siihen pitää panostaa. Laadukas aineistonhallinta vaatii monitahoista vaivannäköä, joka kattaa teknologian, työkalut, aineistopolitiikan, metatiedot ja vakiintuneet aineistokäytännöt. (Mayernik, Batcheller & Borgman 2011). Tiukentuneet vaatimukset aineistonhallinnasta ja aineistonhallintasuunnitelmista eivät välttämät näy vielä organisatorisella tasolla, jossa tehdään päätöksiä resursseista. Tutkimusinfrastruktuurit, menetelmät ja työkalut eivät ole pysyneet vaatimusten perässä, mikä saattaa aiheuttaa ongelmia mm. tutkimusaineiston dokumentoinnin kannalta. (Karimova 2017.)

Tieteellinen tutkimus ei ole standardisoitua eikä tutkimusaineistojen keräämiselle, prosessoinnille ja jakamiselle ole täysin yhtenäisiä käytäntöjä. Tutkijat tuottavat tutkimusaineistoa erilaisilla menetelmillä, eri formaateilla ja eri tieteenaloilla. Jotta tutkimusaineisto olisi tulkittavissa ja käytettävissä, pitää sitä täydentää laadukkailla metatiedoilla. (Mayernik, Batcheller & Borgman 2011.)

[5] Data- ja metatietoarkistot (repositoriot) ovat palveluita, jotka on suunniteltu tutkimusaineiston julkaisemiseen, hakemiseen ja säilyttämiseen. Arkistoja hyödyntämällä tuetaan uudelleen käyttöä ja voidaan lisätä dataviittausten kautta tutkimuksen vaikuttavuutta. Osa arkistoista tukee sekä metatiedon ja varsinaisen tutkimusaineiston tallentamista ja osa pelkästään metatiedon tallentamista ks. https://www.aalto.fi/fi/palvelut/avoimet-data-arkistot-repositoryt

[6] Pitkäaikaissäilytys eli PAS eroaa normaalista tutkimusaineistojen säilyttämisestä lähinnä säilytettävän ajan perusteella. Pitkäaikaissäilytyksen tarkoitus on ylläpitää digitaalisen informaation käytettävyyttä kymmenistä vuosista jopa sataan vuoteen, ks. http://www.digitalpreservation.fi/

Laadukkaalla aineistonhallinnalla varmistetaan tiedon löytyminen ja uudet innovaatiot sekä mahdollistetaan tiedeyhteisölle aineistojen uudelleenkäyttö. Laadukasta aineistonhallintaa ohjaavat FAIR-periaatteet (ks. kappale 2.1.3), joita voidaan soveltaa koko aineistonhallinnan elinkaaren ajan. Hyvästä aineistonhallinnasta hyötyvät erityisesti tutkijat ja tutkijayhteisöt, jotka voivat jakaa ja käyttää hyvin dokumentoituja tutkimusaineistoja sekä saada tutkimusaineiston viittauksista samalla tapaa mainetta kuin esimerkiksi julkaisujen viittauksista (Wilkinson 2016). Merkittävät rahoittajat Yhdysvalloissa ja Euroopassa edellyttävät tutkijoilta aineistonhallintaa rahoituksen vastineeksi. Käytännössä rahoittajat edellyttävät aineistonhallintasuunnitelmaa (Karimova 2017).

Aineistonhallintasuunnitelma osana aineistonhallintaa

Aineistonhallintasuunnitelma on osa aineistonhallintaa. Aineistonhallintasuunnitelmassa kuvataan, miten tutkimusaineistoa kerätään, käytetään ja säilytetään koko tutkimuksen elinkaaren ajan sekä mitä aineistolle tehdään tutkimuksen päätyttyä. Keskeistä aineistonhallintasuunnitelmassa on kuvata tutkimusaineiston dokumentointia eli metatietojen tuottamista. Lisäksi aineistonhallintasuunnitelmassa tulee huomioida etiikka ja tietoturva ja juridiset asiat kuten tietosuoja ja lisensointi mahdollista uudelleenkäyttöä varten. (Freudenberg 2016 ym.; Tietoarkiston aineistonhallinnan käsikirja 2017; Karimova 2017.)

Aineistonhallinta ja siihen liittyvä aineistonhallintasuunnitelma on kriittinen askel kohti monimuotoisempaa aineistojen uudelleenkäyttöä. Aineistonhallintasuunnitelma ei ole staattinen dokumentti vaan sitä on tarkoitus päivittää jatkuvasti. Sen jatkuva päivittäminen koko tutkimusprosessin ajan saattaa johtaa laadukkaampaan aineistonhallintaan ja käytettävämpään tutkimusaineistoon. (Karimova 2017.)

2.1.3 FAIR-periaatteet

Vuonna 2016 julkaistut FAIR-periaatteet ovat muodostuneet osaksi tutkimusaineistojen hallintaa ja avoimuuden edistämistä. Periaatteiden noudattamisesta on EU neuvoston linjaus vuodelta 2016 (Wilkinson ym. 2016; Council of the European Union 2016). FAIR-

periaatteiden tarkoitus on kehittää aineistonhallintaa, vahvistaa tutkimuksen näkyvyyttä ja vaikuttavuutta saattamalla tutkimusaineisto avoimesti käytettäväksi sekä tuoda elinkaariajattelu osaksi tutkimusaineistoja (Fairdata 2018, Kaipainen 2018).

FAIR-periaatteet muodostuvat löydettävyydestä (*findable),* saavutettavuudesta (*accessible)*, yhteentoimivuudesta (*interoperable)* ja uudellenkäytettävyydestä (*re-usable).* Löydettävyydellä tarkoitetaan yksilöivää tunnistetta ja rikasta metatietoa. Saavutettavuudella metatietojen löydettävyys yksilöidyn tunnisteen perusteella ja avoimien tiedonsiirtomenetelmien hyödyntämistä. Yhteentoimivuudella vakiintuneita kontrolloituja sanastoja ja metatietojen koneluettavuutta[7]. Uudelleenkäytettävyydellä kattavaa ja ymmärrettävää metatietoa sekä käyttöehtojen, tekijänoikeuksien ja lisensoinnin huomioimista. (Wilkinson ym. 2016; Force11 2016.)

Kansainvälisesti merkittävistä tutkimusrahoittajista esimerkiksi Euroopan Unioni (EU), jonka alla toimii merkittävä hankeohjelma Horizon2020 on sitoutunut Fair-periaatteisiin (H2020 Programme 2016). Samoin Yhdysvaltain terveysvirasto (*National Institutes of Health, NIH)* on sitoutunut Fair-periaatteisiin (National Institutes of Health 2019). Kansallisista rahoittajista Suomen Akatemia mainitsee FAIR-periaatteet osana aineistonhallintaa (Suomen Akatemia 2019a). Suomen valtioneuvoston tuottamassa tutkimuksessa suositellaan myös FAIR-periaatteeseen pohjautuvaa ajattelua osaksi julkisen hallinnon tietovarantojen hyödyntämisessä (Koski ym. 2017, 57, 67).

FAIR-periaatteet ovat näkyneet myös kansainvälisissä tiedepoliittisissa linjauksissa. Euroopan komission toimenpideohjelma FAIR-periaatteiden käytöstä julkaistiin 2018. Toi-

[7] Koneluettavuus (eng. *Machine-readable data*) tarkoittaa dataa tai metatietoja, jonka tietokone osaa käsitellä. Esimerkiksi jonkin standardoidun tietomallin mukaan järjestetty data (XML, CSV, JSON). ks. http://opendatahandbook.org/glossary/en/terms/machine-readable/

menpideohjelmaa voidaan pitää jatkumona EU-neuvoston linjaukselle FAIR-periaatteiden käytöstä. (Council of the European Union 2016; European Commission Expert Group 2018.)

Suomessa tiedepolitiikasta vastaava Opetus- ja kulttuuriministeriö on sitoutunut FAIR-periaatteisiin. Opetus- ja kulttuuriministeriö on rahoittanut CSC:n[8] toteuttaman Fairdata-palvelun[9], jonka tarkoitus on edistää tutkimusaineistojen hallintaa koko elinkaaren ajan (Fairdata 2018). Suomen yliopistojen rehtorineuvoston (UNIFI RY) on julkaissut Suomen oman toimenpideohjelman avoimesta tieteestä ja datasta. FAIR-periaatteet nähdään tarkentavana käsitteenä tutkimusaineiston ja metatietojen osalta (Unifi 2018).

2.2 Avoin tiede ja tutkimus kirjastoissa

Kansainväliset ja kansalliset tiedepoliittiset linjaukset vaikuttavat korkeakouluissa tehtävään tutkimukseen (ks. EOSC Declaration). Avoimen tieteen ja tutkimuksen toimintamallien omaksuminen sekä rahoittajien vaatimusten kiristyminen aiheuttavat uusia palvelutarpeita, joista osa kohdistuu korkeakoulujen tieteellisiin kirjastoihin. Tieteellisten kirjastojen rooli tutkimuksen tukipalveluina on vakiintunut, mutta uudistuva toimintaympäristö aiheuttaa myös kirjastoille paineita kehittää palveluitaan. (Corrall 2014; Ogungbeni ym. 2016.)

2.2.1 Tutkimuspalvelut tieteellisissä kirjastoissa

Tieteellisen kirjaston rooli tutkimuspalveluiden tuottajana on korkeakouluissa vakiintunut. Tutkimuspalveluita tarvitaan entistä enemmän toimintaympäristön muutoksen

[8] CSC eli tieteen tietotekniikan keskus ks. https://www.csc.fi/fi

[9] Fairdata-palvelut kattavat tällä hetkellä (2019) pitkäaikaissäilytyspalvelun (FAIRDATA-PAS), tutkimusaineiston arkistointipalvelun (IDA), tutkimusaineiston kuvailupalvelun (Qvain) ja tutkimusaineiston metatietojen julkaisu- ja hakupalvelun (Etsin) Fairdata-palveluiden kokonaisuuteen voi tutustua osoitteessa https://www.fairdata.fi/

vuoksi ja kirjastot voivat olla mukana koko tutkimuksen elinkaaren ajan. Kirjasto voitaisiinkin nähdä suuremman roolinsa takia enemmän tutkimuskumppanina kuin pelkkänä tukipalveluna. (Jaguzewski 2013; Corrall, Kennan & Afzal 2013; Corrall 2014.)

Korkeakoulujen tieteellisiltä kirjastoilta odotetaan räätälöityjä tutkimuspalveluita tutkimuksen tueksi. Kirjastojen tarjoamat tutkimuspalvelut ovat perinteisesti kattaneet bibliometrian, tiedonhaun ja tiedonhallinnan (Auckland 2012, 16-31). Tutkimuspalveluihin voidaan lukea myös julkaisu- ja rinnakkaistallennuspalvelut, laajentuneet metriikkapalvelut[10] ja aineistonhallintaan liittyvät palvelut (Corrall 2014). Haasteeksi palvelujen tarjoamisessa on Housewrightin mukaan (ks. Housewright 2013a; Housewright 2013b) kyselytutkimuksissa ilmi tullut tutkijoiden vähäinen arvostus kirjaston tarjoamia tutkimuspalveluita kohtaan. Lisäksi tutkijat saattavat suhtautua kirjastoon edelleen vain kirjojen ja julkaisujen välittäjänä (Koltay 2016).

Useassa korkeakoulussa organisaation sisäistä yhteistyötä on tiivistetty tutkimuspalveluiden kehittämiseksi. Käytännön esimerkkinä voidaan mainita kirjaston ja tietohallinnon yhteistyö. Coxin, Kennanin, Lyonin ja Pinfieldin (2017) tutkimuksessa oli tavoite tutkia myös sitä, millä tavalla kirjastoalan osaamista voidaan hyödyntää tutkimuspalveluissa, millaisia uusia rooleja se tuo kirjastoalan ammattilaisille ja mitä uutta osaamista tutkimuspalveluiden tuottamisen tarvitaan, jotta kirjasto voisi palvella monipuolisesti tutkijoita. Kirjastojen uudet tutkimuspalvelut tuottivat uusia rooleja (esim. uusia työnimikkeitä). Uudet roolit ja osaamistarpeet liittyivät pääasiassa julkaisupalveluihin, julkaisuarkiston hallinnointiin, metriikkapalveluihin, aineistonhallintaan sekä metatietojen laadun mittaukseen. (mts.)

Aineistonhallinta on keskeisessä roolissa, mikäli tutkimusaineistoja halutaan jakaa avoimesti kaikkien saataville. Tieteellisten kirjastojen panos aineistonhallinnan palveluihin

[10] Laajentuneet metriikkapalvelut sisältävät bibliometrian lisäksi altmetriikan, jolla tarkastellaan julkaisujen näkyvyyttä esimerkiksi mediassa ja sosiaalisessa mediassa (ks. Forsman & Englund 2013) sekä datametriikan, jolla seurataan tutkimusaineistojen viittauksia ja vaikuttavuutta (ks. Laine & Nykyri 2018).

on korkeakouluissa suuri ja kirjastoalan osaamista voidaan hyödyntää erityisesti aineistonhallinnan dokumentointiin ja metatietojen tuottamiseen liittyvissä tehtävissä (Yoon & Schultz 2017). Aineistonhallintaan liittyvät palvelut voidaan ajatella jatkeena perinteisille kirjastopalveluille. Käytännössä aineistonhallinnan palvelut koostuvat oppaiden tekemisestä, tutkimusaineistojen viittauskäytäntöjen opastuksesta ja aineistonhallintasuunnitelman tuesta, joka kattaa koulutuksen, konsultoinnin ja ohjauksen. Kirjastot voivat lisäksi tarjota teknistä tukea tutkimusaineiston säilytyspalveluiden käyttöön (Koltay 2016; Yoon & Schultz 2017; Tenopir 2017).

Tärkeimmät tieteellisten kirjastojen tarjoamat aineistonhallinnan palvelut ovat metatietoihin liittyvät palvelut kuten metatietojen tuottamisessa avustaminen, oikean metatietostandardin valitseminen ja metatietojen tuottamiseen liittyvät koulutukset. Tutkijoiden tietoisuus metatietojen tärkeydestä ja metatietostandardien käytöstä on hyvin vähäinen ja sillä on suora vaikutus aineistonhallinnan dokumentoinnin laatuun. (Koltay 2016; Tenopir 2017; Vilar & Kabucevic 2019.)

Tutkijoiden ongelmat metatietojen tuottamisessa ovat monitahoisia. Tutkijat kokevat, että metatietostandardeja on liikaa ja niitä tuotetaan liian nopealla tahdilla, niiden käytettävyys on huono ja niiden hyödyntämiseen ei ole tarvittavia työkaluja. Tämän vuoksi metatietojen tuottaminen voi jäädä kesken tai kokonaan tekemättä. Metatietoja tuotetaan myös epämuodollisesti ilman standardeja, mikä vaikeuttaa tutkimusaineiston tulkintaa myöhemmin. Tutkimusryhmien sisällä voi olla hyvin erilaiset käytäntöjä metatietojen tuottamiselle. Yleensä metatietojen tuottaminen voi olla useamman henkilön vastuulla heidän erikoisosaamiseensa perusteella. Tämä käytäntö johti sekalaisiin metatietoihin. (Mayernik, Batcheller & Borgman 2011.)

Toisaalta on huomioitava, että tässä Mayernikin, Batchellerin & Borgmanin toteuttamassa etnografisessa tutkimuksessa käytänteet ja mielipiteet vaihtelivat suuresti sen mukaan, millainen tutkijaryhmien sitoutumisen taso oli metatietojen tuottamiseen. Tut-

kijat jakoivat tutkimusryhmät kolmeen ryhmään: 1. tilapäinen sitoutuneisuus, 2. epäsäännöllinen sitoutuneisuus ja 3. vakinainen sitoutuneisuus, joista 3. ryhmällä oli positiivisimmat ja 1. ryhmällä negatiivisimmat mielikuvat metatietojen tuottamisesta.

2.2.2 Tutkijat tutkimuspalveluiden käyttäjinä

Tutkimuspalveluiden käyttöä kartoittavassa kyselytutkimuksessa tutkijat kertoivat olevansa valmiita jakamaan tutkimusaineistojaan ja uudelleenkäyttämään muiden tutkimusaineistoja. Ongelmaksi tutkijat kokivat kuitenkin sen, että organisaation tuki ei ole tarpeeksi riittävää. Tutkijat eivät myöskään olleet tyytyväisiä työkaluihin, jolle heidän oli tarkoitus tuottaa metatietoa eikä heillä ollut täyttä ymmärrystä metatietojen tärkeydestä tai metatietostandardeista. (Tenopir 2011.)

Aineistonhallintapalveluihin liittyvässä kyselytutkimuksessa kartoitettiin aineistojenhallintaan kohdistuvia palveluita ja asenteita tutkimusaineiston jakamisesta. Valtaosa vastanneista piti palveluiden kehittämistä tärkeänä. Tutkijat olivat huolissaan tutkimusaineiston väärinkäytöstä ja asianmukaisesta viittaamisesta. Tutkimuksessa tuli myös ilmi tutkijoiden puutteelliset tiedot hyvästä aineistonhallinnasta sekä metatietostandardien käytöstä. (Tenopir 2018.)

Tutkijoiden asenteita tutkimusaineiston avoimuudesta, jakamisesta ja dokumentoinnista selvitettiin kyselytutkimuksessa. Dokumentoinnilla tarkoitettiin aineistonhallintasuunnitelmaa ja metatietojen tuottamista. Suurin huolenaihe tutkimuksessa oli tutkijoiden tietämättömyys oman organisaation aineistonhallintasuunnitelman linjauksesta, sillä vain 9 % tutkijoista oli tietoinen siitä, suositteleeko oma organisaatio aineistonhallintasuunnitelmaa. Metatietojen suhteen 8 % tutkijoista tiesi mitä metatietostandardia organisaatio suosittaa ja suurin osa ei koskaan käyttänyt mitään metatietostandardia. Toisaalta tutkijoista lähes 70 % tiesi mitä metatieto ylipäätään tarkoittaa. Kolmannes tutkijoista uskoi, että aineistonhallintasuunnitelma voisi olla hyödyllinen ja yli puolet uskoi, että koulutus olisi hyödyllistä. Lähes puolet oli kiinnostunut myös metatietoihin liittyvästä koulutuksesta. Vain joka neljäs ei ollut kiinnostunut mistään koulutuksesta. (Vilar & Kabucevic 2019.)

Tutkimusten perusteella kirjastot tarjoavat tai ovat kehittämässä palveluitaan tutkijoiden tarpeisiin. Kirjaston rooli tutkimusprosessissa voisi kattaa koko tutkimusprosessin elinkaaren tutkimuksen suunnittelusta tutkimuksen päättymiseen. Kirjastoalan osaamista voidaan hyödyntää erityisesti tiedon-, aineiston- ja metatietojenhallintaan liittyvissä palvelutarpeissa. Tutkijoiden palvelutarpeisiin vastaaminen vaatii myös kehysorganisaation tukea sekä organisatorisia muutoksia. (Corrall, Kennan & Afzal 2013; Corrall 2014; Tenopir 2017; Cox, Kennan, Lyon & Pinfield 2017; Yoon & Schultz 2017.)

3 METATIETO

Kolmannessa luvussa keskitytään metatietoihin. Ensimmäisessä alaluvussa tutustutaan metatiedon käsitteeseen ja metatiedon merkitykseen tutkimusaineistojen hallinnassa. Toisessa alaluvussa perehdystään metatietostandardeihin ensin yleisemmällä tasolla ja sen jälkeen keskittyen tutkielman kannalta relevantteihin tutkimusaineistojen metatietostandardeihin. Kolmannessa ja neljännessä alaluvussa rakennetaan teoriapohja metatietojen laadulle tutustumalla laadun käsitteeseen sekä kirjallisuudesta poimittuihin metatietojen laatumalleihin.

3.1 Metatieto

Metatieto on yksinkertaisimmillaan rakenteista tietoa tiedosta eli kuvailevaa tietoa tietystä kohteesta tai entiteetistä. Kohde tai entiteetti voi olla esimerkiksi kirja, asiakirja, kuva tai tutkimusaineisto. Metatieto auttaa ymmärtämään sitä kokonaisuutta, johon metatieto on yhteydessä ja kertoo miten, milloin, kenen toimesta ja millä tavalla kuvailtava kokonaisuus on järjestetty (SFS 5895 2001, 3; Digital Curation Centre 2007; Gilliland 2008, 1-2; Palavitsinis 2013, 22-24; Lei Zeng & Qin 2016, 11-14). Metatieto voidaan määritellä muodollisemmin strukturoituna, koodimuotoisena kuvauksena erilaisia tietoja sisältävän entiteetin ominaisuuksista. Entiteetin kuvailu auttaa hahmottamaan sen tunnistettavuutta, löydettävyyttä, ja hallittavuutta (Lei Zeng & Qin 2016, 11-14).

Käytännössä esimerkiksi kirjastojen ylläpitämät tietokannat sisältävät runsaasti metatietoa. Tietokannoissa olevat aineistot kuten kirjat ovat yksityiskohtaisesti kuvailtuja, jotta aineiston hahmottaminen, löytäminen ja hallitseminen on helpompaa. Samoin elintarvikkeiden ainesosaluettelot, jotka kertovat tuotteen sisältämän kalorimäärän ja valmistusaineet, ovat määritelmällisesti metatietoa (Lei Zeng & Qin 2016, s. 3-4, 12). Metatietojen ei siis tarvitse olla digitaalisia (Gilliland 2008, 14). Lisäksi sosiaalisen median palveluita kuten YouTube, Spotify tai Instagram käyttävät ihmiset ovat päivittäin tekemisissä metatietojen kanssa. Jokaisen videon, musiikkikappaleen tai kuvan yhteydessä on aina metatietoa mukana (Riley 2017, 4).

Vuonna 1939 perustettu *National Information Standards Organization* (NISO) kehittää ja määrittelee standardeja digitaaliseen tiedonhallintaan. NISO on jakanut metatiedot neljään tyyppiin niiden käyttötarkoituksen mukaan, joita ovat

- kuvaileva metatieto (*descriptive metadata)*
- hallinnollinen metatieto (*administrative metadata)*
- rakenteellinen metatieto (*structured metadata)*
- merkintäkielet (*markup languages)* (Riley 2017, 6-7).

Kuvailevalla metatiedolla parannetaan kohteen löydettävyyttä lisäämällä metatietoihin tietoja kuten otsikko (*title)*, tekijä (*creator)*, asiasanat *(keywords)* ja luomispäivämäärä (*date of creation).* Hallinnollisella metatiedolla kuvaillaan kohteen käyttöoikeudet (*rights management metadata)*, jotta kohdetta ei mahdollisesti kopioida ja käytetä väärin. Lisäksi hallinnollisiin metatietoihin kuuluu myös teknistä käytettävyyttä kuvaavia tietoja (*technical metadata)* ja pitkäaikaista käyttöä tukevat tiedot (*preservation metadata*). Merkintäkielien, kuten esimerkiksi XML (*extensible markup language)*[11] *,* avulla metatieto voidaan esittää strukturoidusti halutulla tavalla. Käytännössä metatieto voidaan sijoittaa kuvaillun kohteen yhteyteen. Merkintäkielillä lisätään merkittävästi metatietojen yhteentoimivuutta. Lopuksi rakenteellisella metatiedolla kuvataan metatietotietueen elementtien strukturoitu järjestys. (Lei Zeng & Qin 2014, 18-22; Riley 2017, 6-7.)

Taulukossa 1. tarkastellaan yksityiskohtaisemmin Gillilandin (2008, 9), Lei Zengin ja Qin (2014, 18-22) sekä Rileyn (2017, 7) näkemyksiä metatietojen tyypeistä, niiden ominaisuuksista ja käyttötarkoituksista. Tyyppikartoituksen lisäksi on luotu useita laatumalleja, periaatteita ja käytänteitä, jotka helpottavat laadukkaan metatietojen luomisen helpot-

[11] XML-kieltä voidaan pitää myös standardina ks. https://www.w3.org/TR/xml/

tamiseksi. Tässä tutkielmassa laatumalleja käsitellään tarkemmin luvussa 3.4. Yksityiskohtaisista laatumalleista voidaan erottaa yleisluontoisemmat periaatteet ja käytänteet laadukkaiden metatietojen luomiseen.

Taulukko 1. Metatietojen tyypit käyttötarkoituksen mukaan. (Gilliland 2008, 9; Lei Zeng & Qin 2014, 18-22; Riley 2017, 7).

Metatietojen tyyppi	**Sisältö**	**Käyttötarkoitus**
Kuvaileva metatieto	nimeke, tekijä, aihe, asiasanat, julkaisija, julkaisupäivämäärä	löydettävyys, käytettävyys, yhteentoimivuus
Tekninen metatieto	tiedoston tyyppi, tiedoston koko, luomispäivämäärä	yhteentoimivuus, objektien hallinta, säilytys
Säilytyksen metatieto	tarkistussumma[12]	yhteentoimivuus, objektien hallinta, säilytys
Käyttöoikeuksien metatieto	tekijänoikeus, lisenssit, omistajuus	yhteentoimivuus, objektien hallinta
Rakenteellinen metatieto	järjestys, hierarkia	navigaatio
Merkintäkielet	taulukko, otsikko, lista, nimi, päivämäärä	navigaatio, yhteentoimivuus

Vuonna 2007 NISO julkaisi kolmannen painoksen suosituksesta *A Framework of Guidance for Building Good Digital Collections,* jossa käsitellään digitaalisen kokoelman luomista. Suosituksessa käsitellään myös metatietoa ja esitellään kuusi periaatetta hyvän metatiedon luomiseksi. Hyvä metatieto

- on yhteisön standardin mukainen, joka on kokoelmalle tarkoituksenmukainen ja tukee käyttäjien tarpeita nyt ja tulevaisuudessa
- tukee yhteentoimivuutta

[12] ” Tarkistussummat luodaan kryptograafisilla tiivistefunktioilla, jolloin niiden muuttumattomuus matkan varrella voidaan taata.” ks. https://www.yksityisyydensuoja.fi/oikeellisuuden-tarkistaminen

- käyttää kontrolloituja sanastoja, standardeja ja strukturoituja skeemoja objektien tai entiteettien kuvaamiseen
- määrittelee käyttöehdot
- tukee pitkäaikaista käyttöä, hallintaa ja säilytystä
- kertoo myös metatietoa metatiedosta itsestään eli mitä standardeja ja sanastoja käytettiin, ja luotiinko metatieto automaattisesti vai manuaalisesti.

Jaettava metatieto on tietoa, jota voidaan käyttää sujuvasti sen paikallisen ympäristön ulkopuolella. Jaettavasta ja yhteentoimivasta metatiedosta ovat luoneet oman mallin Shreeves, Riley ja Milewicz (2006). Englanniksiheidän mallinsa tunnetaan nimellä 6C (*content, consistency, cohorence, context, communication ja conformance).* Heidän mallinsa koostuu siis sisällön, johdonmukaisuuden, yhtenäisyyden, kontekstin, viestinnän ja vaatimustenmukaisuuden muodostamasta kokonaisuudesta.

Mallin mukaan sisältö on optimoitu jakamiselle ja metatietoa tuotetaan johdonmukaisesti, joka käytännössä tarkoittaa kontrolloitujen sanastojen ja yhtenäisten syntaksien käyttöä[13]. Näin varmistetaan, että metatieto on yhtenäistä. Kontekstisidonnaisuus on jaettavuuden kannalta keskeistä, sillä metatietojen pitäisi olla tulkittavissa myös alkuperäisestä kontekstistaan irrotettuna. Vaatimustenmukaisuudella pyritään varmistamaan yhteiskäyttöisten ja tunnettujen standardien sekä sääntöjen käyttö, jotta metatieto on jaettavaa ja helpommin tulkittavaa. (Shreeves, Riley & Milewicz 2006.)

Metatietojen käytettävyys, jaettavuus ja yhteentoimivuus vaikuttavat olevan keskeisessä asemassa kirjallisuuden perusteella. NISOn (2007) suosituksessa painotetaan vahvasti yhteentoimivuutta ja se on myös suoraan yksi periaatteista. Suositus ei oleellisesti

[13] Yhtenäiseen muotoon koodattua syntaksia, joka tarkoittaa yhtenäistä muotoa. Esimerkiksi päivämäärien esittämisessä. ks. esim. https://www.w3.org/TR/NOTE-datetime

eroa Lei Zeng ja Qin (2014, 18-22) ja Rileyn (2017, 7) näkemyksistä metatietojen tyyppien luokittelusta. Samoin Shreevesin, Rileyn ja Milewiczin (2006) 6C-malli on pitkälti yhteneväinen muiden esiteltyjen suositusten ja periaatteiden kanssa.

Metatieto tutkimusaineistojen hallinnassa

Laadukkaiden ja kattavien metatietojen tuottaminen tutkimusaineistoista ei ole uusi ilmiö. Jo 1970-luvulla dataa tallennettiin magneettinauhoille ja dokumentaatio tuotettiin paperille. Varhaisimpia dokumentaatiomalleja kutsuttiin nimellä koodikirja (*codebook)*. Koodikirjat ovat jäljitettävissä dokumentaatiotyyppinä 1960-luvun puoliväliin[14]. 1970-luvulla *Inter-university Consortium for Political and Social Research* tuotti OSIRIS-koodikirjan, jota voidaan pitää ensimmäisenä konkreettisena standardina tutkimusaineistojen dokumentointiin. OSIRIS-koodikirja sisälsi ohjeistuksen tutkimusaineiston kuvailuun muuttujatasolla. Eri arkistot yrittivät kehittää standardeja, joissa kuvailua tehtäisiin myös ylemmällä tasolla kuten tutkimuksen otoksesta kertominen, aineiston tuottajan yhteystiedot sekä miten tutkimus on toteutettu. Standardien lisäksi 1970-luvulla perustettiin Council of European Social Science Data Archives (CESSDA).[15] (Blank & Rasmussen 2004.)

Kehitys jatkui merkittävästi 1980-luvulla, jolloin Sue Dodd (1982) julkaisi teoksen ”*Cataloging Machine-Readable Data Files. An Interpretive Manual*” (Vardigan 2013). Internetin kehittyminen 80-luvun lopulla vauhditti entisestään tutkimusaineistojen dokumentaatiokäytäntöjä. Vuonna 1994 kaksi turhautunutta tutkijaa (Bretherton & Singley 1994) kirjoittivat metatiedosta ja sen tarpeellisuudesta tutkimusaineistojen kontekstissa, sillä kasvavaa tietomäärää oli hyvin vaikea hallita. Artikkelissaan tutkijat kuvaavat tieteellisen

[14] 1960-luvun puolivälissä julkaistiin kuvailustandardi MARC (*Machine Readable Cataloging)*, joka on edelleen merkittävässä asemassa tietoaineistojen kuvailussa, (ks. Vardigan 2013).

[15] Vuonna 1976 perustettu CESSDA eli ”Eurooppalaisten yhteiskuntatieteellisten tietoaineistoarkistojen tutkimusinfrastruktuuri” ks. https://www.aka.fi/fi/tiedepoliittinen-toiminta/tutkimusinfrastruktuurit/kansainvaliset-infrastruktuurit-joissa-suomi-on-jasenena/cessda--yhteiskuntatieteellisten-data-arkistojen-hajautettu-infrastruktuuri/

metatietoarkiston, jossa metatieto olisi strukturoidussa ja koneluettavassa muodossa (Bretherton & Singley 1994).

Seuraavana vuonna julkaistussa ”*Preserving scientific data on our physical universe: a new strategy for archiving the nation's scientific information resources”* teoksessa todetaan, että metatietojen tulisi määritellä tutkimusaineiston sisältö, formaatti tai esitystapa, struktuuri ja konteksti, ja että huolellisesti suunniteltu dokumentaatio tai laadukkaat metatiedot voivat edistää tutkimusaineistojen käyttöä (Dozier ym. 1995, 4-6). Ennen kuin Bretherton ja Singley (mts). ja Dozier ym. (mts). julkaisivat näkemyksensä, kuvailivat Rewin ja Davisin (1990) prosessin, jossa tutkimusaineistoa voidaan etsiä, luoda ja jakaa. Rewin ja Davisin (mts.) artikkelissa ei mainita kertaakaan sanaa ”metadata”, mutta kuvataan kuitenkin prosessi, jolla tutkimusaineiston sisältö saadaan ymmärrettäväksi ja yhteentoimivaksi eri järjestelmien välillä.

Tällä hetkellä tieteellisiä tutkimusaineistoja tuotetaan enemmän kuin koskaan ennen. Tämä ilmiö avaa uusia mahdollisuuksia ja haasteita. Nykyisellä tekniikalla ongelma ei ole enää suurten tutkimusaineistojen tallentaminen vaan aineistonhallinta ja tutkimusaineistojen tuottaminen ymmärrettävään muotoon. Tutkimusaineiston kuvailu, luettelointi ja indeksointi sekä säilytys ja käyttöoikeuksien hallinta voidaan ymmärtää kaikki osana aineistonhallintaa. Jos tutkimusaineistoista halutaan tehdä löydettävää, saavutettavaa ja yhteentoimivaa, pitäisi metatietojen luomisessa olla kriittisempi kuin koskaan ja panostaa laadukkaisiin metatietoihin. (Greenberg 2009; Lei Zeng & Qin 2016, 429-432.)

Käytännössä laadukas metatieto auttaa siis ymmärtämään ja käyttämään toisen tuottamaa tutkimusaineistoa (Mayernik, Batcheller & Borgman 2011; Lei Zeng & Qin 2016, 429-431; Amorim et al. 2016; Karimova 2017; Vilar & Zabukovec 2019). Erityisesti tutkimusaineiston jaettavuuden, löydettävyyden ja uudelleenkäytön kannalta laadukas metatieto on erittäin tärkeää (Matthews ym. 2010; Rousidis ym. 2014a; Karimova 2017) ja kuvailussa tulisi suosia metatietostandardeja (ks. luku 3.2.1; Amorim et al 2016). Metatietojen laatu on myös merkittävä tekijä FAIR-periaatteiden ja -kokonaisuuden kannalta (Tietoarkisto 2015; Vilar & Zabukovec 2019).

Tutkijat ovat kiinnostuneita käyttämään tutkimusaineistoja, mutta eivät halua kuvailla niitä muita käyttäjiä varten. Tämä johtaakin usein siihen, että jopa perustietojen kirjaaminen on haasteellista puhumattakaan laadukkaista metatiedoista, jolla varmennettaisiin tutkimusaineistojen uudelleenkäyttö (Edwards ym. 2011). Toisaalta tutkijat arvostavat laadukasta kuvailua ja siitä syntyvää metatietoa helpottamaan löydettävyyttä ja uudelleenkäyttöä, mutta eivät välttämättä ole valmiita käyttämään aikaa laadukkaan metatietojen tuottamiseen (Greenberg 2009). Tämä asettaa haasteita tutkimusaineistojen kuvailuun, sillä kuvailun tulee olla hyvin yksityiskohtaista, jotta tutkimusaineistoa voitaisiin käyttää uudelleen. Haasteena nousee myös se, että tutkimusaineisto saattaa olla pelkästään numeroita ja muuttujia, jolloin sen kuvailu muiden kuin tutkimusaineiston kerääjän toimesta ei ole mahdollista. (Amorim et al. 2016).

Kansallisista rahoittajista Suomen Akatemia ja Business Finland edellyttävät metatiedot osana aineistonhallintasuunnitelmaa (Business Finland 2018; Suomen Akatemia 2019). Merkittävistä kansainvälisistä tutkimusrahoittajista EU-rahoitteinen Horizon2020 (suom. Horisontti2020) asettaa tiukat vaatimukset metatiedolle ja sen julkaisemiselle avoimesti (H2020 Programme, 2016).

3.2 Metatietostandardit

Metatietostandardilla tarkoitetaan yhtenäistä mallia aineiston kuvailuun. Yleensä metatietostandardit kehitetään yhteisöjen yhteistyönä kuvaamaan parhaimmalla mahdollisella tavalla kuvailun kohdetta. Metatietostandardit koostuvat elementeistä (*element set),* jotka määrittävät kyseisen standardin struktuurin ja siinä käytetyt elementit (Digital Curation Centre 2007; Palavitsinis 2013, 22-24; Lei Zeng & Qin 2014, 12, 23). Metatietostandardeja on useita riippuen tiedostomuodoista, tieteenalasta, tutkimuksesta ja kuvailun kohteesta. Tämän takia eri data-arkistoilla on käytössä eri metatietostandardeja. Lisäksi metatietostandardissa on huomioitava mahdollinen pitkäaikaissäilytys (Research Data Alliance, 2017).

Metatietostandardit voidaan jakaa käyttötarkoitukseltaan neljään eri tyyppiin. Standardit tietorakenteille (*data structure)*, tietosisällölle (*data content)*, tietoarvoille (*data values)* ja tiedonsiirtoon (*data exchange)* (ks. Taulukko 2.). Metatietostandardit voivat sisältää useamman standardin kokoelman. Metatietostandardi voi siis sisältää sekä tietorakenteiden että tiedonsiirron standardit ja ohjeistukset tai tietorakennestandardi voi sisältää tietosisällön kuvailun standardin ja ohjeistukset (Taulukko 2).

Taulukko 2. Metatietostandardien tyypit käyttötarkoituksen mukaan. (Lei Zeng & Qin 2016, 23-26, 318).

Standardin tyyppi	**Standardeja**	**Käyttötarkoitus**
Tietorakenteet	Dublin Core VRA Core EAD MARC 21	Määrittää käytettävän tietorakenteen ja semantiikan. Valittavan standardin oltava tarkoituksenmukainen. Esimerkiksi kirjastot käyttävät luetteloinnissa MARC21-formaattia ja arkistot EAD-formaattia.
Tietosisällöt	RDA AACR2 CC DACS	Ohjaa metatietojen luomista ja kuvailua määrittäen kuvailutasot[16], käytettävät sanastot ja tarkoituksenmukaiset elementit kohteen kuvailulle.
Tietoarvot	LCSH AAT TGN DDC LCNAF ISO639-2	Ohjaa metatietojen tiedonorganisointia termilistoilla, auktoriteettitietueilla, taksonomioilla ja ontologioilla
Tiedonsiirrot	ISO2709 XML html5 microdata RDFa, RDX/XML JSON	Määrittää käytettävät formaatit yhteismitalliseen tiedonsiirtoon.

Metatietostandardit tutkimusaineistojen hallinnassa

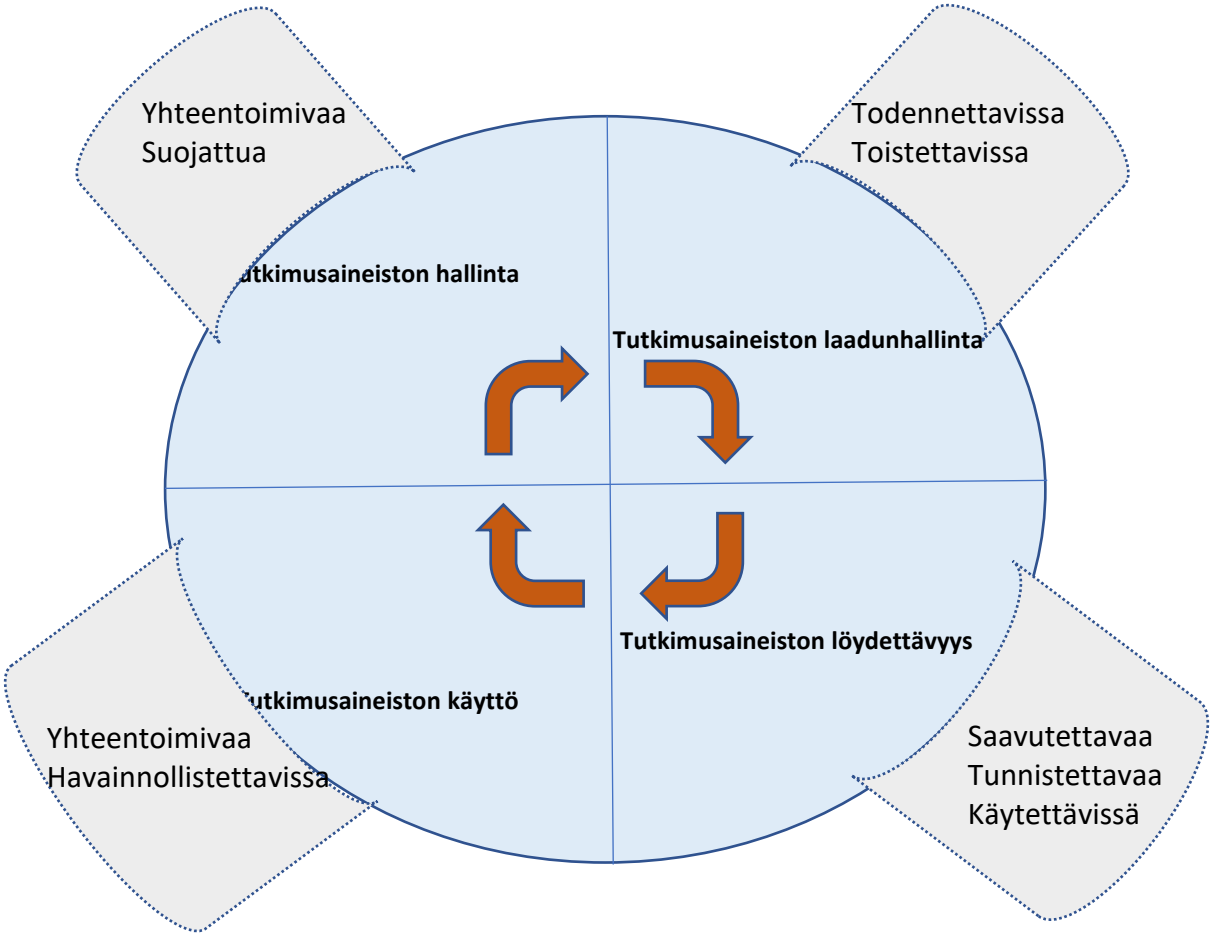
Ensimmäinen metatietostandardi tutkimusaineistolle julkaistiin vuonna 1993 IEEE Meta-Data -työpajan kokoontumisen yhteydessä. Yhdysvalloissa kokoontuneen työryhmän tavoite oli luoda käytänteet suurien aineistojen käyttöön ja aineistojen vaihtamiseen organisaatioiden välillä. Työryhmä tuli siihen tulokseen, että metatieto on tallennettu ja

[16] Esimerkiksi Kansallinen metatietovaranto Melinda käyttää Marc21-standardin tasoa 4 ks. https://www.kiwi.fi/display/melinda/Melindan+kuvailutasot

rakenteinen tietokokonaisuus jostakin kohteesta. Metatietojen tulisi sisältää esimerkiksi säilytykseen, hallinnolliseen tietoon kuten käyttöoikeuksiin liittyvää tietoa sekä tietoa kohteen alkuperästä. (Lei Zeng & Qin 2016, 429-430.)

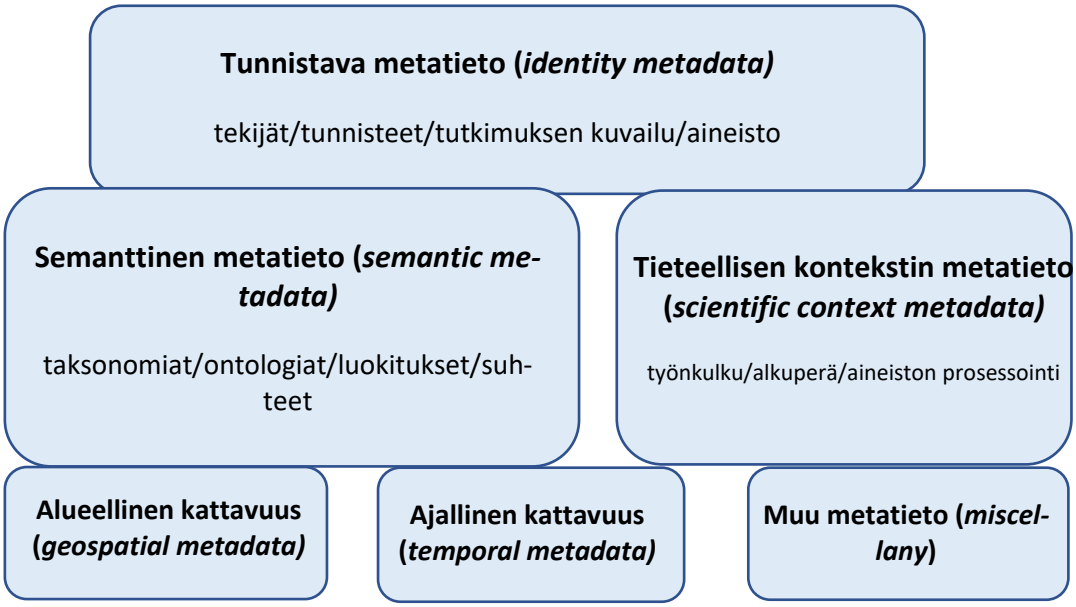
Tutkimusaineiston metatietovaatimukset metatietostandardeille voidaan jakaa neljään osaan, 1) varsinaiseen tutkimusaineiston hallintaan, 2) tutkimusaineiston laadunhallintaan, 3) tutkimusaineiston löydettävyyteen ja 4) tutkimusaineistojen jatkokäyttöön. Tutkimusaineistojen tulisi aina sisältää metatieto alkuperästä, jolloin se voidaan aina yhdistää alkuperäiseen tekijään. Metatietostandardin tulisi mahdollistaa asianmukaisten metatietojen kuvaileminen tutkimusaineistosta. Metatietojen pitäisi myös olla koneluettavia. Tutkimusaineistoa kuvailtaessa metatiedolla pyritään huolehtimaan koko tutkimusaineiston elinkaaresta niin, että tutkimusaineiston käytettävyys, saavutettavuus ja yhteiskäyttöisyys säilyvät koko elinkaaren ajan. (Kuva 1).

Kuva 1. Metatietovaatimukset standardeille tutkimusaineistojen kontekstissa (Lei Zeng & Qin 2016, 430; Qin, Ball & Greenberg 2012, 65).



Tutkimusaineistoissa käytetty metatieto voidaan jakaa edelleen kuuteen kategoriaan[17] niiden käyttötarkoituksen perusteella. Kategoriat ja niiden avulla tuotettu metatieto ovat keskeinen osa tieteellisen tutkimusaineiston kuvausta. Tutkimusaineiston metatietojen jaottelu pienempiin kategorioihin helpottaa kuvailemista ja auttaa hahmottamaan sitä, millä tavalla tutkimusaineistoja kuvaillaan (Kuva 2.).

[17] Qin, Ball & Greenberg (2012) käyttävät kategorioista ensisijaisesti nimitystä *Architechtural view,* joka kääntyy karkeasti suomeksi arkkitehtoninen näkymä. Kyseinen termi liittyy vahvasti arkkitehtuuriin, joten tässä työssä käytetään jaotteluun sopivampaa termiä kategoria.



Kuva 2. Tutkimusaineistojen metatietovaatimukset kategorioittain (Qin, Ball & Greenberg 2012, 65-67).

Jokainen kategoria voi sisältää entiteettejä, jotka taas sisältävät metatietoelementtejä varsinaista kohteen kuvausta varten. Käytännössä tämä tarkoittaa sitä, että esimerkiksi tutkimusaineiston tekijää voidaan kuvata metatietoelementeillä kuten nimellä, organisaatiolla, yhteystiedoilla ja yksilöivällä tunnisteella (esimerkiksi ORCID[18]). Tunnisteellisella metatiedolla voidaan myös yksilöidä kuvailtava tutkimusaineisto DOI:lla (*Digital Object Identifier)*[19]*, URN*:lla (*Uniform Resource Name)*[20] tai URI:lla (*Uniform Resource Identifier)*[21]. Semanttisella metatiedolla kuvataan tutkimusaineiston aihetta, jotta se voidaan linkittää muihin samansisältöisiin aineistoihin ja on helpommin käyttäjän löydettävissä. Tieteellisellä kontekstilla, johon voidaan sisällyttää myös alueellinen ja ajallinen

[18] ORCID on kansainvälinen yksilöivä tutkijatunniste, ks. https://orcid.org/

[19] DOI on yksilöivä tunniste esimerkiksi asiakirjoille, teksteille, audiovisuaalisille aineistoille ja ohjelmistoille, ks. https://www.doi.org/. Lisäksi DOI on ISO-standardisoitu ks. https://www.iso.org/standard/43506.html

[20] ” URN-tunnus on digitaalisen resurssin, kuten verkkojulkaisun, ainutkertainen, pysyvä standarditunnus.” ks. https://www.kansalliskirjasto.fi/fi/palvelut/tiedonkuvailun-asiantuntijapalvelut/urn-tunnukset

[21] https://www.w3.org/Addressing/

kattavuus, täytetään vaatimukset todennettavuudesta ja toistettavuudesta kuvaamalla tutkimusaineiston alkuperää ja työnkulkua aineiston keräyksen ja prosessoinnin suhteen (Kuva 2).

Jotta tutkimusaineiston metatietostandardien vaatimukset ja kategoriat saataisiin kontekstiin, ymmärrettäväksi ja käytännön tasolle, voidaan tutkimusaineiston elinkaariajattelun pohjalta erottaa kymmenen käyttäjän tarvetta tai tehtävää, jotka voidaan linkittää edellä esiteltyihin metatietovaatimuksiin ja luvussa 3.1 esiteltyihin metatietojen eri tyyppeihin (Taulukko 3). Tämän kokonaisuuden pohjalta on jo varsin helppo ymmärtää, mihin tutkimusaineistojen metatietostandardit – ja metatietostandardit ylipäätään - perustuvat. Kaiken lähtökohtana on aina käyttäjän kontekstisidonnainen tarve. Lisäksi huomionarvoista tämän tutkielman kannalta on se, että tämä kokonaisuus voidaan varsin helposti linkittää myös FAIR-periaatteisiin (ks. luku 2.1.3), jotka siis muodostuvat löydettävyydestä, saavutettavuudesta, yhteentoimivuudesta ja uudelleenkäytöstä.

Taulukko 3. Tutkimusaineiston käyttäjien tarve suhteessa metatietojen tyyppeihin ja vaatimuksiin (Qin, Ball & Greenberg 2012, 68-69).

Käyttäjän tarve	Metatietojen tyyppi	Metatietovaatimusten kategoriat
Löytäminen	Kuvaileva metatieto	Tunnistava ja semanttinen metatieto
Tunnistaminen	Kuvaileva metatieto	Tunnistava metatieto
Valitseminen	Kuvaileva metatieto ja tekninen metatieto	Tunnistava, semanttinen, tieteellisen kontekstin ja sekalainen metatieto
Hankkiminen	Kuvaileva metatieto	Tunnistava metatieto
Todentaminen	Kuvaileva metatieto	Tieteellisen kontekstin metatieto
Analysoiminen	Kuvaileva metatieto ja tekninen metatieto	Tieteellisen kontekstin, alueellisen ja ajallisen kattavuuden data
Hallitseminen	Kuvaileva, hallinnollinen, tekninen ja rakenteellinen metatieto	Tunnistava, semanttinen, tieteellisen kontekstin sekä alueellinen ja ajallinen metatieto
Arkistoiminen	Kuvaileva, hallinnollinen, tekninen ja rakenteellinen metatieto	Tunnistava, semanttinen, tieteellisen kontekstin sekä alueellinen ja ajallinen metatieto
Julkaiseminen	Kuvaileva metatieto	Tunnistava, semanttinen, tieteellisen kontekstin sekä alueellinen ja ajallinen metatieto
Viittaaminen	Kuvaileva metatieto	Tunnistava metatieto

Tutkimusaineistoille sopivia metatietostandardeja on yritetty kartoittaa yhtenäiseksi metatietoinfrastruktuuriksi[22] tutkimalla niiden yhteentoimivuutta keskenään. Qin ja Li (2013) tutkivat 16:ta metatietostandardia, joista he saivat 5800 metatietoelementtiä (eli kuvailtavaa metatietokenttää). Metatietoelementit jaoteltiin yhdeksään eri kategoriaan käyttötarkoituksen mukaan. Jaottelu oli hyvin pitkälti sama kuin metatietovaatimusten kategoriat Qin, Ball & Greenberg (2012, 65-67) mukaan. Suurin osa metatietoelementeistä jakautui paikkatiedon, hallinnollisen, tunnistavan, teknisen ja kuvailevan kategorian alle[23]. Tutkijat kiinnittivät huomiota samaa kokonaisuutta kuvaileviin kenttiin, joilla oli kuitenkin metatietostandardien välillä pieniä semanttisia eroja. Lisäksi tutkimuksessa kartoitettiin toistuvien elementtien määrää, joista kuvaus *(description)* ja otsikko (*title)* esiintyivät 10:ssä metatietostandardissa ja julkaisija (*publisher)*, viittaus (*citation)* ja maa (*country)* enää vain 8:ssa metatietostandardissa.

Esitelty Qin ja Lin (2013) tutkimus toi paljon uutta tietoa metatietostandardeista, niiden mahdollisesta yhteensopivuudesta ja muista yhteneväisyyksistä. Vastaavaa tutkimusta ei oltu aiemmin tehty tässä laajuudessa. Qin ja Li (mts). antavat kritiikkiä tieteenala- tai tutkimusaineistokohtaisten metatietostandardien tuottamisesta, koska tämä menetelmällä tuotetut metatietostandardit muodostuvat hankaliksi niin ihmisille kuin koneille. Suurin vaikutus voisi siis olla huono käytettävyys, joka vaikeuttaa tutkimusaineistojen

[22] Qin ja Lin (2013) mukaan metatietoinfrastruktuurilla tarkoitetaan kokonaisuutta, johon koottu metadataelementit, sanastot, entiteetit ja muut metatietojen liittyvät artefaktit sekä näitä tukevat työkalut ja sovellukset.

[23] Ylivoimaisesti eniten metadataelementtejä sijoittui kontekstia kuvaavaan kategoriaan, joka selittyi yhdellä, hyvin yksityiskohtaisella metadatastandardilla, jossa tutkimusaineiston kuvailua tehtiin muuttujatasolla, ts. jokainen tutkimuksessa käytetty muuttuja kuvaillaan, jolloin se aiheutti Qin ja Lin mukaan vääristymää tutkimuksen tuloksiin.

kuvailua. Toisaalta Qin ja Lin (mts). myöntävät, että metatietoinfrastruktuurin kehittäminen vaatii paljon lisätutkimusta, mutta ovat toiveikkaita semanttisen webin[24] ja erilaisten tunnistavien identiteettistandardien kuten FOAF (*Friend of a Friend)*[25], ORCID ja DOI kehittymisestä.

Yhteenvetoa metatietostandardeista tutkimusaineistojen kuvailussa

Tällä hetkellä yhteiskäyttöistä, kaikille tutkimusaineistolle sopivaa metatietostandardia ei siis ole olemassa ja sellaista ei todennäköisesti tule. Useilla tieteenaloilla, yhteisöillä ja tutkimusorganisaatioilla on omat metatietostandardit tutkimusaineiston kuvailua varten ja ne on kehitetty vastaamaan kyseisen alan tutkimusaineistojen tarpeita. Esimerkiksi paikkatietoaineistolle on useampi ISO-standardi (*International Organization for Standardization)* [26] kuvailua varten ja humanistisille tieteenaloille DDI-formaatti[27] (*Data Documentation Initiative)* (Qin, Ball & Greenberg 2012; Lei Zeng & Qin 2016, 429-431).

Myöskään Qin ja Lin (2013) esittämästä metatietoinfrastruktuurista ei ole kovin montaa käytännön esimerkkiä. Mainittavia ja mahdollisesti metatietoinfrastruktuureiksi kutsuttavia kattavia palveluita ovat mm. Dendro ja Suomessa CSC:n toteuttama Fairdata-palvelu. Dendro-alustalla voi kuvailla tutkimusaineiston kontrolloidusti ja strukturoidusti, jonka jälkeen sen voi julkaista metatietoportaalissa kuten EUDAT-palvelussa[28] (Castro, Silva & Ribeiro 2014; Castro ym. 2015; Silva ym. 2016). Suomessa vastaavaa ja mahdollisesti metatietoinfrastruktuuriksi kutsuttavaa palvelua ylläpitää CSC. Palvelun nimi on

[24] " Semanttinen Web (engl. Semantic Web) on Internetin WWW-palvelun laajennus, jonka dokumentit on suunniteltu myös koneita silmällä pitäen." ks. http://www.webopas.net/semanttinen.html

[25] FOAFilla voidaan luoda koneluettavaa tietoa käyttäjistä ja kuvailla heidän aktiviteettejaan, esim. tieteellisten tuotoksien kontekstissa. ks. http://www.foaf-project.org/

[26] Paikkatietoaineistojen kuvailuun kehitetyt ISO-standardit: ISO 19115-2:2009; ISO 19115-1:2014; ISO 19115-3; ISO 19139:2007

[27] DDI-formaatti sopii erityisesti humanististen tutkimusaineistojen (esimerkiksi kyselyaineistot ja litteroidut haastatteluaineistot) kuvailuun ks. https://www.ddialliance.org/

[28] EUDAT (*European Data Infrastructure)* on datainfrastruktuurihanke, joka tarjoaa tutkimusorganisaatioille, tutkimusyhteisöille ja tutkijoille palveluja tutkimusaineiston hallintaa varten (ks. https://www.eudat.eu/eudat-cdi/about ja https://www.csc.fi/fi/-/eudat-datapalvelut-ja-kansainvalinen-tutkimusyhteistyo)

aiemmin tutkielmassa mainittu Fairdata-palvelu, joka kattaa tutkimusaineiston arkistoinnin, metatietojen kuvailutyökalun sekä julkaisu- ja hakupalvelun.

Erilaisia metatietostandardeja tutkimusaineistoille löytyy pääosin digitaalisten aineistojen kuvailemiseen ja kuratointiin erikoistuneiden yhteisöjen sekä avointa tiedettä ja erityisesti datan avoimuutta edistävien tahojen verkkosivuilta. Vuonna 2004 perustettu digitaaliseen kuratointiin ja tutkimusaineistonhallintaan erikoistunut Digital Curation Center (DCC) listaa sivullaan 11 erilaista metatietostandardia, jotka soveltuvat tutkimusaineistojen yleiseen kuvailuun. Lisäksi sivustolle on kuratoitu tieteenalakohtaisia metatietostandardeja, työkaluja ja käyttötapauksia (Digital Curation Centre, 2018). GoFAIR -hanke listaa myös 11 FAIR-periaatteiden mukaista metatietostandardia sivuillaan (GoFAIR, 2018). Research Data Alliance (RDA) on perustettu vuonna 2013 edistämään tieteellisen datan avoimuutta ja uudelleenkäyttöä. Myös RDA on listannut tieteenaloittain sekä yleisesti hyväksyttyjä metatietostandardeja (Research Data Alliance, 2017). Tietoarkiston aineistonhallinnan oppaassa mainitaan useampi metatietostandardi. Tietoarkisto käyttää tutkimusaineistojen kuvailussa DDI-metatietostandardia (Tietoarkisto 2015).

3.3 Metatietojen laatu

Metatietojen keskeinen tehtävä on kuvailla kohteen sisältö käyttäjälle ja se on yleensä ensimmäinen tieto, jonka kanssa käyttäjä on vuorovaikutuksessa eri hakujärjestelmissä. Hyvä metatietojen laatu on keskeinen osa kuvaillun kohteen hyödyllisyydessä, käytettävyydessä ja löydettävyydessä. Huono metatietojen laatu voi aiheuttaa sen, että käyttäjä ei ymmärrä kuvaillun kohteen käyttötarkoitusta tai pahimmassa tapauksessa ei löydä sitä lainkaan. Pienikin puute tai virheellinen tieto metatiedossa voi olla kuvaillun kohteen käyttötarkoituksen suhteen haitallinen. Huono metatietojen laatu voi johtua monista tekijöistä kuten inhimillisistä virheistä ja epäjohdonmukaisesti käytetyistä sanastoista, skeemoista ja standardeista. (Lei Zeng & Qin 2016, 317-319.)

Ongelma metatietojen laadussa ja sen tutkimisessa on, että laadukkaalle metatiedolle ei ole yhtenäistä määritelmää ja metatietojen laadukkuuden mittaamiseen ei ole yhtenäisiä indikaattoreita tai mittareita. Kirjallisuudessa on käyty keskustelua laadukkaasta metatiedosta ja sen laadusta, mutta konsensusta ei ole. (Moen ym. 1998; Bruce & Hillmann 2004; Ochoa & Duval 2006; Ochoa 2009a, 73-74; Palavitsinis 2013, 56-58.)

Yleisellä tasolla termi laatu voidaan määritellä tietyn kohteen sisäisten ominaisuuksien erinomaisuutena. Tämän määritelmän perusteella metatiedoilla olisi mitattavissa olevaa objektiivista laadukkuutta (Ochoa 2009a). Palavitsinis (2014, 56-58) määrittelee väitöskirjassaan metatietojen laadun ISO 11620 -standardin avulla muotoon ”metatietojen ominaisuuksien ja piirteiden kokonaisuus, jolla voidaan tyydyttää ilmaistut ja epäsuorat tarpeet”. Tämä määritelmä ei oleellisesti eroa Ochoan (mts). määritelmästä, sillä Palavitsinisin soveltama määritelmä koskee metatietojen sisäisiä ominaisuuksia.

Metatietoa ei voida kuitenkaan tarkastella pelkästään sen sisäisten ominaisuuksien kautta, joka tekee laadun määrittelystä haastavaa. Metatietojen laadun suhteen pitäisi ensisijaisesti ajatella kohdetta, käyttötarkoitusta ja ympäristöä. Tämä tekee metatietojen laadun täsmällisestä määrittelemisestä vaikeaa, koska metatiedot ovat hyvin paljon kohteen kontekstiin liittyvä asia (Tani ym. 2013; Ochoa 2014). Lisäksi metatietojen laatu on mahdollista määritellä joko kokoelma-, tietue- tai elementtitasolla erilaisten metatietojen laatumallien kategorioiden perusteella (Lei Zeng & Qin 2016, 322).

Metatietojen laatu on moniulotteinen käsite, jolle ei ole yksiselitteistä määritelmää vaan laatu on aina kontekstisidonnainen asia, jota voidaan mitata erilaisilla tarkkuuksilla (kokoelma, yksittäinen tietue tai elementti) ja erilaisilla laatumallien (ks. luku 3.4) indikaattoreilla kuten täydellisyys, tarkkuus ja johdonmukaisuus. Metatietojen laatuongelmia voi esiintyä koko sen elinkaaren ajan metatietoskeeman suunnittelusta varsinaiseen kuvailutyöhön. (Lei Zeng & Qui 2016, 319-324.)

Tutkimuskirjallisuudesta löytyy myös useampia metatietojen laatumalleja[29], joissa laadukkuuden sijaan kuvataan hyviä käytänteitä (Bruce & Hillmann, 2004). Useammat tutkijat ovat myös määritelleet omia laatumalleja laadun arvioimiseen (Ochoa & Duval 2006; Park 2009; Palavitsinis 2013, 62-66). Tutkijoiden kehittämät laatumallit eroavat toisistaan ja useat tutkimukset ovatkin yrittäneet tehdä kattavia yhteenvetoja luodakseen yhtenäisen laatumallin metatietojen laadun mittaukseen. Metatietojen laatumalleista ei ole myöskään saavutettu konsensusta, vaikka erilaisia malleja on tehty ainakin 90-luvulta asti (Ochoa & Duval 2006).

Koska metatietojen laatu on kontekstisidonnainen asia, on tarve määritellä metatietojen laatu myös tämän tutkielman näkökulmasta. Tämän tutkielman kohteena olevien tutkimusaineistojen pääasiallinen tarkoitus on uudelleenkäyttö tai aiemman tutkimusaineistolla toteutetun tutkimuksen tuloksen toistaminen. **Tämän näkökulman pohjalta, tässä tutkielmassa metatietojen laatu ymmärretään löydettävyyden, käytettävyyden ja hyödyllisyyden näkökulmasta [30].**

3.4 Metatietojen laatumallit

Yksi pioneerimaineen saanut laatumalli on kuvailtu Moenin ym. (1998) toteuttamassa tutkimuksessa. Tutkimuksessa Moen ja kumppanit (mts). kävivät kirjallisuutta läpi toteuttaakseen manuaalisen metatietojen laadun mittauksen Yhdysvalloissa 1990-luvulla toimineeseen GILS (*Government Information Locator Service[31])*-järjestelmään. Moen toteutti tutkimuksen manuaalisesti ja kehitti oman laatumallinsa sen mittaukseen.

[29] Kirjallisuudessa ”laatumalleista” käytetään englanninkielisiä nimityksiä kuten ”Metadata Quality Assurance Framework” tai ”Framework for Quality Assurance”. Ruotsalainen (2016) käyttää työssään laatumalli-suomennosta, joten asianmukaisena käännöksenä sitä käytetään myös tässä tutkielmassa.

[30] Englanniksi voidaan käyttää määritelmiä ”fitness for use” tai ”fitness for purpose”, jotka voidaan karkeasti kääntää tutkielmassa mainittuun muotoon tutkimusaineistojen näkökulmasta tarkasteltuna. ks. esim. (Palavitsinis 2013; Palavitsinis, Manouselis, Sanchez-Alonso 2014; Király & Büchler 2018)

[31]GILS eli *Government Information Locator Service* on hajautettu tiedon paikannusjärjestelmä (ks. Kuronen 2000).

Moenin malli koostui neljästä kategoriasta

- **täydellisyys** (*completeness)* perustuu kokonaisvaltaiseen kuvailuun, joka kattaa tietueen pakolliset tiedot kuten nimekkeen, tekijän, tunnisteen sekä käytettyjen metatietokenttien määrää
- **profiili** (*profile)* määrittelee tietueen tyypin, kuten tietueen kenttien määrään vaikuttavan formaatin
- **tarkkuus** (*accuracy*) tarkastelee muotoiluvirheitä ja oikeinkirjoitusta
- **huollettavuus** (*serviceability)* perustuu metatietojen ylläpidettävyyteen kuten tietueen metatietokenttien järjestys, tiedostomuodot, viittauksen lainsäädäntöön ja paikallisesti määriteltyihin metatietokenttiin.

Toisen keskeisen laatumallin työstivät Bruce ja Hillmann (2004), jotka pyrkivät määrittelemään seitsemän erilaista metatietojen laatuominaisuutta, jossa korostuu

- **täydellisyys** (*completeness)*, kuvailun kohde esitellään mahdollisimman kattavasti ja mahdollisimman montaa metatietokenttää käyttäen.
- **saavutettavuus** (*accessibility),* metatieto on luettavissa ja kaikkien ymmärrettävissä.
- **tarkkuus** (*accuracy),* metatietojen pitäisi olla faktuaalista, ne eivät saa sisältää esimerkiksi kirjoitusvirheitä ja kuvailun pitäisi perustua kontrolloituihin sanastoihin.
- **alkuperä** (*provenance),* tietoa metatietojen tuottajasta, luomishetkestä ja tehdyistä muutoksista.
- **loogisuus ja koherenssi** (*logical consistency and coherence)*, metatiedoissa käytetään johdonmukaisesti metatietostandardeja ja keskeisiä käsitteitä.
- **ajantasaisuus** (*timeliness),* metatietojen täytyy muuttua, jos kuvailtu kohde muuttuu.
- **odotuksenmukaisuus** (*conformance of expectations),* metatietojen tulee vastata metatietojen käyttäjien odotuksiin ja palvella käyttäjien tarpeita.

Kolmannen ja paljon viitatun laatumallin kehittivät Stvilia ym. (2004), jotka pyrkivät määrittelemään metatietojen laadun jakamalla laatuominaisuudet kolmeen eri kategoriaan. Pääkategoriat muodostuvat käsitteistä

- **sisäinen laatu** (*intrisic IQ)* kattaa esimerkiksi metatietojen tarkkuuden, täydellisyyden ja johdonmukaisuuden.
- **suhteellinen laatu** (*relational/contextual IQ)* perustuu metatietojen suhteelliseen käytettävyyteen ja kattaa esimerkiksi metatietojen kuvaavuuden, tunnistettavuuden, saavutettavuuden ja täsmällisyyden.
- **maineeseen perustuva laatu** (*reputational* IQ*)* perustuu metatietojen tuottajan auktoriteettiin (Stvilia 2004, Appendix A).

Stvilia ja muut lajittelivat vielä erilaiset laatuominaisuudet em. kategorioihin. Yhteensä laatuominaisuuksia on 19 kappaletta. (Stvilia 2004, Liite A). Vuonna 2007 Stvilia uudisti malliaan. Pääkategoriat pysyivät samoina, mutta laatuominaisuuksia tuli kolme kappaletta lisää (Stvilia ym. 2007).

Europeanan[32] tuottamassa raportissa (Europeana 2015) pyrittiin kehittämään ja määrittelemään metatietojen laatua erityisesti Europeanaan tallennetuissa kohteissa. Europeanan työryhmä (Europeana 2015, 6) koosti raportissaan laadukkaan metatietojen keskeiset ominaisuudet. Työryhmä koosti seitsemän ominaisuutta, joihin laadukas metatieto perustuu. Nämä ominaisuudet olivat

- **luotettava prosessi** (*trusted process)* perustuu monivaiheeseen laadunvarmistukseen metatietojen tuottajan, välittäjän, kerääjän sekä lopulta Europeanan toimesta.

[32] Europeana on Euroopan digitaalisen kulttuurin ja taiteen keskus. ks. https://www.europeana.eu/portal/en

- **löydettävyys** (*findable)* perustuu saavutettavuuteen ja käytettävyyteen, jotka muodostuvat relevanteista asiasanoista ja uniikeista tiedoista ja tunnisteista.
- **luettavuus** (*human and machine readable)* tarkoittaa sitä, että metatieto on sekä ihmisen että koneen luettavissa ja ymmärrettävissä.
- **standarsoitu** (*standardised)* tarkoittaa metatietostandardien ja kontrolloitujen asiasanastojen käyttöä.
- **tarkoituksenmukaisuus** (*meaningful to users)* perustuu metatietojen tuottamiseen käyttäjälähtöisesti ja käyttäjien odotuksiin ja tarpeisiin peilaten.
- **uudelleenkäyttö** (*clear on re-use*) kerrotaan käyttäjille kohteen käyttöoikeuksista.

Yhteenveto laatumalleista

Näistä malleista kaksi (Bruce & Hillmann 2004; Stvilia ym. 2004) on tehty teoreettisessa mielessä metatietojen laadun määrittelyyn ja mittaukseen, ja kaksi muuta (Moen ym. 1998; Europeana 2015) tietyn organisaation tietovarantojen metatietojen laadun mittauksen. Kaikilla malleilla on yhteneväisyyksiä ja pieniä eroja, mikä on linjassa aiemmin kirjallisuudessa esille tulleen metatietojen kontekstisidonnaisuudesta. Esitellyistä malleista Brucen ja Hillmannin (mts). malli on saavuttanut suuren suosion laatumallien perustana ja heidän mallinsa on toiminut useiden metriikoiden pohjana (kts. mm. Ochoa & Duval 2009; Palavitsinis 2013) (Taulukko 4).

Taulukko 4. Yhteenveto laatumalleista

Moen ym. 1998	
Kategoria	**Ominaisuudet**
Täydellisyys	nimeke, tekijä, tunniste
Profiili	tietueen tyyppi, kenttien määrä
Tarkkuus	oikeinkirjoitus, sanastot
Huollettavuus	ylläpidettävyys, struktuuri

Bruce & Hillmann 2004	
Kategoria	**Ominaisuudet**
Täydellisyys	kattavuus
Saavutettavuus	käytettävyys, ymmärrettävyys
Tarkkuus	oikeinkirjoitus, sanastot
Alkuperä	tekijä, luomispäivämäärä
Loogisuus ja koherenssi	käytettävät standardit, käsitteet
Ajantasaisuus	ajantasainen tieto
Odotuksenmukaisuus	käyttökonteksti

Stvilia ym. 2004	
Kategoria	**Ominaisuudet**
Sisäinen laatu	tarkkuus, täydellisyys, johdonmukaisuus
Suhteellinen laatu	käytettävyys, kuvaavuus tunnistettavuus, täsmällisyys
Maineeseen perustuva laatu	tuottajan auktoriteetti

Europeana 2015	
Kategoria	**Ominaisuudet**
Luotettava prosessi	laadunvarmistus: tuottaja, käyttäjä, kerääjä ja Europeana
Löydettävyys	saavutettavuus, käytettävyys
Luettavuus	koneen ja ihmisen luettavissa
Standardisoitu	standardit, kontrolloidut sanastot
Tarkoituksenmukaisuus	käyttäjälähtöisyys, konteksti
Näkyvyys	löydettävissä

Lisäksi huomionarvoista on se, että Moenin ym. (1998) ja Europeanan (2015) mallien välillä ei löytynyt sinänsä merkittäviä eroja, vaikka niiden julkaisemisen välillä on kulunut 17 vuotta. Mittaustavat (automaattinen vs. manuaalinen) eroavat toisistaan ja kyseiset laatumallit on luotu aivan erilaisten järjestelmien tarkasteluun. Europeanan mallissa korostuu ehkä Moenin mallia enemmän koneluettavuus ja yhteiskäyttöisyys, joka voi johtua tietotekniikan kehittymisestä ja sitä myöden digitaalisten arkistojen yleistymisestä.

Laatumallien merkitys metatietojen laadun mittaamisessa on muodostunut hyvin tärkeäksi ja konkreettiseksi asiaksi. Tutkielmassa esitelty Brucen ja Hillmannin (2004) luoma laatumalli on luonut perustan hyvin monille metatietojen laatua kartoittavalle tutkimuksille.

4 AIEMPI TUTKIMUS

Neljännessä luvussa tehdään katsaus aiempaan tutkimukseen metatietojen laadusta. Ensimmäisessä alaluvussa tutustutaan manuaaliseen laadun mittaukseen. Toisessa ja kolmannessa alaluvussa tutustutaan automaattiseen mittaukseen ja neljännessä yhteenvetoon.

4.1 Metatietojen manuaalinen laadun mittaus

Yksi merkittävin metatietotutkimuksen alkuaikojen tutkimus on paikannusjärjestelmä GILSin metatietojen laatua analysoinut tutkimus, joka toteutettiin kaksivaiheisena Moen ym. (1998) toimesta (Ochoa 2014, 73). Ensimmäisessä vaiheessa kerättiin 80 metatietotietuetta mahdollisimman monipuolisesti ja johdonmukaisesti. Tietueet edustivat erilaisia tietolähteitä kuten tietokantoja tai luetteloita. Toisessa vaiheessa toteutettiin samoilla kriteereillä 83 tietueen kerääminen eri ajankohtana. Kerätyissä tietuissa oli yli 3500 metatietokenttää. Tiedot syötettiin manuaalisesti tietokoneelle ja tämän jälkeen tiedot analysoitiin (Moen ym. 1998).

Tutkijat etsivät kirjoitus- ja muotoiluvirheitä sekä vääriä päivämääräformaatteja. Nämä kuuluivat laatumallin tarkkuuden kategoriaan. Noin 10 % tietueista sisälsi ongelmia tarkkuudessa. Täydellisyyden suhteen tutkijat tarkastelivat pakollisen metatietojen esiintyvyyttä. GILS-järjestelmän tietomallissa[33] on 67 kenttää, joista osa on pakollisia ja osa vapaaehtoisia. Kerätyissä tietuissa käytettiin keskimäärin 42 kenttää. Tietueista 36 % sisälsi tyhjiä kenttiä, 96 %:ssa oli käytetty pakollisia kenttiä. Vain 12 % tietueista oli käytetty vapaaehtoisia, kontrolloituja sanastoja. (Moen ym. 1998.)

[33] Tietomalli on strukturoitu/rakenteistettu malli siitä, mitä tietoja halutaan kerätä ja mikä kerättyjen tietojen suhde toisiinsa. Esimerkiksi ks. https://www.ibm.com/support/knowledgecenter/en/SS9UM9_9.1.1/com.ibm.datatools.logical.ui.doc/topics/clogmod.html

Huollettavuuden kategoriassa tarkasteltiin metatietojen ylläpidettävyyttä. Tässä tutkimuksessa huollettavuuden tutkiminen perustui enemmän tutkijoiden henkilökohtaiseen arvosteluun ja tulkintaan kuin täydellisyydessä ja tarkkuudessa. Esimerkiksi jos nimeke on kuvailtu tutkijan mielestä huonosti, se on tulkittu puutteelliseksi. Huollettavuutta tutkiessa esiin nousivat huonosti kuvaillut nimekkeet, joita oli 25 %. (Moen ym. 1998.)

Metatietoja voidaan tutkia myös manuaalisen ja automaattisen tutkimusmenetelmän yhdistelmällä. Stvilia ym. (2005) tutkivat tieteellisten kirjastojen metatietoja sekä manuaalisesti että automaattisesti (ks. luvut 4.2 ja 4.3). Osana automaattista metatietojen laadun mittausta Stvilia ym. (mts). keräsivät manuaaliseen tarkasteluun 150 tietuetta. Yksikään tietueista ei sisältänyt Dublin Coren kahdeksaa keskeistä metatietokenttää. Suurimmassa osassa oli vain nimeke (*title),* tekijä (*creator),* tyyppi (*type)*, kieli (*language),* tunniste (*identifier)* ja lähde (*source).* Lisäksi päivämäärä (*date)* kenttä oli yhdistetty lähteen loppuun, jolloin se on ihmisen luettavissa, mutta ei koneluettavissa ilman ylimääräistä työtä.

4.2 Metatietojen automaattinen laadun mittaus kansainvälisesti

Automaattisella laadun mittauksella tarkoitetaan automatisoitua eli yleensä ohjelmallisesti (*programmatic)* tehtävää mittaamista. Metatietojen kontekstissa tämä tarkoittaa tietyssä arkistossa tai muussa vastaavassa kohteessa julkaistujen metatietojen systemaattista läpikäyntiä tietyillä ehdoilla. Käytännössä siis luodaan ehto sille, mikä on laadukasta ja sen jälkeen käydään automatisoidusti valittu kohde tai kohteet läpi. Yleensä ehdolla tarkoitetaan automaattisessa metatietojen laadun mittauksessa metriikkaa (*metrics)* tai metriikoiden yhdistelmää. Metatietojen laatua mittaavat metriikat pohjautuvat useissa tutkimuksissa luvussa 3.4 esiteltyihin laatumalleihin. (Ochoa & Duval 2009b.)

Stvilia (2004) tutki automatisoidusti 152 782 metatietotietuetta[34] eri metatietovarannosta sekä lisäksi 150 metatietotietuetta manuaalisesti (ks. luku 3.3). Metatiedot olivat kirjastojen, tieteellisten kirjastojen ja museoiden. Suurimmat ongelmat olivat puutteellisissa tiedoissa, turhassa metatiedossa, selkeyden puutteessa ja väärin käytetyissä Dublin Coren-skeeman elementeissä tai semanttisessa epäjohdonmukaisuudessa. Suurimmat yksittäiset ongelmat olivat päivämäärissä, joita oli annettu vaihtelevissa muodoissa.

Shreeves (2005) tutki tieteellisten kirjastojen metatietoja. Tutkimuksessaan hän analysoi yli 14000 metatietotietuetta sekä tieteellisen kirjaston ja yleisen kirjaston yhteistä metatietovarantoa, jossa metatietotietueita oli 1599. Tutkimuksen laadun mittaamisen Shreeves toteutti vertaamalla metatietokenttiä kahdeksaan Dublin Core-skeeman yleisimmin käytettyihin metatietokenttiin: nimeke, tekijä, aihe, kuvaus, päivämäärä, formaatti, tunniste ja käyttöoikeudet. Shreeves löysi aineistostaan epätäydellisiä metatietotietueita 71 % ja pienemmästä 1599 metatietotietueen kokoelmasta 43 %. Lisäksi Shreeves tutki suurta digitalisaatioprojektia, jossa metatietotietueita oli 27444 ja epätäydellisiä metatietotietueita löytyi 69 %.

National Science Digital Library (NSDL) –kirjaston metatietovarantoa tutkivat Bui ja Park (2006). Heidän toteuttamansa automaattinen keräysmenetelmä keräsi metatietoa lähes 18 kuukautta. Lopputuloksena oli yli miljoona metatietotietuetta, joita analysoitiin verrattuna Dublin Coren 15 metatietokentän standardiin. Yleisimmin käytettyjä olivat nimeke (99,92 % metatietotietueista), yksilöivä tunnus (99,35 %), päivämäärä (86,04 %), kuvaus (83,40 %) ja tekijät (83,34 %). Buin ja Parkin mukaan tärkeimmät kuusi Dublin Coren kenttää ovat kuvaus, aihe, nimeke, tunnus, tyyppi ja tekijä. Tämän perusteella Bui ja Park nostivat esille tekijän odotettua alhaisemman esiintyvyyden, joka oli 83,34 %.

[34] Metatietotietue on strukturoitu/rakenteistettu kokoelma metatietoja, joka yleensä noudattaa jotain metatietostandardia ks. https://www.sciencedirect.com/topics/computer-science/metadata-record

Myös arkistojen metatietojen laatua on tutkittu. Texas Archival Repository Online (TARO)[35] arkiston metatietojen laatua tutkittiin vuonna 2013, jolloin arkistossa oli 35 kokoelmaa, joissa oli yhteensä 8729 kuvailtua kohdetta XML-muodossa. Kuvailtujen kohteiden sisältöä tarkasteltiin verkkoselaimessa käytettävällä visuaalisella työkalulla (VADA, Visualizing Archival Data system). Arkistosta löytyneet laatuongelmat olivat pääasiassa puuttuvia tietoja ja epäjohdonmukaisuutta kokoelman sisällä. (Francisco-Revilla ym. 2014.)

Kokoelmaa kuvaava <repository> elementti puuttui kokonaan 305 metatietotietueesta viidessä eri kokoelmassa. Jos elementtiä oli käytetty, sen käytössä oli jopa saman kokoelman sisällä epäjohdonmukaisuutta. Käytännössä esimerkiksi oli käytetty vanhaa arkiston tai kokoelman nimeä. Samoin yksittäisten kohteiden kuvailuun tarvittavaa <dsc> elementtiä ei ollut kaikissa tietueissa. Se puuttui 230:stä metatietotietueesta. Kohteen fyysistä olemusta kuvaillaan <physdesc> tiedolla, jota oli käytetty tutkijoiden mukaan epäjohdonmukaisesti jokaisessa tietueessa. Käytännössä tämä tarkoittaa sitä, että kuvailussa ei ollut käytetty kontrolloituja standardeja. Epäjohdonmukaisuutta oli myös XML-tiedostojen hierarkiassa ja koodauksessa (*encoding).* (Francisco-Revilla ym. 2014.)

Kirjastojen metatietovarantojen ja arkistojen lisäksi myös opinnäytetöihin keskittyvien julkaisuarkistojen metatietoa on tutkittu (Thompson ym. 2019). Tutkijat latasivat Houstonin yliopiston DSPACE-pohjaisen[36] julkaisuarkiston metatietojen ja analysoivat ne[37]. Löytyneiden ongelmien lisäksi tutkijat esittivät korjausehdotukset tulevaisuuden varalle.

---

[35] TARO on Yhdysvaloissa toimiva arkisto, jota ylläpitää Teksasin yliopisto Austinissa. Arkiston nimi näyttää jälkeenpäin muuttuneen muotoon Texas Archival Resources Online, ks. https://legacy.lib.utexas.edu/taro/

[36] Dspace on avoimen lähdekoodin pohjalta toteutettu julkaisuarkistoalusta. Suomessa sitä käytetään useassa julkaisuarkistossa kuten ammattikorkeakoulujen yhteisessä julkaisuarkisto Theseuksessa (https://www.theseus.fi) ja Tampereen korkeakouluyhteisön Trepossa (https://trepo.tuni.fi) ks. lisää Dspacesta https://www.dspace.com/en/pub/home.cfm

[37] Tutkijat eivät valitettavasti kerro kuinka monta tietuetta he latasivat ja kuinka monesta löytyi virheitä vaan he käyvät löytyneet virheet läpi yleisellä tasolla.

Ongelmat löytyivät pääosin duplikaateista eli samaa kenttää oli käytetty useampaan kertaan samassa metatietotietueessa. Tunnistetietojen verkko-osoitteissa oli toimimattomia linkkejä, jotka eivät johtaneet mihinkään. Lisäksi sisällönkuvailussa oli käytetty epäjohdonmukaista ja standardoimatonta muotoa tekijän, tieteenalan tai koulutusohjelman nimessä. Myös päivämäärän kirjaaminen oli epäjohdonmukaista ja standardoimatonta. (mts.)

Dryad Digital Repositoryn[38] tutkimusaineistoja ja niiden metatietoa on tutkinut Rousidis ym. (2014a, 2014b). Tutkimusryhmä tutki pääasiassa kolmea metatietokenttää, jotka olivat tekijä (*creator),* päivämäärä (*date)* ja tyyppi (*type).* Metatietotietueita tutkittiin yli 16 000 kappaletta. Ongelmiksi nousivat tekijä-kentän kohdalla nimikirjainten tai pelkän etukirjaimen käyttö. Päivämäärä-kentässä käytettiin hyvin erilaisia formaatteja, mikä esimerkiksi saattoi estää hakutulosten järjestämisen päivämäärän mukaan. Rousidiksen (mts). tutkimuksen jatkotutkimuksissa Rousidis ym. (2015) ja Balatsoukas, Rousidis ja Garoufallou (2018), keskittyivät tutkimaan Dryad Repositoryn ”Aihe”-metatietokentän laatua. Suurimmiksi epäkohdiksi nousivat puutteelliset tiedot tekijä, päivämäärä ja tyyppi -metatietokentistä. Myöhemmin tutkitussa ”Aihe”-kentän tarkemmassa tarkastelussa ongelmaksi nousi se, että kenttää täyttäessä ei ollut käytetty kontrolloituja asiasanastoja. Silloin ongelmaksi tulee käytettyjen termien moninaisuus eikä kuvailu ole johdonmukaista ja yhteiskäyttöistä. Tämä laski heidän mukaansa metatietojen laatua.

Avoimen datan portaalit ovat keskeisessä asemassa tutkimusaineistojen ja muiden avointen aineistojen (kuten esimerkiksi valtioiden tuottaman avoimen aineiston) jakamisessa. Avoimen datan portaaleja tutkittiin vuonna 2016, kun Neumaier, Umbrich ja Polleres (2016) tutkivat 261 portaalin metatietoja. Portaalit sisälsivät pääosin valtioiden tuottamaa avointa dataa. Tutkijat loivat geneerisen mallin, jonka avulla tarkasteltiin eri

[38] Dryad Digital Repository on kuratoitu data-arkisto tutkimusjulkaisulle ja tutkimusaineistoille ks. https://datadryad.org/pages/organization

dataportaaleja riippumatta niiden julkaisualustoista tai metatietoformaateista. Käytännössä tutkijat kartoittivat[39] dataportaalien metatietoskeemat DCAT-yhteensopivaksi[40], jolloin metatietoja voitiin verrata yhteismitallisesti. Tutkimuksessa arvioitiin yli 2,1 miljoonaa metatietotietuetta.

Tutkijat testasivat ensin ovatko metatietotietueiden aineistot saatavilla. Noin 98 % aineistoista oli ladattavissa metatiedosta saadulla linkillä. Jäljelle jäänyt 2 % (8026) antoi tutkijoille virhekoodin HTTP 403, joka tarkoittaa, että tutkijoilta puuttui käyttöoikeus. Kyseiset aineistot saattavat tutkijoiden mukaan olla saatavilla vain pyynnöstä. Seuraavaksi tutkijat tarkastelivat metatietokenttien sisältöä. Tutkijoiden mielestä erittäin huolestuttavaa oli yhteystietojen ja lisenssitietojen puuttuminen noin puolesta metatietotietueita. Jos yhteystietoja ei ole saatavilla, aineiston uudelleenkäyttö vaikeutuu huomattavasti, koska alkuperäiseen tekijään ei välttämättä saada yhteyttä. Lisenssitietojen puuttuminen myös avoimen datan kontekstissa oli myös hyvin huolestuttavaa, koska se on aineiston uudelleenkäytön kannalta keskeisin vaatimus. Esimerkiksi CC-BY-4.0 -lisenssillä[41] julkaistuja aineistoja oli vain 6,6 %. Lisäksi tutkijoita huolestutti se, että aineistoa oli julkaistu eniten PDF-muodossa (23 %), joka ei ole koneluettavaa ja varsinaisesti avoimeen aineistoon sopiva formaatti. (Neumaier, Umbrich & Polleres 2016.)

[39]Vastaa englanninkielistä sanaa mapping: ”mapping **vastaavuus** (yt). Tarkoittaa sitä, että todellisuus ja sen säädin ovat esimerkiksi ”samassa järjestyksessä”, noudattavat samaa hierarkiaa tai niistä vain löytyy vastaavat elementit.” (MOT IT-Ensyklopedia). Käytännössä Neumaier, Umbrich ja Polleres järjestivät siis metatietotietueet DCAT-hiearkian mukaiseksi.

[40] ”DCAT on W3C:n laatima formaatti tai tarkemmin ”Data Catalog Vocabulary”, joka on tarkoitettu kaikentyyppisten data-aineistojen kuvailuun. Se on paljon yksinkertaisempi väline kuin DDI, mutta DCAT:in tavoitteena ei ole niinkään yksityiskohtaisen vaan yhteismitallisen kuvailun mahdollistaminen.” ks. https://tietolinja.kansalliskirjasto.fi/2016-3/1603-rda2/

[41] CC-BY-4.0 eli ”Nimeä 4.0 Kansainvälinen” tarkoittaa, että aineistoa voi vapaasti jakaa, kopioida ja levittää missä tahansa välineessä ja muodossa sekä remiksata ja muokata aineistoja sekä luodan sen pohjalta uusia aineistoja missä tahansa tarkoituksessa, myös kaupallisesti ks. https://creativecommons.org/licenses/by/4.0/deed.fi

4.3 Metatietojen automaattinen laadun mittaus Suomessa

Suomessa metatietojen laatua on tutkittu varsin vähän. Viimeisin tutkimus on Ruotsalaisen (2016) toteuttama tutkimus, jossa hän tutki julkaisuarkistojen metatietojen laatua. Ruotsalainen kävi läpi 14 julkaisuarkistoa ja keräsi 209 407 metatietotietueen aineiston. Metatietokenttiä verrattiin Kansalliseen metatietoformaattiin opinnäytetöille, joka on opinnäytetyön kuvailussa käytetty standardi. Aineiston analyysissa käytettiin täydellisyyden ja painotetun täydellisyyden mittareita.

Täydellisyyden metriikassa verrattiin aineiston metatietokenttiä suoraan Kansalliseen standardiin. Painotetun täydellisyyden metriikassa verrattiin tiedonhaun kannalta olennaisia kenttiä kuten tekijä, nimeke, aihe ja kuvailu Kansalliseen standardiin. Tutkielma hyödynsi Ochoan ja Duvalin (2009b) artikkelissa esitettyjä metriikoita. Tutkielman tuloksissa Ruotsalainen esittää korkeakoulukohtaisen yhteenvedon laatumetriikoiden pohjalta.

4.4 Yhteenvetoa metatietojen laadunarvioinnin tutkimuksista

Tässä tutkielmassa esiteltiin 10 erilaista metatietojen laadunarvioinnin tutkimusta (Taulukko 5.). Jokaisesta tutkimuksesta löytyi heikkolaatuista metatietoa. Tutkimuksien tyyppeinä olivat manuaaliset ja automaattiset tutkimukset. Tutkimukset keskittyivät paikannusjärjestelmiin, kirjaston metatietovarantoihin, perinteisiin arkistoihin, opinnäytetöiden julkaisuarkistoihin ja tämän tutkielman kannalta relevantteihin tutkimusaineistojen tallennuspalveluihin. Esiteltyjen tutkimusten ajallinen kattavuus oli hieman yli 20 vuotta. Tämä aikaväli kattaa hyvin pitkälti metatietotutkimuksen koko elinkaaren, sillä esimerkiksi Lei Zeng ja Qin (2016, 379-384) jakavat metatietotutkimuksen neljään vaiheeseen, joista ensimmäinen oli 1995 – 1998, toinen 1998 – 2003, kolmas 2003 – 2008 ja neljäs 2008 -.

Taulukko 5. Yhteenveto metatietojen laadunarvioinnin tutkimuksista

Tutkimus	Tutkimus-menestelmä	Tutkimuksen kohde	Tutkittu määrä	Löytyneet ongelmat
Moen ym. (1998)	Manuaalinen	GILS - Paikannusjärjestelmä	163 tietuetta	puuttuvat tiedot, kirjoitusvirheitä, väärät päivämääräformaatit
Shreeves ym. (2005)	Manuaalinen	Tieteellisten kirjojen metatietovaranto	35 tietuetta	puuttuvat tiedot, väärät päivämääräformaatit
Stvilia (2004)	Manuaalinen+Automaattinen	Muistiorganisaatioiden metatietovaranto	35 tietuetta manuaalisesti, 16000 tietuetta automaattisesti	puuttuvat tiedot, semanttinen epäjohdonmukaisuus, väärät päivämääräformaatit
Bui & Park (2006)	Automaattinen	NSDL-kirjaston metatietovaranto	> 1 000 000 tietuetta	puuttuvat tiedot
Francisco-Revilla ym. (2014)	Automaattinen	Arkisto	8729 tietuetta	puuttuvia tietoja, epäjohdonmukaisuutta, metatietoelementtien väärinkäyttö
Thompson ym. (2019)	Automaattinen	Opinnäytetöiden julkaisuarkisto	-	duplikaatit, tunnistetietojen ongelmat, epäjohdonmukaisuus kuvailussa, väärät päivämääräformaatit
Ruotsalainen (2016)	Automaattinen	Opinnäytetöiden julkaisuarkistot	209 407 tietuetta	puuttuvat tiedot, metatietoelementtien väärinkäyttö
Rousidis ym. (2014a, 2014b)	Automaattinen	Data-arkisto tutkimusjulkaisuille ja tutkimusaineistoille	16567 tietuetta	metatietoelementtien väärinkäyttö, väärät päivämääräformaatit
Rousidis ym. (2015), Balatsouskas ym. (2018)	Automaattinen	Aihe-kenttä Dryad-palvelussa	21809 tietuetta	puuttuvat tiedot, ei käytetty kontrolloituja asiasanastoja
Neumair, Umbrich ja Polleres (2016)	Automaattinen	Avoimen datan portaalit	261 portaalia	puuttuvat tiedot, formaatit, jotka eivät tue koneluettavuutta

Yhteenvetona esitellyistä tutkimuksista voidaan todeta, että tutkimusten toteutukset ovat linjassa metatietojen laatua ja laatumalleja (ks. luvut 3.3 ja 3.4) kuvaavan kirjallisuuden kanssa. Jokaisella tutkimuksella oli oma kontekstisidonnainen tutkimusote tutkimuksen toteuttamiseen. Esimerkiksi perinteistä arkistoa tutkitaan omien standardien kautta ja paikannusjärjestelmää sisäisen käytön näkökulmasta, kun taas tutkimusaineistojen laadunmittauksessa tärkeää oli toteuttaa tutkimus löydettävyyden, uudelleenkäytettävyyden ja yhteentoimivuuden näkökulmasta. Yhtenäisenä tekijänä tutkimuksissa toistui myös tavoite parantaa metatietojen laatua, joka taas nostaa kuvailtujen kohteiden käytettävyyttä riippumatta tutkimuskohteesta (Taulukko 5.).

5 TUTKIMUSASETELMA

Tämä tutkielma on empiirinen tutkimus, jonka tutkimusasetelma koostuu tutkimusongelmasta, tutkimusaineistosta ja tutkimusmenetelmästä.

Tutkimusongelma ja kysymykset esitellään ensimmäisessä alaluvussa. Seuraavassa alaluvussa esitellään tutkimusaineisto, jossa käydään läpi tutkimusaineiston keruuta ja prosessointia analysointia varten. Viimeisessä alaluvussa kerrotaan tutkimusmenetelmistä, ja käytetyistä metriikoista metatietojen laadunmittaukseen.

5.1 Tutkimusongelma

Tässä tutkielmassa käyään läpi avoimen tieteen ja tutkimuksen käsitettä ja sen vaikutusta tiedepolitiikkaan, tiedeyhteisöön ja tieteelliseen tutkimukseen (ks. Luku 2). Avoimen tieteen ja tutkimuksen johdosta muuttuvat tiedepoliittiset linjaukset ja rahoittajien vaatimukset vaikuttavat suoraan tieteellisen tutkimuksen käytänteisiin. Uudet tieteelliset käytänteet perustuvat entistä vastuullisempaan tieteeseen, jonka avainsanoja ovat saatavuus, luotettavuus, läpinäkyvyys ja toistettavuus. Taustalla vaikuttaa myös tekniikan nopea kehitys ja sen johdosta tapahtuva tietoyhteiskunnan kehittyminen.

Avoimen tieteen ja tutkimuksen aiheuttamat tiedepoliittiset linjaukset vaikuttavat myös Suomeen. Opetus- ja kulttuuriministeriö on muuttanut tiedepoliittisia linjauksiaan avoimempaan suuntaan. Samoin on tehnyt osa tutkimusrahoittajista kuten esimerkiksi Suomen Akatemia (luku 2.1).

Tämän tutkielman tutkimusongelman kannalta keskeisin muutos koskee tutkimusaineistoja ja niihin liittyvää aineistonhallintaa (luku 2.1.2). Uudistuneet tiedepoliittiset linjaukset edellyttävät rahoituksen vastineeksi huolellista tutkimusaineistojen dokumentoimista sekä tutkimusaineistojen avaamista kaikkien käytettäväksi. Metatietojen laatu on keskeisessä asemassa tutkimusaineistojen uudelleenkäytössä. Puutteelliset metatiedot voivat tehdä tutkimusaineiston tulkinnasta vaikeaa tai jopa mahdotonta (luku 3.3).

Tutkielman luvuissa 4.1 – 4.4 käytiin läpi metatietojen laatua mittaavia tutkimuksia, joissa jokaisessa ilmeni laatuongelmia tutkittavasta kohteesta, käytetyistä menetelmistä, metatietojen tuottajista tai tutkimuksien näkökulmista riippumatta. Esitellyt tutkimukset olivat kansainvälisiä ja kansallisia tutkimuksia.

Muuttuvien tiedepoliittisten linjauksien ja tutkielmassa esiteltyjen tutkimuksien pohjalta on aiheellista kartoittaa data-arkistoissa tai muissa vastaavissa tietovarannoissa julkaistujen tutkimusaineistojen metatietojen laatua ja selvittää mahdollisia ongelmakohtia metatietojen tuottamisessa.

Tässä tutkielmassa tutkimuskysymyksinä ovat:

1. **Millaista on tutkimusaineistojen metatietojen laatu automaattisille menetelmillä mitattuna?**
2. **Millaisia metatietojen laatuongelmia tutkimusaineistoissa ilmenee?**

Esitellyistä tutkimuksista sekä tutkielmassa aiemmin esitellystä kirjallisuudesta (luku 3.1 ja 3.4) käy myös ilmi metatietojen laatua kartoittavien tutkimusten kontekstisidonnaisuus.

5.2 Tutkimusaineisto

Tämän tutkielman tutkimusaineisto koostuu metatietotietueista. Tutkimusaineisto kerättiin data- ja metadata-arkistojen API-rajapinnoista. Tutkielmaan valitut data- ja metadata-arkistot ovat MetaX, The Dryad Digital Repository ja Zenodo. Työkaluina metatietojen lataamiseen, käsittelyyn ja analysointiin käytettiin Python-ohjelmointikieltä[42].

MetaX

MetaX on CSC:n tuottama tietovarantopalvelu, joka toimii Etsin-palvelun[43] metatietovarantona. MetaX-palvelu haravoi[44] tietoja myös Tietoarkistosta, Kielipankista ja SYKE-palvelusta. Tietomallit määrittävät MetaXiin syötettävän metatietotietueen rakenteen ja ne ovat verkossa vapaasti saatavilla[45].

Tutkimusaineisto hankittiin automaattisilla menetelmillä hyödyntäen MetaXin API-ohjelmointirajapintoja (*Application Programming Interface*)[46]. Kaikki MetaXin metatietotietueet ladattiin syksyllä 2019. Aineiston koko oli 9684 metatietotietuetta. MetaXista saadaan ladattua metatietotietue kahdessa tiedostomuodossa, XML tai JSON (*JavaScript Object Notation)*-muodossa[47]. Aineisto ladattiin JSON-muodossa (Kuva 3).

---

[42] Python on korkean tason tulkattava ohjelmointikieli, jonka kehitys on alkanut vuonna 1990. Kielen on kehittänyt Guido van Rossum. Ks. esim. https://docs.python.org/3/tutorial/index.html

[43] Etsin on tutkimusaineistojen hakupalvelu, joka sisältää tutkimusaineistojen metatietoja, ks. (https://www.fairdata.fi/etsin/)

[44] Tiedonharavointi tarkoittaa koneluettavissa tai ihmisen luettavissa olevan tiedon automatisoitua keräämistä (ks. http://www.tsk.fi/tsk/termitalkoot/hakemistot-267.html?page=get_id&id=ID331&vocabulary_code=TSKTT)

[45] Tietomallit ovat osoitteessa https://tietomallit.suomi.fi/model/mrd/CatalogRecord/ sekä Githubissa osoitteessa https://github.com/CSCfi/metax-api/blob/master/src/metax_api/api/rest/base/schemas/att_dataset_schema.json

[46] API on ohjelmointirajapinta, jonka avulla ohjelmat voivat vaihtaa tietoja ja keskustella keskenään. Ohjelmointirajapintojen avulla voidaan vaihtaa suuria tietomääriä helposti. Voidaan puhua siis käyttöliittymästä käyttäjän ja palveluntarjoajan välillä. API-esimerkkeinä voidaan mainita esimerkiksi REST-API ja OAI-PMH. (ks. https://www.webopedia.com/TERM/A/API.html ja http://avoinrajapinta.fi/)

MetaX-palvelun API-rajapinta on osoitteessa https://metax.fairdata.fi/docs/

[47] JSON on avoimesti standardoitu tiedostomuoto tiedonvälitykseen (ks. https://www.json.org/).

The Dryad Digital Repository

The Dryad Digital Repository on vuonna 2008 julkaistu kansainvälinen data-arkisto, joka on keskittynyt tutkimusaineistojen tallentamiseen ja jakamiseen[48]. Vuonna 2018 Dryad julkaisu lähes 5000 uutta tutkimusaineistoa[49]. The Dryad Digital Repository tarjoaa pääsyn metatietotietueisiin API-rajapinnan kautta. Saatavilla on uusittu rajapinta[50], joka tarjoaa tuen JSON-tiedostomuodolle. Vanhempi rajapinta perustui OAI-PMH-protokollaan[51] ja tarjosi tuen vain XML-tiedostomuodolle. Vanha rajapinta on edelleen saatavissa, mutta sitä ei enää päivitetä[52]. Aineisto ladattiin uudempaa API-rajapintaa hyödyntäen marraskuussa 2019. Ladatun aineiston koko oli 27 365 metatietotietuetta ja ne ladattiin JSON-muodossa.

Zenodo

Zenodo on vuonna 2013 julkaistu arkisto, joka syntyi osana The OpenAIRE-projektia Euroopan komission toimesta. Arkiston taustalla toimii CERN[53] ja rahoittajana toimii Euroopan Unioni[54]. Zenodo ei ole keskittynyt pelkästään tutkimusaineistoihin vaan siellä voi julkaista myös tieteellisiä julkaisuja, seminaariesityksiä ja ohjelmistoja. Zenodo tarjoaa aineistoihinsa REST- ja OAI-PMH pohjaiset API-rajapinnat, josta metatietotietueet voidaan automatisoidusti ladata. Tutkielman tekijän tekniset haasteet REST-rajapinnan toimivuuden kanssa ajoi käyttämään OAI-PMH rajapintaa, josta metatietotietueet ladattiin XML-muodossa (Kuva 4.). Metatietotietueita ladattiin 52 869 kappaletta.

---

[48] https://datadryad.org/stash
[49] Ks. vuosittaiset raportit https://github.com/CDL-Dryad/dryad/blob/master/public/docs/DryadAnnualReport2018.pdf
[50] ks. https://datadryad.org/api/v2/docs/
[51] ” OAI-PMH on yksinkertainen HTTP-pohjainen protokolla, joka ei ota kantaa esim. haravoitavan metadatan formaattiin. Ideana OAI-PMH:ssa on se, että ensimmäisellä kerralla haravoidaan kaikki tietueet repositorystä, ja jatkossa vain edellisen haravoinnin jälkeen tapahtuneet lisäykset, muutokset ja poistot.” ks. https://www.kiwi.fi/display/Finna/Finna+ja+OAI-PMH
[52] ks. http://v1.datadryad.org/oai
[53] CERN on hiukkasfysiikan tutkimuskeskus ks. https://home.cern/
[54] https://about.zenodo.org/

```
{
    "id": 10753,
    "identifier": "c5f3cafa-0f68-48a0-b3ce-82b4c01e3ed7",
    "data_catalog": {
        "id": 13,
        "identifier": "urn:nbn:fi:att:data-catalog-harvest-fsd"
    },
    "deprecated": false,
    "metadata_owner_org": "fairdata.fi",
    "metadata_provider_org": "fairdata.fi",
    "metadata_provider_user": "harvest@fairdata.fi",
    "research_dataset": {
        "title": {
            "en": "ISSP 2017: Social Networks and Social Resources III: Finnish Data",
            "fi": "ISSP 2017: sosiaaliset verkostot ja voimavarat III: Suomen aineisto"
        },
        "creator": [{
            "name": {
                "en": "International Social Survey Programme (ISSP)",
                "fi": "International Social Survey Programme (ISSP)"
            },
            "@type": "Organization"
        }, {
            "name": "Melin, Harri",
            "@type": "Person",
            "member_of": {
                "name": {
                    "en": "University of Tampere. School of Social Sciences and Humanities",
                    "fi": "Tampereen yliopisto. Yhteiskunta- ja kulttuuritieteiden yksikk\u00f6"
                },
                "@type": "Organization"
            }
        }],
        "keyword": ["sosiaaliset verkostot", "avuntarve", "yst\u00e4v\u00e4t", "luottamus", "huolenpito", "perhe", "sukulaiset", "tapaamiset", "ulkopuolisuus", "osallistumin
        "spatial": [{
            "place_uri": {
                "in_scheme": "http://www.yso.fi/onto/yso/places",
                "identifier": "http://www.yso.fi/onto/yso/p94426",
                "pref_label": {
                    "en": "Finland",
                    "fi": "Suomi",
                    "sv": "Finland",
                    "und": "Suomi"
                }
            },
            "geographic_name": "Finland"
        }],
        "language": [{
            "title": {
                "en": "Finnish language",
                "fi": "Suomen kieli",
                "sv": "finska",
                "und": "Suomen kieli"
            },
            "identifier": "http://lexvo.org/id/iso639-3/fin"
        }],
        "modified": "2018-03-06T00:00:00-00:00",
        "temporal": [{
            "end_date": "2017-12-31T23:59:59-00:00",
            "start_date": "2017-09-20T00:00:00-00:00"
        }],
        "publisher": {
            "name": {
                "en": "Finnish Social Science Data Archive",
```


Kuva 3. JSON-tiedostomuodossa oleva metatietotietue MetaXista

Kuva 4. Zenodon XML-tiedostomuodossa oleva metatietotietue

Tutkimusaineisto kokonaisuutena

Tutkimusaineisto koostui 89 918 metatietotietueesta. Metatietotietueista 37 049 oli JSON-muodossa ja 52 869 XML-muodossa.

Taulukko 6. Tutkimusaineisto kokonaisuutena.

Arkisto	**Metatietotietueita**	**Tiedostomuoto**
MetaX	9684	JSON
The Dryad Digital Repository	27 365	JSON
Zenodo	52 869	XML
Yhteensä	89 918	

Täydellisyyden ja painotetun täydellisyyden laatuarvojen lisäksi tarkasteltiin tilastollisia keskilukuja, joita olivat keskiarvo ja mediaani. Lisäksi tarkasteltiin tilastollisia tunnuslukuja, joita olivat pienin ja suurin laatuarvo sekä vaihteluväli.

5.3 Tutkimusmenetelmä

Tilastotiede on menetelmätiede, jolla tehdään päätelmiä empiirisistä havainnoista. Tilastotieteellä pyritään analysoimaan kerättyjä aineistoja ja tekemään niiden pohjalta johtopäätökset (Metsämuuronen 2002a, 6; Nummenmaa, Holopainen & Pulkkinen 9-10, 2014). Tilastollisten menetelmien avulla voidaan erottaa systemaattiset ja satunnaiset tekijät sekä arvioida näiden välisiä yhteyksiä. Tilastotiedettä voidaan pitää apuvälineenä oman aineiston tulkitsemiseen (Metsämuuronen 2002a, 6).

Tilastollinen tutkimus eli kvantitatiivinen tutkimus perustuu tilastotieteeseen, numeroihin ja niiden tarkasteluun matemaattisten toimenpiteiden avulla. Tilastollisen tutkimuksen aineisto koostuu havaintoyksiköistä, joista pyritään löytämään johdonmukaisuuksia. (Heikkilä 2008, 13). Havaintoaineistosta voidaan tehdä tilastollisen päättelyn avulla johtopäätöksiä. Tilastollisen päättelyn johtopäätösten mielekkyys pitäisi kuitenkin aina tulkita tutkimuskysymyksen kontekstissa. Tilastollisissa menetelmissä tutkimuskysymyksen empiiriseen ja teoreettiseen tarkasteluun sekä tulkintaan tarvitaan mittaustuloksia (Ketokivi 2015, 17-18, 96-97).

Tilastollinen tutkimus voidaan jakaa kokonaistutkimukseen ja otantatutkimukseen. Kokonaistutkimuksella tarkastellaan koko perusjoukkoa eli jokaista sen jäsentä. Otantatutkimuksella tutkitaan vain osaa perusjoukosta eli otosta. Otantatutkimus sopii tutkimustyypiksi, jos koko perusjoukkoa ei ole mahdollista tutkia. Esimerkiksi Suomessa tehdään paljon tutkimuksia puolueiden kannatuksista, joissa tarkastellaan tietyn perusjoukon osan äänestyskäyttäytymistä. Kokonaistutkimus olisi lähes mahdoton toteuttaa edes Suomen kokoiselle populaatiolle. (Nummenmaa, Holopainen & Pulkkinen 2014, 24-28.)

Tilastollista tutkimusta suunniteltaessa on määriteltävä mikä joukko on tutkimuksen kohteena. Määritelty joukko on perusjoukko ja se muodostuu tilastoyksiköistä, joita kerätään mittaamalla (Holopainen & Pulkkinen 2008, 15). Mittaamisella tarkoitetaan mittasymbolin tai mittaluvun liittämistä havaintoihin jonkin säännöstön mukaisesti. Mittaamisen avulla tilastoyksiöihin voidaan liittää havaintoja eli arvoja (Holopainen & Pulkkinen 2008, 15-17; Nummenmaa, Holopainen & Pulkkinen 2014, 17-18). Mittaaminen tehdään tutkimukselle sopivalla mittarilla. Tutkimuksessa voidaan käyttää valmista mittaria tai luoda mittari itse. Tutkimusten tulosten luotettavuus on riippuvainen mittarista, joten se kannattaa suunnitella huolella tai käyttää valmista mittaria, jonka luotettavuus on tutkittu (Metsämuuronen 2002a, 10-11).

Mittauksen ja mittarin luotettavuutta voidaan kuvata reliabiliteetilla ja validiteetilla. Reliabiliteetilla tarkoitetaan yleensä luotettavuutta eli onko tutkimus esimerkiksi toistettavissa (Metsämuuronen 2002a, 11; Nummenmaa, Holopainen & Pulkkinen 2014, 20). Reliabiliteettia arvioidessa ei siis oteta kantaa siihen mitä mitataan vaan siihen, voidaanko mittaustulos toistaa (Ketokivi 2015, 100). Validiteetilla tarkoitetaan pätevyyttä eli onko mittauksen tulos se, mitä oli tarkoitus mitata (Metsämuuronen 2002, 11; Nummenmaa, Holopainen & Pulkkinen 2014, 20). Validiteettia voidaan pitää tutkimuksen kannalta jopa tärkeämpänä kuin reliabiliteettia, koska sen avulla mitataan sitä, mitä halutaan (Ketokivi 2015, 104). Lisäksi Ketokivi (2015, 104) toteaa ”Mittari on sisällöllisesti validi (*content valid*) silloin, kun tutkija on uskottavasti perustellut indikaattorien kattavan käsitteen teoreettista sisältöä riittävästi.”

Haluttua tietoa mittaavan mittarin rakentaminen perustuu tutkimuksen teoriaan, tutkittavaan ilmiöön ja tutkimuskysymykseen. Teorian ja ilmiön käsittelyn pohjalta muodostuneet käsitteet ja niiden pohjalta tuotetut operationalisoinnit[55] toimivat mittarin perustana. Operationalisoinnin avulla tutkimukselle saadaan mitattavissa oleva määritelmä. Tutkimuksen validiteetin kannalta onkin elintärkeää, että operationalisointi onnistuu ja tutkimukselle onnistutaan valitsemaan relevantit mittauskohteet. (Metsämuuronen 2002b, 22-23, 25; Ketokivi 2015, 95-98.)

5.4 Metatietojen laadun mittaus tilastollisena tutkimuksena

Tutkielman tutkimusongelma, tutkimuskysymykset ja niistä muodostuneet tavoitteet määrittivät käytettävän tutkimusmenetelmän. Tässä tutkielmassa metatietojen laatua mitattiin automaattisilla menetelmillä, jotka perustuivat tilastollisiin menetelmiin. Tutkimusmenetelmä perustui tilastolliseen analyysiin, jossa sovellettiin tämän tutkielman näkökulmasta tehtyä metatietojen laadun määritelmää (ks. luku 3.3), metatietojen laatumalleja (ks. luku 3.4) ja niistä johdettuja metriikoita, jotka toimivat tutkielman mittareina.

Tutkielma tutkimuksellinen osuus toteutettiin harkinnanvaraisena näytteenä data- ja metatietoarkistoista. Tämän tutkielman tilastoyksiköitä ovat eri arkistojen tutkimusaineistoja kuvaavat metatietotietueet. Näistä tilastoyksiköistä muodostuu perusjoukko. metriikoiden ja vertailumittarin avulla metatietotietueiden laatu voidaan muuttaa numeraaliseen muotoon (ts. operationalisoidaan), jolloin niistä saadaan mitattavia suureita.

[55] ” Tieteellisessä tutkimuksessa tarvitaan teoreettisia käsitteitä ja niiden empiirisiä vastineita. Luonnontieteissä ja määrällisessä tutkimuksessa puhutaan operationalisoinnista, joka tarkoittaa teoreettisten käsitteiden muuttamista empiirisesti mitattavaan muotoon.” (ks. https://www.fsd.uta.fi/menetelmaopetus/kvali/L2_3_2_2.html)

Tutkimusaineiston analyysissa sovelletaan vertailumittaria (Taulukko 7) ja luvussa 5.4.1 esiteltyjä metriikoita. Käytännössä tämä toteutettiin niin, että ladatuista metatietotietueista etsittiin rivi kerrallaan, onko vertailumittarissa määriteltyä tietoa saatavilla vai ei. Mikäli tieto löytyi, paluuarvona saatiin luku 1 ja mikäli tietoa ei ollut, paluuarvona oli luku 0. Ainoana poikkeuksena olivat asiasanat, joihin sovelletaan asiasanojen määrään painottuvaa metriikkaa (ks. Kaava 3). Jokaisesta tutkimusaineiston metatietotietueesta kirjoitettiin rivi paluuarvoja ja ne tallennettiin arkistokohtaisena csv-tiedostona. Analyysin työvälineenä toimi Python-ohjelmointikieli. Pythonin data-analyysiin soveltuvista kirjastoista käytettiin erityisesti Pandas[56]- ja NumPy[57]-ohjelmakirjastoja. Tulosten visualisointiin käytettiin Matplotlib[58]-kirjastoa.

5.4.1 Tutkielman metriikat

Tässä tutkielmassa metriikkana käytettiin täydellisyyden (*completeness*) metriikkaa. Aiemmin tutkielmassa määritelty täydellisyys tarkoittaa metatietojen kontekstissa (luku 3.4) sitä, että ” kuvailun kohde esitellään mahdollisimman kattavasti ja mahdollisimman montaa metatietokenttää käyttäen.” (Bruce & Hillman 2004) ja ”kohteen metatiedot perustuvat kokonaisvaltaiseen kuvailuun, joka kattaa tietueen pakolliset tiedot kuten nimekkeen, tekijän, tunnisteen sekä käytettyjen metatietokenttien määrää.” (Moen ym. 1998). Yksinkertaisimmillaan täydellisyyden metriikalla vertaillaan metatietotietueiden kuvailtujen elementtien määrää määriteltyyn vertailumittariin, jolla saadaan mitattava lukema metatietojen kattavuudelle (Ochoa & Duval 2009b, 70-71; Margaritopoulous ym. 2012; Palavitsinis 2013, 66).

Täydellisyyden metriikka voidaan jakaa esimerkiksi

1. pakollisten elementtien,

[56] Pandas-kirjaston dokumentaatio ja esittely, ks. https://pandas.pydata.org/
[57] NumPy-kirjaston dokumentaatio ja esittely, ks. https://numpy.org/
[58] Matplotlib-kirjaston dokumentaatio ja esittely, ks. https://matplotlib.org/

2. ”suositeltujen” elementtien tai
3. valinnaisten/vapaaehtoisten elementtien tarkasteluun. (Gavrilis ym 2015).

Miksi tässä tutkielmassa käytetään täydellisyyden metriikkaa ja sen variaatioita metatietojen laadun mittaamiseen? Esimerkiksi Margaritopoulous ym. (2012) ja Park (2009) toteavat, että täydellisyyden metriikka on yksi keskeisimpiä ja yksinkertaisimpia tapoja mitata metatietojen laatua ja useat tutkimukset hyödyntävät sitä. Täydellisyyden metriikka on varsin suoraviivainen ja helposti ymmärrettävä sekä tuottaa mitattavia lukuja halutuista metatietotietueista erityisesti automaattisilla menetelmillä (Ochoa & Duval 2009b, 71; Margaritopoulous 2012 ym.) Lisäksi täydellisyyden metriikka taipuu Gavrilisia ym. (2015), Ochoaa ja Duvalia (2009b) ja Margaritopoulousia ym. (2012). mukaillen mm. painotettuun täydellisyyden metriikkaan, jossa annetaan suurempi painoarvo tietyille metatietoelementeille, joiden katsotaan olevan keskeisiä kuvaillulle kohteen käyttötarkoitukselle.

Täydellisyys toistuu useammassa tässäkin tutkielmassa esitellyssä metatietojen laatumallissa (ks. luku 3.4) sekä automaattisissa laatua kartoittavissa tutkimuksissa (ks. luvut 4.2 ja 4.3). Jos metatietojen laatua mitataan täydellisyyden metriikalla, pitäisi metriikka ja sen kehys asettaa metatietotutkimuksille tyypillisesti kontekstiin eli huomioida ensisijaisesti kuvaillun kohteen käyttötarkoitus (Margaritopoulous ym. 2012; Lei Zeng & Qin 2016, 325-326). Jotta voitaisiin käyttää täydellisyyden metriikkaa, pitäisi olla jokin määritelty mittari, johon metatietotietueiden elementtejä voidaan vertailla eli vertailumittari. Vertailumittarinkin suhteen pitäisi ensisijaisesti huomioida kuvaillun kohteen käyttötarkoitusta, jotta metatietojen laatua voidaan mitata luotettavasti (Margaritopoulous ym. 2012).

Tämän tutkielman tutkimusongelma ja tutkimuskysymykset muodostuivat tutkimusaineistojen kontekstissa, joten käytettävän vertailumittarin pitäisi tukea sitä lähtökohtaa. Kuten tämän tutkielman luvuista 2.1.2, 3.1 ja 3.2 nähdään, tutkimusaineistojen dokumentoinnissa pitäisi huomioida koko tutkimusaineiston elinkaari, joka sisältää löydettävyyden, saavutettavuuden, käytettävyyden ja yhteiskäyttöisyyden.

Kuten metatietojen laatua mittaavasta tutkimuskirjallisuudesta nähdään (ks. Luku 4), täydellisyyden metriikkaa ja sen vertailumittari voidaan käyttää hyvin monella tapaa varioiden ja eri näkökulmista. Voidaan todeta, että täydellisyyden metriikka sopii joustavuutensa ja helppokäyttöisyytensä vuoksi myös tähän gradututkielmaan. Tässä tutkielmassa sovelletaan Ochoan & Duvalin (2009b) ja Margaritopouloksen (2012) kehittämiä täydellisyyden metriikoita. Työssä käytetään Ruotsalaisen (2016) tapaan yksinkertaista täydellisyyden metriikkaa ja painotettua täydellisyyden metriikkaa.

Täydellisen metriikka (Com) mitataan kaavalla (Kaava 1), jossa n kuvaa metatietokenttien määrää ja P(i) on *i*:s metatietokenttä, jossa ei ole tyhjää arvoa, jolloin kaava palauttaa P(*i*):lle arvon 1, muutoin arvon 0. Lopuksi $\sum P(i)$ jaetaan *n*:llä eli kenttien määrällä.

Kaava 1. Täydellisyyden metriikka (Ochoa & Duval 2009b).

$$Com = \sum_{i=1}^{n} P(i) \ / \ n$$

Yksinkertaistettu käytännön esimerkki täydellisyyden mittarista: jos oletetaan, että 4 metatietokenttää on täytetty kuvailutiedolla 10 kentän metatietotietueessa ja loput 6 kenttää ovat tyhjiä, on täydellisyyttä kuvaava luku silloin 4 / 10 eli 0,4 (tai 40 %). Jos taas kaikki metatietokentät ovat täytetty kuvailutiedolla, on täydellisyyttä kuvaava luku 1 ja jos metatietokentissä ei ole mitään kuvailua, on täydellisyyttä kuvaava luku 0.

Painotetun täydellisyyden (ComW) metriikka mitataan kaavalla (Kaava 2), jossa sovelletaan muuten täydellisen metriikan kaavaa (Kaava 1), mutta a_i määrittelee kentän painoarvon. Tämän tutkielman kontekstissa a määrittelee metatietokenttien suhteellisen painoarvon tutkimusaineistojen näkökulmasta.

Kaava 2. Painotetun täydellisyyden metriikka (Ochoa & Duval 2009b).

$$ComW = \sum_{i=1}^{n} (a_i * P(i)) \ / \ \sum_{i=1}^{n} a_i$$

Yksinkertaistettu käytännön esimerkki painotetun täydellisyyden mittarista: jos oletetaan, että 4 metatietokenttää, joiden painoarvoksi määritellään 1, on kuvailtu 10 kentän

metatietotetueessa ja loput 6 ovat tyhjiä ja painoarvoltaan 0,5, on painotettua täydellisyyttä kuvaava luku silloin 4 / 7 = 0,57 (tai 57 %). Jos taas kaikki metatietokentät ovat täytetty kuvailutiedolla, on täydellisyyttä kuvaava luku 1 ja jos metatietokentissä ei ole mitään kuvailua, on täydellisyyttä kuvaava luku 0.

Näiden kahden metriikan avulla voidaan mitata parhaiten sellaisia metatietokenttiä, joiden kardinaalisuus on 1[59]. Esimerkiksi tekijän nimi voi olla moniosainen, mutta se ymmärretään silti vain yhtenä metatietokentän alkiona. Sen sijaan esimerkiksi asiasanoista (*subject headings)* muodostuva kokonaisuus on moniarvoinen metatietokenttä. (Margaritopoulous ym. 2012). Saman ongelman havaitsivat Bui & Park (2006), jotka huomasivat, että Aihe-kentän sisältöä tulisi tarkastella monimuotoisemmin, koska samalla kentällä voi olla useampi alkio.

On siis perusteltua tarkastella moniarvoisia metatietokenttiä vielä omalla metriikallaan, jotta metatietotietueen laadusta saadaan tarkempi kuva. Esimerkiksi hakupalveluissa tehtävät haut kohdistuvat mm. metatietotietueiden asiasanoihin, jolloin suurempi asiasanojen määrä voi tehdä metatietotietueesta löydettävämmän.

Jos metatietokentän kardinaalinen arvo on suurempi kuin 1, voidaan sen täydellisyyttä mitata Margaritopoulouksen ja kumppaneiden (2012) kehittämällä kaavalla (Kaava 3)., jossa *f* kuvaa tutkittavaa kenttää, *m* alkioiden määrävaatimusta, a_i painotusta ja $P_i(f)$ lukua 1, mikäli kenttä *f* on olemassa *i*:ssä, ja lukua 0, mikäli indikaattoria *f* ei ole *i:*ssä.

Kaava 3. Metriikka moniarvoisten metatietokenttien mittaamiseen (Margaritopoulous ym. 2012).

$$Com(f) = \sum_{i=1}^{m(f)} (a_i * P_i(f))$$

[59] Kardinaaliteetti eli joukon mahtavuus. ” määre, joka ilmaisee, kuinka monta kertaa tietyn tietoelementin voi liittää tai kuinka monta kertaa se tulee liittää kohteeseen tai metatietoon” (ks. http://www.tsk.fi/tepa/fi/haku/kardinaliteetti)

Yksinkertaistettu käytännön esimerkki moniarvoisten metatietokenttien mittaamisesta: jos oletetaan, että asiasanoja (f = ”asiasana”) on kaksi kappaletta mutta vaatimus asiasanojen määrän suhteen on 3 (*m(f)* = 3) ja painoarvo (a_i) jokaiselle asiasanalle on sama, silloin saadaan $P_{1(f)=1}$, $P_{2(f)=1}$ ja $P_{3(f)=0}$, jolloin voidaan soveltaa ylläolevaa kaavaa (Kaava 3). ja laskea:

$Com(f) = a_1(f) * P_1(f) + a_2(f) * P_2(f) + a_3(f) * P_3(f)$ = 0.33*1 + 0.33*1 + 0.33*0 = 0.66

Tämän tutkielman kontekstissa *m(f)* eli vaatimus asiasanojen määrästä oli 5. Käytännössä tämä siis tarkoittaa sitä, että mikäli metatietotietueen asiasanakentässä oli 5 asiasanaa, oli täydellisyyden metriikan palauttama arvo 1. Jos kentässä oli vain 1 asiasana, oli täydellisyyden metriikan palauttama arvo 0,20. Jos asiasanoja ei ollut lainkaan, olikentän arvo 0. Painotetun täydellisyyden metriikassa pätee muutoin sama periaate, mutta painoarvon ollessa 0,5, oli metriikan tuottama vaihteluväli 0 – 0,5.

Vaatimus asiasanojen määrästä perustui kahden tutkimuksen keskiarvoihin asiasanojen määrästä. Tavoite oli saada karkea kuva tutkijoiden tuottamien asiasanojen määrästä erilaisissa tieteellisissä konteksteissa kuten tutkimusaineistojen tai tutkimusjulkaisujen kuvailu. Tutkimukset olivat aiemmin tutkielmassa esitelty Balatsoukas, Rousidis ja Garoufalloun (2018), jossa keskimääräinen arvio asiasanojen määrästä oli 4,79 per tutkimusaineiston metatietotietue ja tieteellisten julkaisujen asiasanamäärää tutkineet Babaii ja Taase (2013), jotka päätyivät keskiarvolukuun 4,8.

5.4.2 Tutkielman vertailumittari metriikoille

Jotta operationalisointi onnistuisi metriikoiden avulla, käytettiin tutkielmaa varten kehitettyä vertailumittaria. Tutkielmassa käytettiin itse kehitettyä vertailumittaria metatietojen ja tutkimusaineistojen vahvan kontekstisidonnaisuuden takia. Luvussa 3.2.1 kerrotaan metatietostandardien moninaisuudesta ja siitä, että mikään metatietostandardi ei ole vakiintunut tutkimusaineistojen kuvailussa. Dublin Core (ISO 15836-1:2017) on todennäköisesti suosituin kuvailustandardi, joten sitä hyödynnettiin myös tämän tutkielman vertailumittarin kehittämisessä. Vertailumittarin tarkoitus on kartoittaa pienin mahdollinen metatietojen tarve, jotta tutkimusaineisto olisi löydettävää, saavutettavaa

ja hyödyllistä. Tällä tarkoitetaan tunnistavaa metatietoa tutkimusaineistojen konteksissa. Semanttista tai tieteellisen kontekstin metatietoa ei tässä tutkielmassa arvioida (ks. Kuva 2).

Vertailumittarin luomisessa sovelletaan FAIR-periaatteita (Force11 2016), Dublin Core (ISO 15836-1:2017) -standardia ja tutkimusaineistojen kuvailuun luotua minimimetatietomallia (Tutkimuksen tietoaineistot 2013), metatietostandardien vaatimuksia (Qin, Ball Greenberg 2012, 65-67; Lei Zeng ja Qin 2016, 429-432; ks. myös Kuva 2) ja tutkimusaineistojen käyttäjien tarpeita (Qin, Ball ja Greenberg 2012, 68-69; ks. myös Taulukko 3) sekä luvussa 3.3 tehtyä laatumääritelmää tämän tutkielman kannalta: ”***Tässä tutkielmassa metatietojen laatu ymmärretään löydettävyyden, käytettävyyden ja hyödyllisyyden näkökulmasta*”.** Mittarin luomisen teoreettisena pohjana käytettiin Metsämuurosen (2002b, 22-30) ohjeistuksia mittarin luomisesta.

Näiden lähteiden pohjalta koottiin vertailumittaria kuvaava taulukko (Taulukko 7.). Vertailumittari koostui 9:stä parametrista. Näistä 9 parametrista käytetään tähdellä merkittyjä 4 parametria (Tekijä, Tekijän yksilöivä tunniste, Kuvaus ja Lisenssi) painotetussa täydellisyyden metriikassa painoarvolla 1 ja muita parametreja painoarvolla 0,5 (Taulukko 6.).

Esitellyistä parametreistä kaksi ei esiinny Dublin Coressa tai minimimetatietomallissa. Nämä ovat tutkimusaineiston jakelijan yksilöivä tunniste ja organisaatio. Huomionarvoista on se, että mikäli tutkimusaineiston tekijän yhteystietoja ei ole saatavilla tutkimusaineiston metatietotietueessa, on hyvin vaikea ottaa yhteyttä alkuperäiseen jakelijaan. Tämä saattaa olla elintärkeää, mikäli tutkimusaineistossa on jotain epäselvää. Tällainen tilanne voi pahimmillaan estää tutkimusaineiston analyysin. Minimimetatietomallissa todetaan: ”Yhteystietoja tarvitaan, jos aineiston uudelleenkäyttöön liittyy haasteita, joita käyttäjä ei pysty itse ratkaisemaan. Tilanteissa, joissa aineiston saatavuus, laatu tai ymmärrettävyys aiheuttavat käyttäjälle hankaluuksia, on yhteyshenkilöön turvautuminen usein ainoa vaihtoehto.” (Tutkimuksen tietoaineistot 2013). Suorien yhteystietojen jakaminen voi olla haitallista, joten yksilöivät tunnisteet kuten ORCID ja organisaatiotiedot voivat toimia ainakin välillisesti yhteystietoina tutkimusaineiston jakelijan

saavuttamiseksi[60]. Yksilöivän tunnisteen ja organisaatiotiedon muut edut ovat vaikuttavuuden ja näkyvyyden kasvaminen sekä tekijyyden määritteleminen (Nygård 2018).

[60] Suorien yhteystietojen jakaminen avoimesti on harvinaista. Sähköpostiosoitteen jakaminen avoimesti verkossa voi johtaa suureen määrään roskapostia, hakkerointiyrityksiin ja muihin vastaaviin haittoihin. Tämän johdosta osa data-arkistoista ei edes tue suorien yhteystietojen jakamista tietomalleissaan. Tässä tutkielmassa käytetyistä data- ja metadata-arkistoista MetaXiin voi antaa sähköpostin, mutta se ei tule julkisesti saataville vaan sivuston kautta voi lähettää viestin tutkimusaineiston tuottajalle. The Dryad Digital Repository-arkistossa voi jakaa sähköpostiosoitteensa suoraan. Zenodossa ei ole mahdollista antaa sähköpostiaan yhteystiedoksi vaan tietomallissa voi määritellä yhteyshenkilön, jolle voidaan antaa vain yhteystiedoiksi ORCID-tunnisteen ja affiliaation.

Taulukko 7. Vertailumittari.

Mitattava metatieto	**Painoarvo painotetussa täydellisyyden metriikassa**	**Dublin Core**	**Minimimetatietomalli**	**FAIR-periaate**	**Käyttäjän tarve**	**Tutkimusaineiston metatietovaatimus**
Tekijä* (*creator*)	1	Sisältyy (*creator*)	Sisältyy (*toimija*)	Löydettävyys	Löytäminen	Tunnistava metatieto
Tekijän yksilöivä tunniste (*identifier*)*	1	Ei sisälly	Ei sisälly	Löydettävyys, Yhteentoimivuus	Viittaaminen, Löytäminen	Tunnistava metatieto
Nimi (*title)*	0.5	Sisältyy (*title*)	Sisältyy (*aineiston nimi*)	Löydettävyys	Löytäminen, Viittaaminen, Tunnistaminen	Tunnistava metatieto
Aihe (*subject,/subject headings)*	0,5	Sisältyy (*subject)*	Sisältyy (*aihe)*	Löydettävyys	Löytäminen, Tunnistaminen	Tunnistava metatieto,
Kuvaus* (*description)*	1	Sisältyy (*description)*	Ei sisälly	Löydettävyys, saavutettavuus	Hallitseminen, Analysoiminen, Löytäminen, Valitseminen	Tunnistava metatieto, Tieteellisen kontekstin metatieto
Tutkimusaineiston yksilöivä tunniste (*identifier)*	0,5	Sisältyy (*identifier)*	Sisältyy (*aineiston tunniste)*	Löydettävyys, saavutettavuus	Viittaaminen, Todentaminen	Tunnistava metatieto
Lisenssi* (*license)*	1	Sisältyy (*rights)*	Sisältyy (*käyttöehdot)*	Uudelleenkäytettävyys	Valitseminen, Hankkiminen	Tunnistava metatieto
Organisaatiotieto (*affiliation)*	0,5	Ei sisälly	ei sisälly	Löydettävyys, uudelleenkäytettävyys	Löytäminen, Valitseminen	Tunnistava metatieto,
Päivämäärä (*language)*	0,5	Sisältyy (*date)*	Sisältyy (*muokkaamisaika*)	Uudelleenkäytettävyys	Löytäminen, Tunnistaminen	Tunnistava metatieto

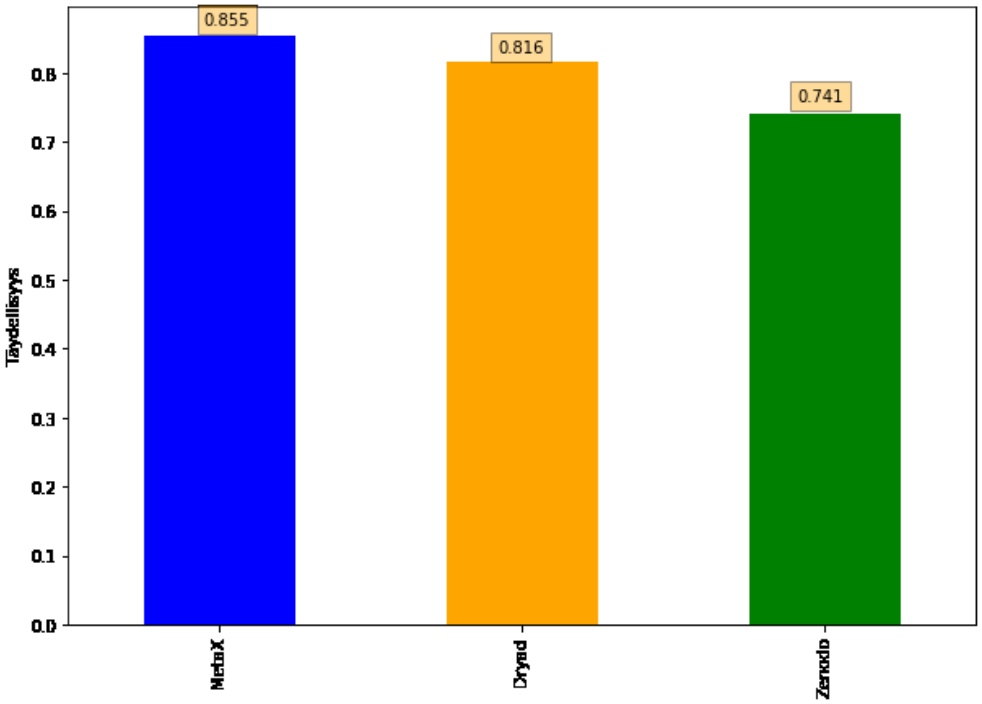
6 TULOKSET

Ennen tulosten läpikäyntiä on oleellista tuoda esiin tutkielmassa mitattujen data- ja metatietoarkistojen yhteneväisyyksiä tietomallien kannalta. Jokaisessa tutkissa arkistossa oli pakollisena tietona tutkimusaineistojen tekijä, tutkimusaineiston otsikko ja lisenssi. Lisäksi jokainen tutkittu arkisto muodostaa automaattisesti tutkimusaineistolle pysyvän tunnisteen. Nämä neljä mainittua kenttää saavat siis aina maksimaalisen arvon, jotka nostavat arkistojen saamia laatuarvoja. Käytännössä tämä tarkoittaa, että jokaisella metatietotietueella on laatuarvo vähintään 0,40 eli 40 % täydellisyyden metriikalla ja painotetun täydellisyyden metriikalla 46%. Tietomallien pakottamat tiedot olivat ennen aineiston keräystä tehdessä tiedossa[61]. Tulosten kannalta asetelma oli merkityksellinen, mutta ei estä vastaamasta tutkielman luvussa 5.1 esitettyihin tutkimuskysymyksiin.

6.1 Metatietojen laatu täydellisyyden metriikalla mitattuna

Täydellisyyden metriikalla saatiin luvut data- ja metatietoarkistojen metatietotietueiden laadusta. Parhaiten metatieto- ja data-arkistoista menestyi MetaX, jonka täydellisyyttä kuvaava laatukeskiarvo oli 0,855. The Dryad Digital Repository (jatkossa Dryad) sai laatukeskiarvoksi 0,816 ja heikoimman arvon Zenodo, joka sai 0,741. Arkistojen keskiarvo oli 0,80 (Kuva 5.)

[61] Teknisestä näkökulmasta tarkasteltuna metatietomalleihin perustuvien tietomallien on tallennettava aina jotain tietoa. Metatietomallien ja tietomallien suunnittelu tehdään aina käyttötarpeen mukaan kuten tutkielman luvussa 3 esitetään. Pakollisten kenttien suunnittelusta ja toteuttamisesta metatietomalleissa voi lukea lisää esimerkiksi osoitteessa https://www.sciencedirect.com/topics/computer-science/mandatory-field



Kuva 5. Metatietotietueiden laatu täydellisyyden keskiarvolla mitattuna

Keskiarvon ohella tarkasteltiin myös mediaania ja keskihajontaa. Mediaanilla ilmoitetaan lajitellun jakauman keskimmäinen arvo, ja keskihajonnalla kuinka kaukana havainnot olivat keskimäärin keskiarvosta. MetaXin mediaani oli 0,88 ja keskihajonta 0,567, Dryadin mediaani 0,77 ja keskihajonta 0,59, ja Zenodon mediaani 0,69 ja keskihajonta 0,80. MetaXin mediaani oli hieman keskiarvoa suurempi, jolloin jakauma on vasemmalle vino. Dryadilla ja Zenodolla mediaani oli keskiarvoa pienempi, jolloin jakauma on oikealle vino. Zenodon keskihajonta oli suurempi kuin Zenodon keskiarvo ja selvästi suurempi kuin MetaXilla ja Dryadilla. Tämä on mielenkiintoinen tulos ja kertoo hyvin suuresta vaihtelusta metatietotietueiden laadun välillä (Taulukko 8).

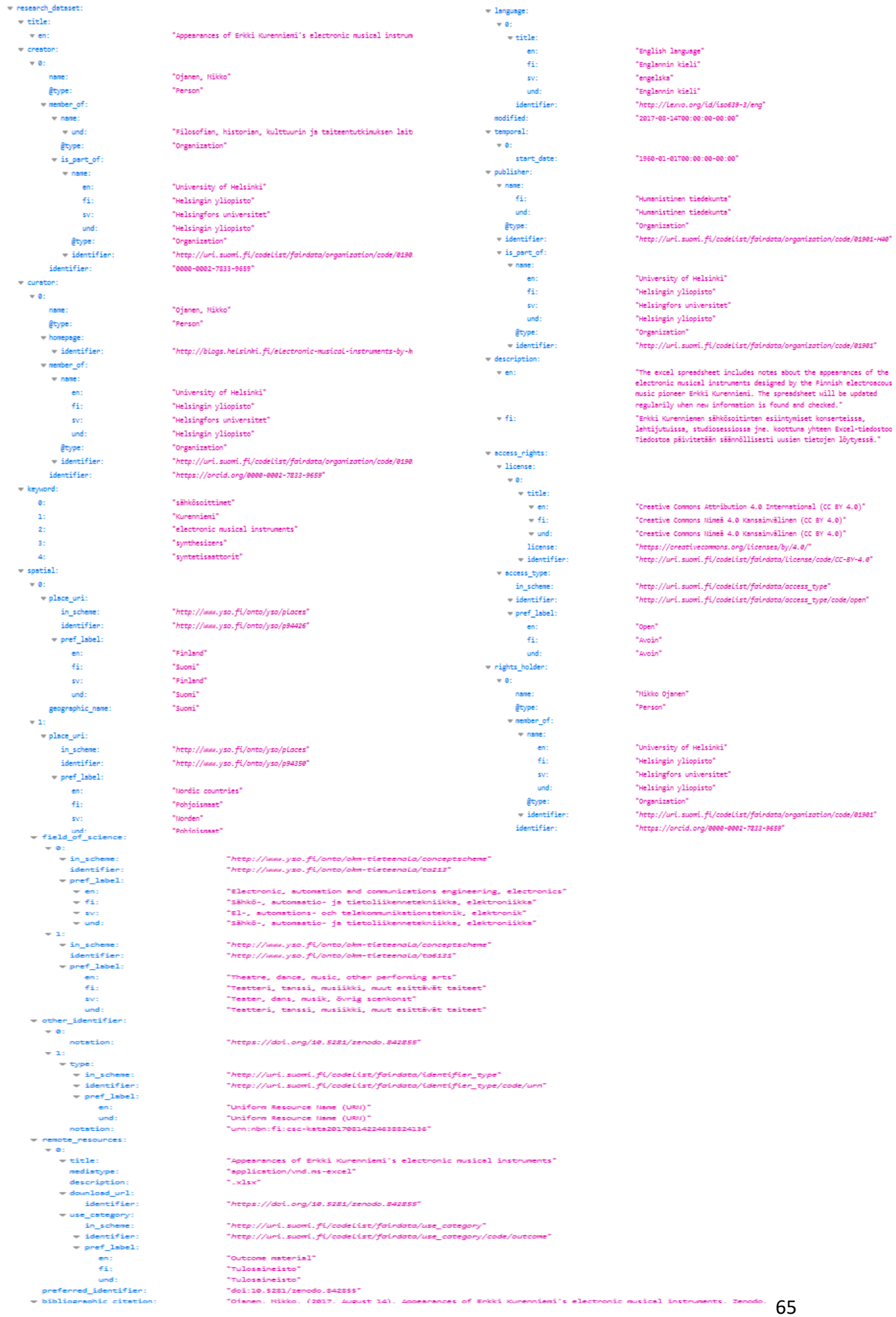
Laatuarvoja voidaan myös tarkastella mittariston parametri kerrallaan. Heikoiten arkistoihin oli kirjattu tietoja tekijän yksilöivästä tunnisteesta. Dryad-arkistosta saatiin vain lukema 0,02 ja on hyvin heikko verrattuna parhaaseen laatuarvoon, joka oli MetaXin 0,63. Arkistojen väliin sijoittui Zenodo laatuarvolla 0,13, joka sekin on hyvin heikko. Keskimäärin tästä saadaan laatuarvoksi 0,26 eli käytännössä vain noin joka neljännessä tietueessa on tutkijan yksilöivä tunnus. Toinen heikosti kirjattu tieto olivat asiasanat.

Dryad-arkistossa oli selvästi paras laatuarvo 0,86. MetaXissa arvo oli 0,21 ja Zenodossa alhaisin 0,10. Keskiarvoksi asiasanojen suhteen laatuarvoksi saadaan 0,39.

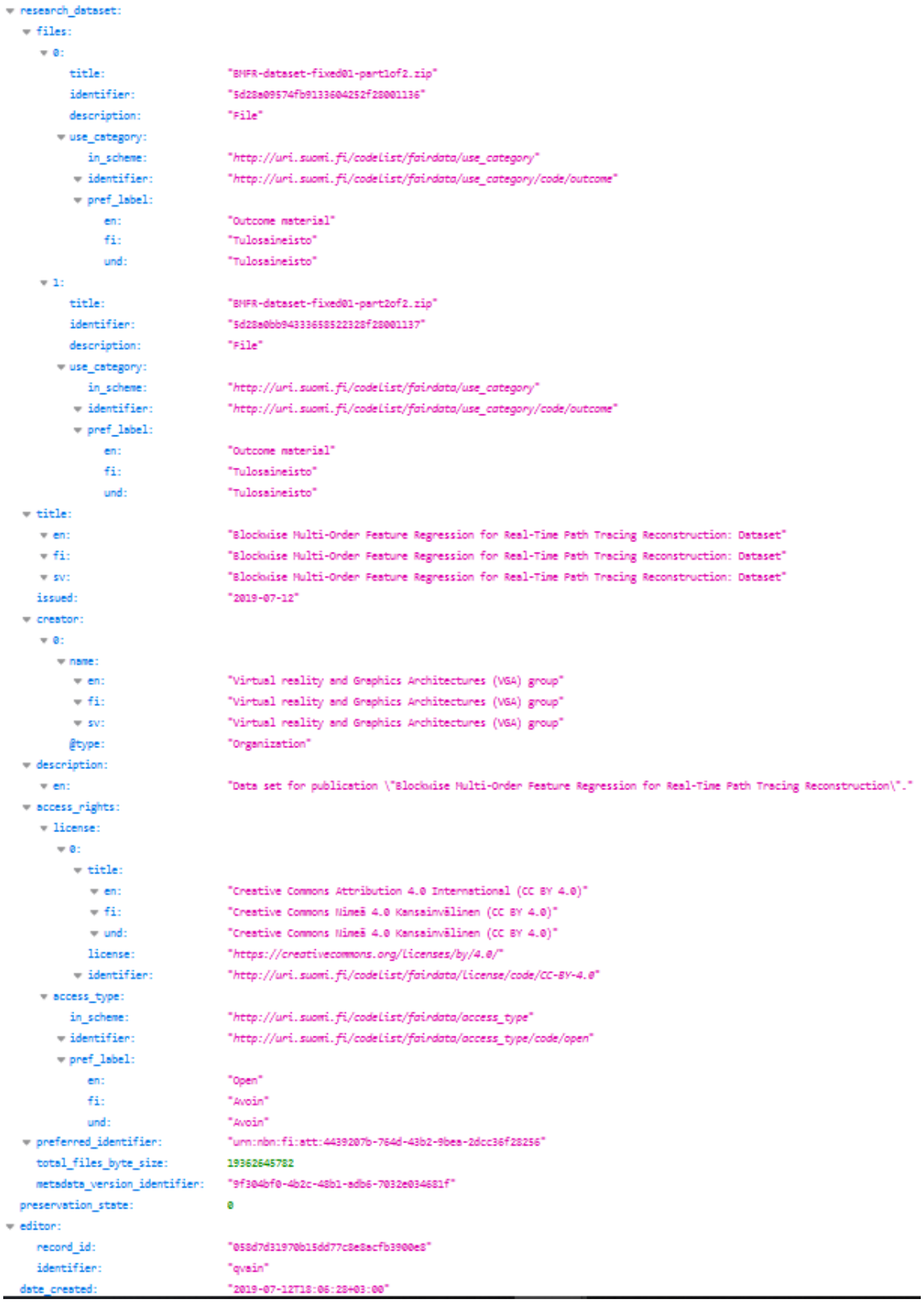
Taulukko 8. Tutkimusaineistojen metatietojen laatu täydellisyyden laatuarvoilla

Mittari	MetaX	Dryad	Zenodo
Tutkimusaineiston nimi	1,00	1,00	1,00
Tutkimusaineiston tekijä	1,00	1,00	1,00
Tekijän yksilöivä tunniste	0,63	0,02	0,13
Tekijän organisaatiotieto	0.87	0.48	0.43
Asiasanat	0.21	0.86	0.10
Kuvaus tutkimusaineistosta	1,00	0,97	1,00
Päivämäärä	0,99	0,97	0,95
Lisenssi	1,00	1,00	1,00
Tutkimusaineiston tunniste	1,00	1,00	1,00
Keskiarvo	0,855	0,816	0,741
Keskihajonta	0,57	0,59	0,80
Mediaani	0,88	0,77	0,69
Minimiarvo	0,55	0,55	0,67
Maksimiarvo	1,00	1,00	1,00
Vaihteluväli	0,45	0,45	0,33

Jokaisesta arkistosta löytyi täydellisiä metatietotietueita, jotka saivat arvon 1. Laadultaan heikoimmat metatiedot saivat vain arvon 0,55 ja vaihteluvälit MetaX- ja Dryad-arkistossa olivat korkeat, 0,45 yksikköä. Zenodossa minimiarvo oli 0,67 ja vaihteluväli arkistojen pienin, 0,33 yksikköä. Tulos on mielenkiintoinen, sillä Zenodo oli kuitenkin keskiarvolla ja mediaanilla mitattuna arkistoista heikoin (Taulukko 8). Esimerkit parhaan mahdollisesta arvon saaneesta metatietotietueesta (Kuva 6.) ja heikoimman arvon saaneesta (Kuva 7.) ovat nähtävissä alla.



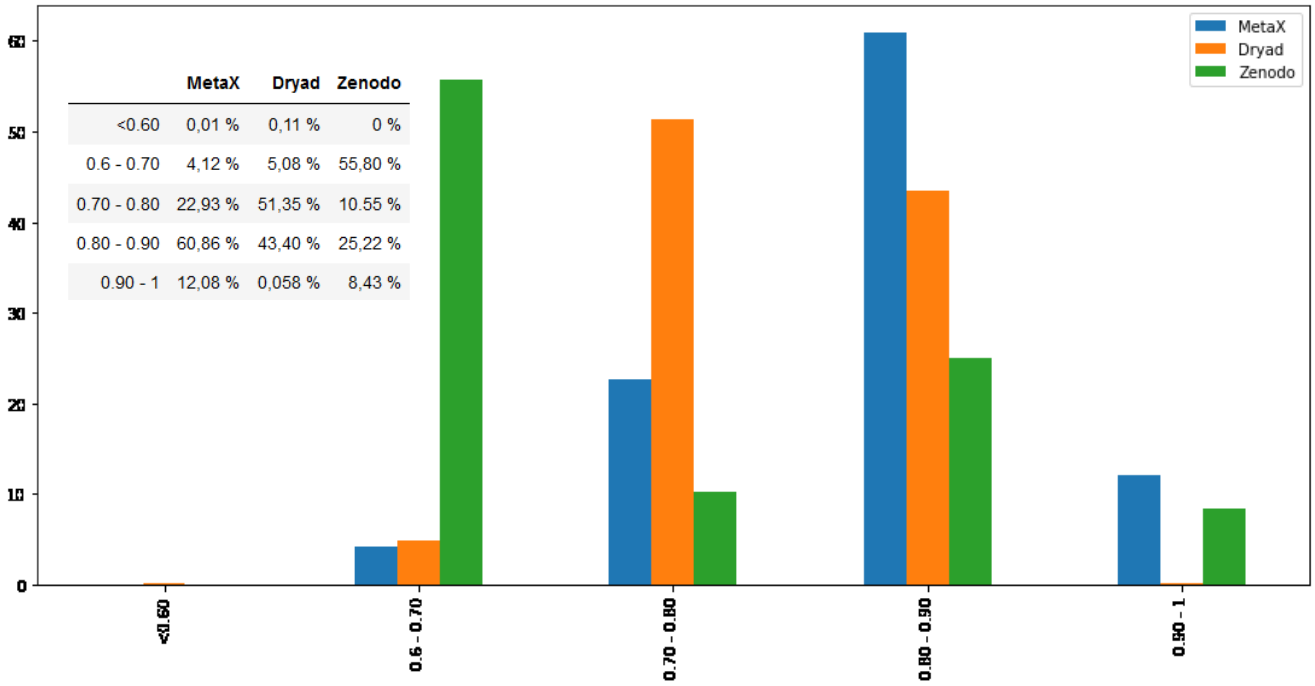
Kuva 6. Parhaimman mahdollisen laatuarvon (1,00) saanut metatietotietue MetaXista



Kuva 7. Heikoimman laatuarvon (0,55) saanut metatietotietue MetaXista

Laatuarvojen jakauma antaa tärkeää tietoa siitä, miten laatuarvot jakautuvat määriteltyjen luokkien välillä. Käytännössä voidaan esimerkiksi tarkastella sitä, kuinka suuri osa metatietotietueista kuului alimpaan luokkaan ja kuinka suuri osa sijoittuu ylimpään luokkaan. Luokitus tehtiin 0,10 asteen välein, pois lukien alin luokka, joka kattoi kaikki alle 0,60 arvon saaneet metatietotietueet (Kuva 8).

Heikoimmassa luokassa oli vain noin 30 metatietotietuetta koko aineistosta, joka sisälsi 89 918 tietuetta. Hyvään tulokseen vaikuttaa tietysti se, että osa metatietotietueiden kentistä oli tietomallin pakottamia kuten tämän luvun alustuksessa kerrotaan. Toiseksi alimpaan luokkaan sijoittui 31 290 metatietotietuetta eli reilu kolmannes koko aineistosta. Huomionarvoista on se, että Zenodon metatietotietueista 55,80 % (29 500) sijoittui toiseksi alimpaan luokkaan. MetaXista ja Dryadista vain 4,12 % ja 5,08 % sijoittuivat toiseksi alimpaan luokkaan.



	MetaX	Dryad	Zenodo
<0.60	0,01 %	0,11 %	0 %
0.6 - 0.70	4,12 %	5,08 %	55,80 %
0.70 - 0.80	22,93 %	51,35 %	10.55 %
0.80 - 0.90	60,86 %	43,40 %	25,22 %
0.90 - 1	12,08 %	0,058 %	8,43 %

Kuva 8. Metatietotietueiden täydellisyyden metriikan laatuarvot luokiteltuna

Keskimmäiseen luokkaan sijoittui 21 848 metatietotietuetta eli hieman alle neljännes koko aineistosta. Dryadin metatietotietueista 51,35 % sijoittui kolmanteen luokkaan, kun taas Zenodolta vain 10,55 %. MetaX sijoittui näiden arkistojen keskelle 22,93 % osuudellaan. Poikkeuksellista aineistossa on MetaXin erinomaiset luvut. Vain noin 27 % MetaXin metatietotietueista sijoittui kolmeen alimpaan luokkaan. Dryadilla vastaava luku on 56,54 % ja Zenodolla 66,35 %.

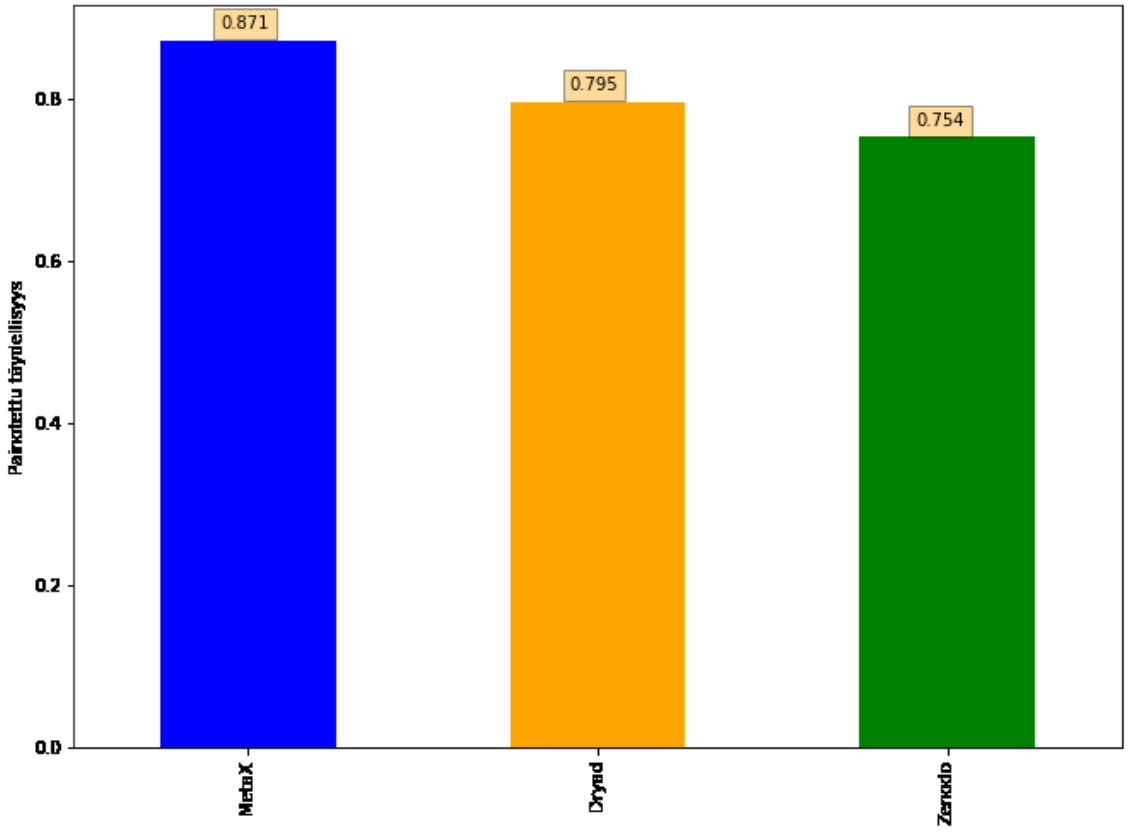
Toiseksi ylimpään luokkaan sijoittui 24 295 metatietotietuetta eli hieman yli neljännes koko aineistosta. MetaXin aineistoista 60,86 % sijoittui toiseksi ylimpään luokkaa. Tätä toisaalta selittää jälleen aiemmin mainittu Helsingin fysiikan tieteenalan tuottamat 7000 metatietotietueen kokoelma, jotka ovat sisällöltään ja laatuarvoltaan lähes samanlaiset. Ylimpään luokkaan sijoittui lopulta 5783 metatietotietuetta, joka on 6,5 % koko aineistosta. Huomionarvoista ylimmässä luokassa on Dryadin metatietotietueiden vähäinen määrä. Vain 0,058 % eli noin 158 tietuetta Dryadista sai parhaaseen luokaan kuuluvan laatuarvon.

Yhteenvetona täydellisyyden metriikasta voidaan todeta, että noin 33 % eli kolmannes kaikista metatietotietueista sijoittui kahteen ylimpään luokkaan, jotka käytännössä kertovat tutkimusaineiston metatietotietueen 80 % - 100 % täydellisyydestä. Lähes 35 % sijoittui kahteen alimpaan luokkaan, mutta on huomioitava, että alin luokka oli alle 60 % laatuarvon saaneet ja toiseksi alin 60 % - 70 %, joten ei voida kuitenkaan puhua huonosti kuvailluista tutkimusaineistoista.

Arkistojen kannalta mielenkiintoista oli erityisesti Zenodon laadunvaihtelu arkiston sisällä. Tämä näkyy keskihajontana, joka oli suurempi kuin keskiarvo. Lisäksi pitää huomioida se, että Zenodon vaihteluväli oli pienin ja minimiarvo 0,67. Kolmas luokitus alkaa arvosta 0,70, joten Zenodolla on 55% tietueistaan 0,67 ja 0,70 arvon välillä. Selitys tähän saattaa olla samankaltainen kuin MetaXin 7000 lähes samanlaisen metatietotietueen kokoelma.

6.2 Metatietojen laatu painotetun täydellisyyden metriikalla mitattuna

Painotetun täydellisyyden metriikalla saatiin luvut data- ja metatietoarkistojen metatietotietueiden laadusta antamalla suurempi painoarvo tärkeille metatietokentille. Parhaiten metadata- ja data-arkistoista menestyi MetaX, jonka painotettua täydellisyyttä kuvaava keskiarvo oli 0,871. Dryad sai keskiarvoksi 0,795 ja Zenodo 0,754. Arkistojen keskiarvo oli 0,80 (Kuva 5.)



Kuva 9. Metatietotietueiden laatu painotetun täydellisyyden keskiarvolla mitattuna

Painotetulla täydellisyyden metriikalla mitattuna MetaXin mediaani oli 0,92 ja keskihajonta 0,5, Dryadin mediaani 0,77 ja keskihajonta 0,33, ja Zenodon mediaani 0,71 ja keskihajonta 0,54. Huomionarvoista on Zenodon ja Dryadin selvästi pienempi keskihajonta kuin täydellisyyden metriikalla. MetaXilla keskihajonta on vain hieman pienempi kuin täydellisyyden metriikalla mitattuna. Tämä ei kuitenkaan vaikuttanut keskiarvon jakaumaan ne pysyivät lähes samoina molemmilla mittaustavoilla (Taulukko 9).

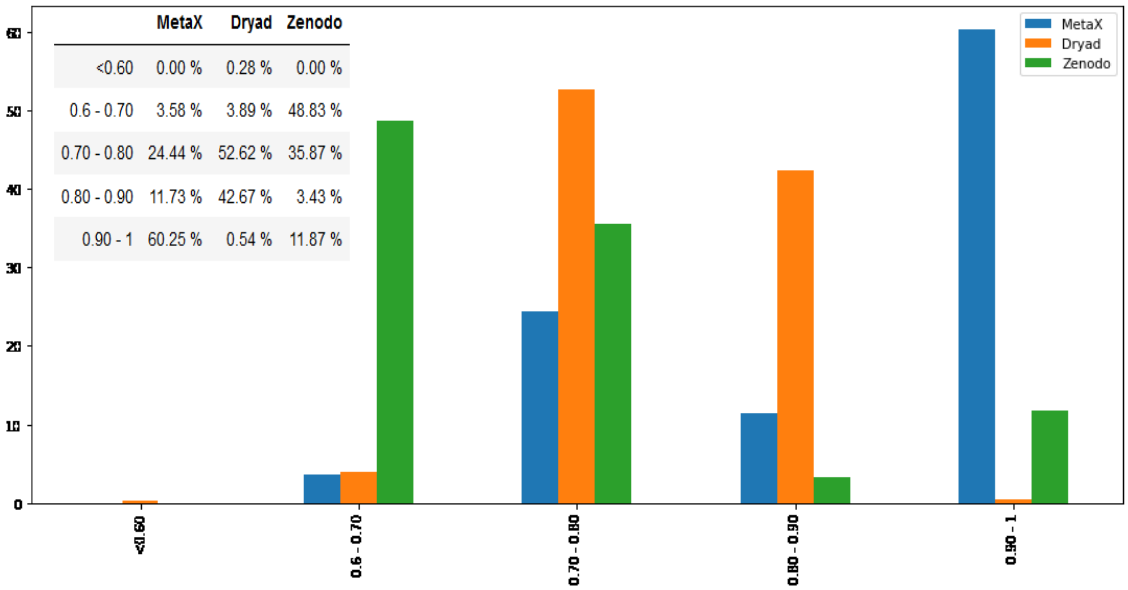
Metatietotietueiden minimilaatuarvoissa nähtiin pieniä muutoksia. MetaXin minimilaatuarvo oli painotetulla täydellisyydellä 0,62, kun se oli täydellisyydellä 0,55. Eroa oli 0,07 yksikköä. Dryadilla muutos meni toiseen suuntaan, painotetun täydellisyydellä minimiarvo oli 0,46 ja täydellisyydellä 0,55. Eroa oli 0,09 yksikköä. Zenodon muutos oli hyvin pieni, 0,02 yksikköä. Tärkeiden metatietokenttien painotus vaikutti siis positiivisesti MetaXiin ja Zenodoon, mutta negatiivisesti Dryadiin. Myös Dryadin keskiarvo laski 0,02 yksikköä. Muutos oli pieni, joten tästä ei voi tehdä suurempia johtopäätöksiä.

Taulukko 9 Tutkimusaineistojen metatietojen laatu painotetun täydellisyyden laatuarvoilla

Mittari	MetaX	Dryad	Zenodo
Tutkimusaineiston nimi	0,50	0,50	0,50
Tutkimusaineiston tekijä	1,00	1,00	1,00
Tekijän tunniste	0,63	0,02	0,13
Tekijän organisaatio	0,44	0,24	0,22
Asiasanat	0,11	0,43	0,05
Kuvaus tutkimusaineistosta	1,00	0,97	1,00
Päivämäärä	0,49	0,47	0,45
Lisenssi	1,00	1,00	1,00
Tutkimusaineiston tunniste	0,50	0,50	0,50
Keskiarvo	0,87	0,79	0,75
Keskihajonta	0,5	0,33	0,537
Mediaani	0,92	0,77	0,71
Minimiarvo	0,62	0,46	0,69
Maksimiarvo	1,00	1,00	1,00
Vaihteluväli	0,38	0,54	0,30

Vaihteluväleissä nähtiin mielenkiintoinen muutos. MetaXin vaihteluväli pieneni 0,07 yksikköä ja Dryadilla nousi 0,09 yksikköä. Täydellisyyden metriikalla MetaXilla ja Dryadilla oli täsmälleen sama vaihteluväli (0,45), mutta painotetulla täydellisyydellä luvut ovat 0,38 ja 0,54. Tässä tulee esiin painotetun täydellisyyden periaate, jossa painotetaan etukäteen valittuja, tärkeitä metatietokenttiä. MetaXin ja Dryadin tapauksessa erot selittyvät MetaXin vahvalla ja Dryadin heikolla laatuarvolla tekijän tunnisteesta. Lisäksi Dryadin vahvin osa-alue oli asiasanat, jotka ovat painotetussa täydellisyyden metriikassa pienemmällä painotuksella. Samalla myös MetaXin ja Dryadin keskiarvojen erotus kasvoi 0,04 yksikköä MetaXin eduksi.

Painotetun täydellisyyden osalta laatuarvojen jakautuminen eri luokkiin toteutettiin samalla tavalla kuin täydellisyyden laatuarvojen. Alimpaan luokkaan sijoittui 76 metatietotietuetta, jotka olivat kaikki Dryadista. Toiseksi alimpaan luokkaan sijoittui 27 231 metatietotietuetta eli hieman alle kolmannes koko aineistosta. Arkistojen suhteen ei tapahdu suuria muutoksia. Zenodon metatietotietueista 48,83 % oli toiseksi alimmassa luokassa, kun taas MetaXilta ja Dryadilta hieman alle 4 %. (Kuva 10).



Kuva 10. Metatietotietueiden painotetun täydellisyyden metriikan laatuarvot luokiteltuna

Keskimmäiseen luokkaan sijoittui 35 730 metatietotietuetta, joka on 40 % koko aineistosta. Keskimmäisen luokan suhteen painotetun täydellisyyden ja täydellisyyden metriikan välillä oli merkittävä ero, sillä täydellisyyden metriikalla n. 25 % metatietotietueista sijoittui keskimmäiseen luokkaan. Arkistojen suhteen tarkasteltuna suurin muutos tapahtui Zenodolla, jolla muutos oli peräti 25,3 % prosenttiyksikköä täydellisyyden metriikan ja painotetun täydellisyyden metriikan välillä.

Toiseksi ylimpään luokkaan sijoittui 14 625 metatietotietuetta, joka on 16 % koko aineistosta. Täydellisyyden metriikalla vastaava lukema oli 27 %, joten erot ovat merkittävät. Eroa selittävät Zenodon laskeneet laatuarvot, sillä täydellisyyden metriikalla Zenodolla oli toiseksi ylimmässä luokassa 25,22 % ja painotetulla täydellisyyden metriikalla vain 3,43 %.

Ylimpään luokkaan sijoittui lopulta 12 256 tietuetta, joka on lähes 14 % koko aineistosta. MetaX menestyi arkistoista parhaiten myös painotetun täydellisyyden metriikalla. Peräti 60,25 % MetaXin metatietotietueista sijoittui ylimpään luokkaan.

6.3 Täydellisyyden ja painotetun täydellisyyden metriikoiden erot

Keskiarvoja vertailtaessa erot olivat pienet eri metriikoiden välillä. Suurin muutos tapahtui Dryadilla, jonka keskiarvo laski 0,03 yksikköä. MetaXin ja Zenodon keskiarvot taas nousivat 0,01 yksikköä. Arkistojen välillä suurin ero keskiarvossa oli 0,12 yksikköä MetaXin ja Zenodon välillä. Tämä ero toteutui molemmilla metriikoilla. Kaikkien arkistojen keskiarvo oli molemmilla metriikoilla 0,80. Mediaaniarvoissa muutokset olivat myös pieniä. Suurin muutos tapahtui MetaXilla, jonka mediaani nousi 0,04 yksikköä. Dryadilla mediaani pysyi täsmälleen samana ja Zenodolla nousi 0,02 yksikköä. Arkistojen välillä suurin ero mediaanista löytyi painotetun täydellisyyden metriikalla, jossa MetaXin ja Zenodon ero oli peräti 0,21 yksikköä (Taulukko 10).

Taulukko 10. Yhteenveto tutkielman tuloksista

	Täydellisyys			Painotettu täydellisyys		
Mittari	**MetaX**	**Dryad**	**Zenodo**	**MetaX**	**Dryad**	**Zenodo**
Aineiston nimi	1,00	1,00	1,00	0,50	0,50	0,50
Aineiston tekijä	1,00	1,00	1,00	1,00	1,00	1,00
Tekijän tunniste	0,63	0,02	0,13	0,63	0,02	0,13
Tekijän organisaatio	0,87	0,48	0,43	0,44	0,24	0,22
Asiasanat	0,21	0,86	0,10	0,11	0,43	0,05
Kuvaus aineistosta	1,00	0,97	1,00	1,00	0,97	1,00
Päivämäärä	0,99	0,97	0,95	0,49	0,47	0,45
Lisenssi	1,00	1,00	1,00	1,00	1,00	1,00
Aineiston tunniste	1,00	1,00	1,00	0,50	0,50	0,50
Keskiarvo	0,86	0,82	0,74	0,87	0,79	0,75
Keskihajonta	0,57	0,59	0,80	0,50	0,33	0,54
Mediaani	0,88	0,77	0,69	0,92	0,77	0,71
Minimiarvo	0,55	0,55	0,67	0,62	0,46	0,69
Maksimiarvo	1,00	1,00	1,00	1,00	1,00	1,00
Vaihteluväli	0,45	0,45	0,33	0,38	0,54	0,30
Luokitus						
< 0,6	0,01 % (1)	0,11 % (*30*)	0,00% (*0*)	0,00 % (*0*)	0,28% (*76*)	0,00% (*0*)
0,6 - 0,7	4,12 % (*399*)	5,08 % (*1390*)	55,80 % (*29 400*)	3,58 % (*347*)	3,89% (*1064*)	48,83% (*25 815*)
0,7 - 0,8	22,93 % (*2220*)	51,35 % (*14 001*)	10,55 % (*5477*)	24,44% (*2366*)	52,62 % (*14 400*)	35,87 % (*18 964*)
0,8 - 0,9	60,86 % (*5893*)	43,40 % (*11 776*)	25,22 % (*13 233*)	11,73% (*1136*)	42,67 % (*11 676*)	3,43 % (*1813*)
0,9 – 1	12,08 % (*1170*)	0,058 % (*158*)	8,43 % (*4357*)	60,25 % (*5836*)	0,54 % (*149*)	11,87 % (*6276*)

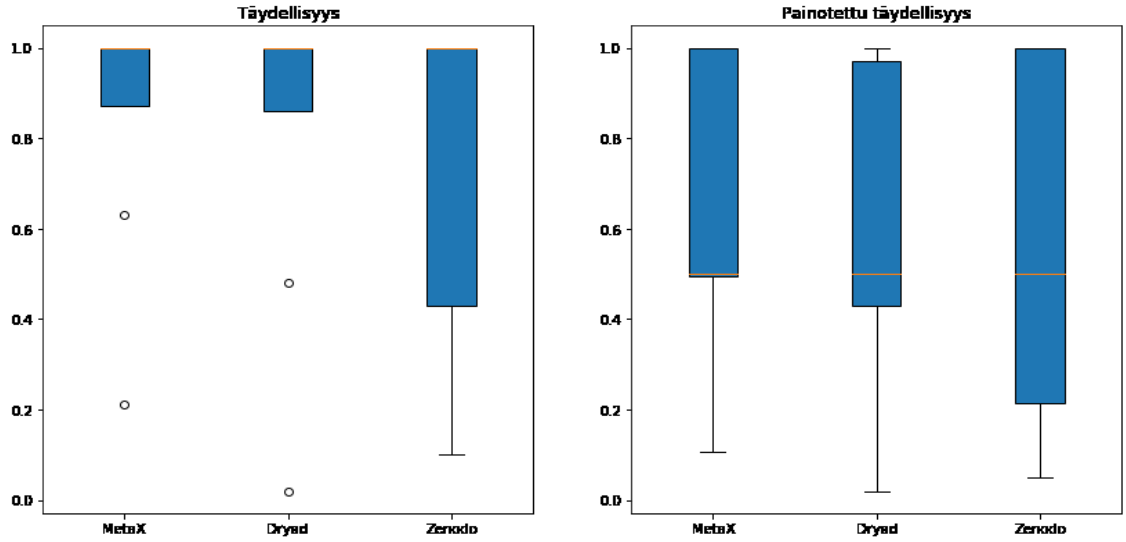
Mielenkiintoista on se, että Zenodossa oli suhteellisesti eniten täydellisiä tietueita. Yhteensä 406 Zenodon tietuetta sai laatuarvon 1, joka on 0,76 % koko Zenodon aineistoista. Tämäkin tulos kertoo Zenodon laajasta hajonnasta läpi koko arkiston ja on ilmiönä mielenkiintoinen.

Täydellisyyden ja painotetun täydellisyyden mittaustuloksissa oli myös selviä eroja, kun metatietueiden laatu jaettiin viiteen eri luokkaan 0,10 yksikön välein. Täydellisyyden metriikalla kahteen alimpaan käytettyyn luokitukseen (< 0,60 ja 0,60 – 0,70) sijoittui 35% metatietotietueista, kun taas painotetun täydellisyyden metriikalla 30 %. Tämän perusteella voitaisiin todeta, että painotettu täydellisyys hieman nostaisi metatietotietueiden tasoa. Vastakkaista näyttöä saadaan kuitenkin kahden ylimmän luokituksen (0,80 – 0,90 ja 0,90 – 1) tuloksista. Painotetulla täydellisyyden metriikalla kahteen ylimpään luokkaan sijoittui 30 % metatietotietueista, mutta täydellisyyden metriikalla 3 prosenttiyksikköä enemmän eli 33 %. Keskimmäisessä luokassa (0,7 – 0,8) oli myös merkittävä ero. Täydellisyyden metriikalla keskimmäiseen luokkaan sijoittui 24 % ja painotetun täydellisyyden metriikalla noin 40 % metatietotietueista (Taulukko 10).

Vertailumittarin (ks. Taulukko 7) parametreissa nähtiin suuria eroja. MetaXissa tekijän tunniste sai laatuarvon 0,63, kun toiseksi paras lukema oli Zenodon 0,13. Eroa oli peräti 0,50 yksikköä. Samoin tekijän affiliaatio sai MetaXissa laatuarvon 0,87, kun toiseksi parhaan laatuarvon saaneessa Dryadissa se oli 0,48. Eroa oli 0,39 yksikköä. Asiasanoissa taas Dryadin laatuarvo oli 0,86 ja toiseksi parhaan laatuarvon saaneen MetaXin 0,21 (Taulukko 10).

Erot olivat merkittäviä. Mistä näin suuret erot johtuvat? Yksi keskeinen selitys MetaXin korkeille arvoille tekijän tunnisteen ja organisaatiotiedon suhteen on se, että aineisto oli kooltaan pienin (9684 metatietotietuetta) ja aineisto sisälsi 7400 metatietotietuetta Helsingin fysiikan laitokselta. Näissä tietueissa oli käytetty hyvin useasti tekijän tunnistetta ja tekijän organisaatiotietoa, mutta asiasanoitus oli puutteellinen. Dryadissa oli taas käytetty läpi koko aineiston poikkeuksellisen hyvin asiasanoja. Zenodossa vaihtelu näiden kolmen metatiedon suhteen oli hyvin suurta. Näiden kolmen metatietokentän kirjaaminen oli selvästi puutteellisinta.

Hajonta vertailumittarissa määriteltyjen parametrien (ks. Taulukko 7) suhteen oli varsin suurta. Tätä voidaan havainnollistaa laatikko-jana (*boxplot)* -kuviolla, jolla voidaan tarkastella muuttujien jakaumaa ja hajontaa. Kuvasta nähdään, että kaikilla arkistoilla on minimiarvoja selvästi keskiarvosta poikkeavina sekä täydellisyyden että painotetun täydellisyyden metriikalla mitattuna. MetaXilla ja Dryadilla oli varsin samanlainen hajonta parametrien kesken riippumatta käytetystä metriikasta. Nämä näkyvät poikkeavina pisteinä kuviossa. Zenodossa oli arkistoista selvästi eniten hajontaa parametrien kesken molemmilla metriikoilla (Kuva 11).



Kuva 11. Vertailumittarin parametrien hajonta täydellisyyden ja painotetun täydellisyyden metriikoilla

Zenodon suuresta hajonnasta tekee mielenkiintoisen se, että Zenodolla oli molemmilla metriikoilla mitattuna suurin minimiarvo ja pienin vaihteluväli (ks. Taulukko 10). Tämä selittyy hyvin pitkälti sillä, että Zenodossa on todella paljon samoilla laatuarvoilla tuotettuja metatietotietoja.

7 JOHTOPÄÄTÖKSET

Tämän tutkielman keskiössä oli tutkimuskysymys, jonka tavoite oli selvittää, millaista on tutkimusaineistojen metatietojen laatu automaattisilla menetelmillä mitattuna. Toisen tutkimuskysymyksen tavoite oli selvittää, millaisia metatietojen laatuongelmia tutkimusaineistoissa ilmenee.

Tutkielmassa käytetty tutkimusaineisto on näyte data- ja metatietoarkistoista. Tutkimus on tilastollinen tutkimus. Tutkielman tulokset eivät ole harkinnanvaraisen näytteen vuoksi yleistettävissä. Tutkielman tutkimusaineisto kerättiin automaattisin menetelmin, koska se mahdollisti otannan kohdistamisen useampaan data- ja metatietoarkistoon. Tutkimuksessa käytettiin täydellisyyden ja painotetun täydellisyyden metriikoita, jotka pohjautuivat Ochoan ja Duvalin (2009b) ja Margaritopoulouksen (2012) kehittämiin laadun mittausmenetelmiin. Lisäksi tutkielmaa varteen luotiin vertailumittari, joka määritti, mitä metatietoja aineistosta haluttiin löytyvän.

Tulosten perusteella laadussa on paljon hajontaa arkistojen kesken ja arkistojen sisällä. Kaikkien otannassa olleiden arkistojen laatukeskiarvo oli 0,80 eli 80 %. Laatukeskiarvo oli sama molemmilla metriikoilla. Keskiarvoilla mitattuna erot arkistojen välillä olivat enimmillään 0,12 yksikköä ja mediaanilla 0,21 yksikköä. Arkistojen sisällä metatietotietueiden hajonta oli varsin suuri. Keskihajonta oli Zenodon kohdalla keskiarvoa korkeampi ja vaihteluvälit olivat keskimäärin suuret. Arkistojen erot olivat merkittävät myös siinä mielessä, että vertailumittarin parametrien määrä oli vähäinen. Vertailumittari luotiin siitä näkökulmasta, että se sisältäisi tieteenalasta riippumattomat metatiedot, jotka löytyvät jokaisen tutkimusaineiston metatiedoista.

Aineistojen hajontaa tarkasteltiin tarkemmin luokittelemalla metatietotietueet niiden laatuarvon mukaan 0,10 yksikön välein. Luokitus oli 5-portainen ja alkoi metatietotietueista, joiden laatuarvo oli pienempi kuin 0,60 ja päättyi ylimpään mahdolliseen laatuluokitukseen, joka oli väliltä 0,90 – 1. Täydellisyyden metriikalla 40 % kaikista metatietotietueista sijoittui kahteen ylimpään luokkaan ja sai laatuarvon väliltä 0,8 – 1. Painotetun

täydellisyyden metriikalla vastaava luku oli 30 %. Kahteen alimpaan luokkaan sijoittui täydellisyyden metriikalla 35 % tietueista ja painotetun täydellisyyden metriikalla 30 %.

Tutkimusaineistojen metatietojen keskimääräistä laatua voidaan tämän tutkimuksen perusteella pitää hyvänä, vaikka hajonta oli suurta ja vertailumittari oli hyvin kapea. Tämän tutkimuksen tulokset ovat linjassa aiemman tutkimuksen kanssa siinä mielessä, että puuttuvia tietoja löytyi lähes kaikista metatietotietueista. Ainoastaan 0,56 % metatietotietueista oli täydellisiä. Tutkimuksen keskimääräisellä metatietotietueen laatuarvolla saavutetaan kuitenkin tässä tutkielmassa annetut tutkimusaineistojen metatietojen laatuvaatimukset, jotka olivat löydettävyys, käytettävyys ja hyödyllisyys.

Yksi syy hyviin laatuarvoihin on se, että arkistojen tietomallit pakottivat vertailumittarissa käytettyjen tärkeiden metatietojen kirjaamisen. Tämä edisti huomattavasti tutkimusaineistojen laatua ja osoittaa myös sen, että arkistojen tietomallien suunnitteluun kannattaa panostaa. Tärkeiden tietojen kirjaaminen kannattaa siis asettaa pakolliseksi arkistojen tietomalleissa, jolloin metatietojen laatu, löydettävyys, käytettävyys ja hyödyllisyys on parempaa.

Suurimmat ongelmat metatietojen kirjaamisessa liittyivät niihin metatietokenttiin, jotka eivät olleet tietomalleissa pakollisena. Yksittäisissä metatietokentissä ongelmia esiintyi erityisesti tutkijoiden yksilöivän tunnisteeseen, organisaatiotietoihin ja asiasanoituksiin. Tutkijoiden yksilöivän tunnisteen suhteen voidaan nostaa esiin mahdollinen kyseisen käsitteen tuntemattomuus tutkijapiireissä. Yksilöivän tunnisteen käyttö on yleistynyt vasta viime vuosina. Tutkielmassa käytetty aineisto oli kuitenkin sen verran uutta, että yksilöivän tunnisteen pitäisi olla ainakin käsitteenä tuttu. Ongelmallisten metatietojen kirjaamiseen ja laadun parantamiseen voi auttaa tutkijoiden intensiivisempi koulutus ja tutkijoille kohdistettu viestintä tutkimusorganisaatioiden taholta. Lisäksi arkistojen ohjeistus ja tietomallien suunnittelu voisivat auttaa näiden metatietojen kirjaamisessa.

Molempien tutkimuskysymysten taustalla käsiteltiin arkistojen tietomalleja. Kuten edellä esitettiin, on hyvin todennäköistä, että tietomallit ja erityisesti pakolliset kentät vaikuttavat vahvasti metatietojen laatuun. Olisiko siis tarpeen luoda tutkijoille sellainen

kuvailutyökalu tai tietomalli, että tärkeitä metatietokenttiä ei voida ohittaa vaan ne pitää aina kirjata? Toisaalta on myös huomioitava se, että mitään yhtenäistä metatietostandardia tutkimusaineistojen kuvailuun ei ole ja standardien määrä on suuri. Tämä vaikeuttaa tietomallien suunnittelemista ja luomista. Huomionarvoista on kuitenkin se, että tunnistavan metatiedon suhteen hajonta eri tieteenalojen välillä ei läheskään ole niin suurta kuin varsinaista tutkimusaineistoa kuvailevan tieteellisen metatiedon suhteen. Kuvailevan tiedon yhteismitallisuus ei todennäköisesti ole läheskään yhtä hankalaa toteuttaa kuin tieteellisen metatiedon suhteen.

Tutkimuksen vertailumittarin suunnittelu tehtiin avoimen tieteen ja tutkimusaineistojen kuvailun kontekstissa, jotta laatuarvoista saataisiin mahdollisimman relevantteja tutkimuksen kannalta. Painopiste oli avoimen tieteen käytänteissä ja periaatteissa, jolla pyritään edistämään yhteiskunnallista vaikuttavuutta, näkyvyyttä ja yhteiskäyttöisyyttä. Tutkimusta varten suunniteltu vertailumittari toimi hyvin laadun mittauksen välineenä ja sillä saatiin relevantti kokonaiskuva näytteeseen kuuluvien arkistojen metatietojen laadusta.

Tutkielman tutkimuksen rajoitteet kohdistuivat täydellisyyden ja painotetun täydellisyyden metriikkoihin. Tässä tutkimuksessa mitattiin, löytyykö haluttu vertailumittarin tieto metatietotietueesta eli millä tavalla kuvailukenttiä oli hyödynnetty. Tutkielmassa ei tehty laatuanalyysia metatietojen sisällöstä. Metatietojen sisällön analysoiminen tarkkuuden ja johdonmukaisuuden metriikoilla voisi tuoda lisäarvoa laadun analysointiin. Jatkotutkimuksen aihe voisi olla samaan aineistoon kohdistettu sisällönanalyysi, jossa hyödynnetään tarkkuuden ja johdonmukaisuuden metriikoita. Näin saataisiin kattavampi kuva metatietojen varsinaisesta sisällöstä ja tulokset voisivat olla hyvinkin erilaisia.

Toinen tutkimuksen rajoite liittyy harkinnanvaraiseen näytteeseen. Tutkimuksen näytteessä oli vain kolme data- ja metatietoarkistoa. Mikäli haluttaisiin tehdä yleistettävissä olevia tuloksia kuten Neumaier, Umbrich ja Pollenes (2016) yli 260 arkiston otannassaan, pitäisi otannan olla kattavampi. Jatkotutkimus voisi toteuttaa samalla tutkimusasetelmalla, mutta huomattavasti laajemmalla otannalla.

Tutkimusaineistojen kuvailulla on yllättävänkin pitkä historia, joka alkaa jo 1970-luvulta. Tästä huolimatta tutkimusaineistojen kuvailu on yleistynyt merkittävästi vasta viimeisten 10 vuoden aikana, jolloin avoimen tieteen vaatimukset ja uudistuvat tiedekäytännöt ovat lisänneet tutkijoiden painetta aineistonhallintaan ja tutkimusaineistojen kuvailuun. Kuten kirjallisuuskatsauksessa tuli ilmi, tutkimusorganisaatioiden tuki tutkijoille ei ole ollut riittävällä tasolla.

Metatietoja ja niiden laatua on tutkittu hieman yli 20 vuotta. Tutkimusaineistojen metatietojen laatua vielä huomattavasti vähemmän aikaa. Suomessa tutkimusaineistojen metatietojen laatua ei ole juurikaan tutkittu. Tässä tutkielmassa toteutettua pro gradu -tutkimusta voidaan Suomessa pitää pelinavauksena tutkimusaineistojen metatietojen laadunarviointiin ja kartoittamiseen. Tulevaisuudessa tutkimusaineistojen metatietoja on tarpeellista tutkia avoimen tieteen käytäntöjen jalkautuessa yhä useampaan tutkimusorganisaatioon, kansallisiin tieteentekemisen periaatteisiin ja tutkijan arkeen.

Tutkimusaineistojen kuvailutyökalujen kehittäminen ja tutkijoiden kouluttaminen sekä tutkimusaineistojen laadun seuraaminen ovat ensisijaisesti tutkimusorganisaatioiden ja erilaisten tukipalveluiden kuten kirjastojen kehittämien tutkimuspalveluiden vastuulla. Näiden asioiden huomioiminen ja niihin panostaminen tutkimusorganisaatioiden toimesta nostaa todennäköisesti tutkimusaineistojen metatietojen laatua.

LÄHTEET

Ala-Kyyny, J. 2016. Yliopistokirjastojen rooli avoimessa julkaisemisessa. Pro Gradu. Tampereen yliopisto. Saatavilla: http://urn.fi/URN:NBN:fi:uta-201610122420

Ala-Kyyny, J., Korhonen, T. & Roinila, M. 2017. Tutkimusdatan avaamisen esteet: haastattelututkimus Helsingin yliopistossa. Signum 49(4), 25-29. Luettu 26.3. Saatavilla: https://doi.org/10.25033/sig.69198

Amorim, R., Aguiar, C., Rocha, J. & Ribeiro, C. 2016. A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society. Luettu 1.3.2019 Saatavilla https://www.researchgate.net/publication/303918099_A_comparison_of_research_data_management_platforms_architecture_flexible_metadata_and_interoperability

Archambault, É. Amyot, D. Deschamps, P., Nicol, A. Provencher, F., Rebout, L. & Roberge, G. 2014. Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013. European Commission Available Luettu Saatavilla 26.3. http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

Auckland, M. 2012. Re-skilling for research: An investigation into the roles and skills of subjectand liaison librarians required to effectively support the evolving information needs of researchers. London: RLUK Research Libraries UK. Saatavilla: https://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf

Avoin tiede ja tutkimus. 2019. Tutkimusorganisaatioiden avoimuuden linjaukset ja ohjeistukset. TSV. Luettu 15.3.. Saatavilla https://avointiede.fi/fi/linjauksia/ohjeita-ja-linjauksia/tutkimusorganisaatioiden-avoimuuden-linjaukset-ja-ohjeistukset

Babaii, E. &Taase, Y. 2013. Author-assigned Keywords in Research Articles: Where Do They Come From? Iranian Journal of Applied Linguistics (IJAL), Vol. 16, No. 2. Luettu 5.9. Saatavilla: https://ijal.khu.ac.ir/article-1-1786-fa.pdf

Balatsoukas, P., Rousidis, D., & Garoufallou, E. 2018. A method for examining metadata quality in open research datasets using the OAI-PMH and SQL queries: the case of

the Dublin Core 'Subject' element and suggestions for user-centred metadata annotation design. IJMSO, 13, 1-8.

Baker, M. 2016. 1,500 Scientists Lift the Lid on Reproducibility. Nature 533(7604):, s. 425-454. doi:10.1038/533452a Saatavilla: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Blank, G. & Rasmussen, K.B. 2004. The Data Documentation Initiative: The Value and Significance of a Worldwide Standard. Social Science Computer Review, 22(3), 307–318. https://doi.org/10.1177/0894439304263144 Luettu 22.10.2019

Borg, S. & Kuula, A. 2007. Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Yhteiskuntatieteellisen Tietoarkiston julkaisuja 6, 2007. Luettu 30.3. Saatavilla: http://urn.fi/urn:isbn:978-951-44-6999-2

Borg, S. 2014. Työmaana tutkimusdatan avoimuus. Signum, (6). Luettu 30.3. Saatavilla: https://journal.fi/signum/article/view/40769

Bruce, T. & Hillmann, D. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. Metadata in Practice. ALA Editions. Luettu 24.3. Saatavilla: https://www.researchgate.net/publication/247818823_The_Continuum_of_Metadata_Quality_Defining_Expressing_Exploiting

Buck, S. Solving reproducibility. Science, (348), 6242, s. 1403. DOI: 10.1126/science.aac8041 Saatavilla: https://science.sciencemag.org/content/348/6242/1403

Bui, Y., & Park, J.-R. (2006). An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository. Proceedings of the Annual Conference of CAIS. Luettu 29.4. Saatavilla https://www.researchgate.net/publication/28674964_An_Assessment_of_Metadata_Quality_A_Case_Study_of_the_National_Science_Digital_Library_Metadata_Repository

Business Finland. 2018. Business Finland edellyttää avointa tieteellistä julkaisemista. Luettu 20.3. Saatavilla https://www.businessfinland.fi/globalassets/finnish-customers/01-funding/04-research-organization/business-finland-edellyttaa-avointa-tieteellista-julkaisemista.pdf

Castro, J., Rocha da Silva, J. & Riberio, C. 2014. Creating Lightweight Ontologies for Dataset Description Practical Applications in a Cross-domain Research Data Management Workflow. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 10.1109/JCDL.2014.6970185. https://www.researchgate.net/publication/281436453_Creating_lightweight_ontologies_for_dataset_description_Practical_applications_in_a_cross-domain_research_data_management_workflow

Castro, J., Perrotta, D., Amorim, R., Rocha da Silva. & Ribeiro, C. 2015. Ontologies for Research Data Description: A Design Process Applied to Vehicle Simulation. 10.1007/978-3-319-24129-6_30. Luettu 10.9. Saatavilla: https://www.researchgate.net/publication/287782285_Ontologies_for_Research_Data_Description_A_Design_Process_Applied_to_Vehicle_Simulation

cOALition S. 2019. Accelerating the transition to full and immediate Open Access to scientific publications. Luettu 16.6. Saatavilla: https://www.coalition-s.org/wp-content/uploads/PlanS_Principles_and_Implementation_310519.pdf

Council of the European Union. 2016. Council conclusions on the transition towards an open science system. Luettu 22.3. Saatavilla http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

Corrall, S., Kennan, M-A. & Afzal, W. 2013. Bibliometrics and Research Data Management Services: Emerging trends in Library Support for Research. Library Trends. Luettu 28.3. Saatavilla http://d-scholarship.pitt.edu/18948/

Corrall, S. 2014. Designing Libraries for Research Collaboration in the Network World: An Exploratory Study. *LIBER* Quarterly, 24(1), pp.17–48. Luettu 28.3. Saatavilla: http://doi.org/10.18352/lq.9525

Cox, A. & Pinfield, S. 2014. Research data management and libraries: Current activities and future priorities. J Librariansh Inf Sci.

Davis, P. 2011. Open access, readership, citations: a randomized controlled trial of scientific journal publishing. FASEB Journal. (25) 2129–2134. Luettu 26.3. Saatavilla: https://www.fasebj.org/doi/full/10.1096/fj.11-183988?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed

Digital Curation Centre. 2007. What are metadata standards. Luettu 26.3. Saatavilla: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards

Digital Curation Centre. 2018. General Research Data. Luettu 25.3. Saatavilla http://www.dcc.ac.uk/resources/subject-areas/general-research-data

Ding, W. 2010. The Impact of Information Technology on Academic Scientists' Productivity and Collaboration Patterns. Management Science, 56(9), 1439-1461.

Dodd, S. 1982. Cataloguing machine-readable data files. An interpretive manual. Chicago, Illinois, American Library Association, 1982.

Dozier, J., Alexander, S., Courain, M., Dutton, J., Emery, W. & Gritton, B. 1995. Preserving scientific data on our physical universe: A new strategy for archiving the nation's scientific information resources. Washington, D.C.: National Academy Press.

Farnham, A ym. 2017. Early career researchers want Open Science. Genomy Biology 18:221. Luettu 24.3. Saatavilla: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1351-7

Fecher B., Friesike S. & Hebing M. 2015. What Drives Academic Data Sharing?. PLoS ONE 10(2) Luettu 30.3. Saatavilla doi:10.1371/journal.pone.0118053.

Freudenberg, M., Brümmer, M., Rücknagel, J., Ulrich, R., Eckart, T., Kontokostas, D. & Hellmann, S. 2016. The Metadata Ecosystem of DataID. Teoksessa (Toim. Garoufallou E., Subirats Coll I., Stellato A. & Greenberg J) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672, s. 317-332. Springer, Cham. Saatavilla: https://svn.aksw.org/papers/2016/MSOR_DataID2/public.pdf

Edwards, P., Mayernik, M., Batcheller, A., Bowker, G. & Borgman, C. 2011. Science friction: Data, metadata, and collaboration. Social Studies of Science, 41(5), pp. 667-

690. doi:10.1177/0306312711413314 Saatavilla: https://www.researchgate.net/publication/51874125_Science_Friction_Data_Metadata_and_Collaboration

Enwald, H. (2018). Tutkimusaineiston avoin jakaminen – tutkimusorganisaatioiden jäsenten käsityksiä, kokemuksia ja mielipiteitä. Informaatiotutkimus, *37*(4). Luettu 30.3. Saatavilla: https://doi.org/10.23978/inf.77411

ERC Scientific Council. 2018. Open Research Data and Data Management Plans, version 2.0. Luettu 18.3. Saatavilla https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf

Europeana. 2015. Report and Recommendations from the Task Force on Metadata Quality. Luettu 31.3. Saatavilla: https://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf

European Commission Expert Group. 2018. Final Report and Action Plan from the European Commission Expert Group on FAIR Data. Luettu 27.3. Saatavilla: https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

EOSC. 2017. European Open Science Declaration. 2017. Saatavilla: https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

Fairdata. 2018. Fair-periaatteet. Fairdata.fi. Luettu 19.3. Saatavilla https://www.fairdata.fi/miksi-fairdata/fair-periaatteet/

Force11. 2016. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0. Luettu 25.3. Saatavilla https://www.force11.org/fairprinciples

Forsman, M., & Englund, J. (2014). Altmetriikka – bibliometriikan uusi suuntaus. Signum, (6). Saatavilla: https://journal.fi/signum/article/view/40768

Francisco-Revilla, L., Trace, C., Li, H. & Buchanan, S. 2014. Encoded Archival Description: Data Quality and Analysis. Proceedings of the American Society for Information Science and Technology (51:1), 1-4.

Gavrilis, D., Makri, D-N., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C. & Constantopoulos, P. 2015. Measuring Quality in Metadata Repositories. Teoksessa Research and Advanced Technology for Digital Libgaries (toim. Sarantos, K., Mazurek, C. & Werla, M.). Springer International Publishing, Cham. Luettu 26.8. Saatavilla: https://link.springer.com/book/10.1007%2F978-3-319-24592-8

Gilliland, AJ. 2008. Setting the Stage. Teoksessa Introduction to metadata 2nd ed. (toim. Baca, M). Los Angeles, CA: Getty Research Institute, 2008. Luettu 22.7. Saatavilla http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf

GoFair. 2018. R1.3: (Meta)data meet domain-relevant community standards. Luettu 25.3. Saatavilla https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/

Greenberg, J., White, H., Carrier, S. & Scherle, R. 2009. A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata, 9(3-4), pp. 194-212. doi:10.1080/19386380903405090

Haffar, S., Bazerbachi, F. & Murad, H. 2019. Peer Review Bias: A Critical Review. Mayo Clinic Proceedings. Luettu 25.3. Saatavilla https://www.mayoclinicproceedings.org/article/S0025-6196(18)30707-9/fulltext#sec3.5

Heikkilä, T. 2008. Tilastollinen tutkimus. Helsinki: Edita Oy.

Housewright, R., Schonfeld, R.C. & Wulfson, K. 2013a. Ithaka S+R US faculty survey 2012. New York: Ithaka S+R. Luettu 28.3. Saatavilla http://www.sr.ithaka.org/research-publications/us-faculty-survey-2012

Housewright, R., Schonfeld, R. & Wulfson, K. .2013b. Ithaka S+R/JISC/RLUK US survey of academics 2012. New York: Ithaka S+R. Luettu 28.3. Saatavilla: http://sr.ithaka.org/research-publications/ithaka-sr-jisc-rluk-uk-survey-academics-2012

H2020 Programme. 2016. Guidelines on FAIR Data Management in Horizon 2020. Horizon 2020. Luettu 20.3. Saatavilla http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

H2020 Programme. 2017. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Horizon 2020. Luettu 20.3. Saatavilla http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Holopainen, M. & Pulkkinen, P. 2008. Tilastolliset menetelmät. 5. uud. p. Porvoo ; Helsinki: WSOY Oppimateriaalit.

ISO 15836-1. 2017. Information and documentation -- The Dublin Core metadata element set -- Part 1: Core elements. International Organization for Standardization.

ISO/DIS 15836-2. 2019. Information and documentation — The Dublin Core metadata element set — Part 2: DCMI Properties and classes. International Organization for Standardization. Luettu 5.7. Saatavilla: https://www.iso.org/obp/ui#iso:std:iso:15836:-2:dis:ed-1:v1:en

Jaguszewski, J.M., & Williams, K. (2013). New roles for new times: Transforming liaison roles in research libraries. Washington, DC: Association of Research Libraries. Luettu 28.3. Saatavilla: https://www.arl.org/publications-resources/2893-new-roles-for-new-times-transforming-liaison-roles-in-research-libraries#.XJyk46RS-uU

Jytilä, R., & Laakso, M. 2019. Avoin tiede muuttaa vertaisarvioinnin käytäntöjä. Vastuulline tiede – tutkimusetiikka ja tiedeviestintä Suomessa. Luettu 27.3. Saatavilla: https://www.vastuullinentiede.fi/fi/julkaiseminen/avoin-tiede-muuttaa-vertaisarvioinnin-kaytantoja

Kaipainen, T. 2018. Datan elinkaari haltuun ja FAIR-periaatteet. Kreodi. Luettu 27.3. Saatavilla: https://www.kreodi.fi/en/25/Matkakertomuksia/472/Datan-elinkaari-haltuun-ja-FAIR-periaatteet.htm

Karimova, Y ym. 2017. Promoting Semantic Annotation of Research Data by Their Creators: A Use Case with B2NOTE at the ENd of the RDM Workflow. Teoksessa Metadata and Semantic Research 2017 (toim. Garoufallou, E., Virkus, S., Siatri, R. & Koutsomiha, D), s. 112-122. Springer International Publishing, Cham.

Ketokivi, M. 2015. Tilastollinen päättely ja tieteellinen argumentointi. 2. laaj. laitos. [Helsinki]: Gaudeamus.

Kim, J. 2013. Data sharing and its implications for academic libraries. New Library

World, (114)494–506. Luettu 29.3. Saatavilla: https://doi.org/10.1108/NLW-06-2013-0051.

Király, P. & Büchler, M. 2018. Measuring Completeness as Metadata Quality Metric in Europeana. Luettu 26.8. Saatavilla http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2018/12/7.Kiraly.pdf

Koltay, T. 2016. Are you read? Tasks and roles for academic libraries in supporting Research 2.0. New Library World 117 (1-2), 94-104. Luettu 30.3.

Koski, H. 2017. Avoimen datan hyödyntäminen ja vaikuttavuus. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 40/2017. Valtioneuvoston kanslia. Luettu 9.7. Saatavilla: https://www.etla.fi/julkaisut/avoimen-datan-hyodyntaminen-ja-vaikuttavuus/

Laine, H. 2018a. Avoimen julkaisemisen Plan S: paljon voimaa, pieniä heikkouksia. Think open -blogi. Luettu 12.3. Saatavilla: **https://blogs.helsinki.fi/thinkopen/plan-s/**

Laine, H (toim.). 2018b. Tracing Data - Data Citation Roadmap for Finland. Helsinki, Finland: Finnish Committee for Research Data. Luettu 29.3. Saatavilla: http://urn.fi/URN:NBN:fi-fe201804106446

Laine, H., & Nykyri, S. (2018). Dataviittaamisen tiekartta tutkijalle. Informaatiotutkimus, *37*(2). Luettu 29.3. Saatavilla: https://doi.org/10.23978/inf.72999

Lee, C., Sugimoto, C., Zhang, G. & Cronin, B. 2013. Bias in Peer Review. Journal of the American Society for Information Science & Technology 64(1):2-17. Luettu 27.3. Saatavilla: https://www.researchgate.net/publication/260409966_Bias_in_peer_review

Lei Zeng, M. & Qin, J. 2016. Metadata. Second edition. London: Facet Publishing.

Lowndes, J., Best, B., Scarborough, C., Aflerbach, J., Frazier, M., O'hara, C., Jiang, N. & Halpern, S. 2017. Our path to better science in less time using open data science tools. *Nature Ecology & Evolution, 1*(6). doi:10.1038/s41559-017-0160 Saatavilla: https://www.nature.com/articles/s41559-017-0160/#ref6

Manchikanti, L., Kaye, A. D., Boswell, M., & Hirsch, J. A. 2015. Medical journal peer review: Process and bias. Pain Physician, 18(1), E1-E14. Luettu 25.3. Saatavilla http://www.painphysicianjournal.com/linkout?issn=1533-3159&vol=18&page=E1

Matthews, B. 2010. Using a Core Scientific Metadata Model in Large-Scale Facilities. International Journal of Digital Curation, 5(1), pp. 106-118. Luettu 5.7. Saatavilla: https://www.semanticscholar.org/paper/Using-a-Core-Scientific-Metadata-Model-in-Matthews-Sufi/77d0d5b917361933676876354f47e0656df66e1a

Margaritopoulos, M.; Margaritopoulous, T., Mavridis, I. & Manitsaris, A. 2012. Quantifying and measuring metadata completeness. Journal of the American Society for Information Science and Technology, 63(4), pp. 724-737. doi:10.1002/asi.21706

Mayernik, M. 2011. Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. Saatavilla: https://pdfs.semanticscholar.org/1fab/bf61e1ad05ee2fcd424681da61a78f3d82af.pdf

Mayernik, M., Batcheller, A. & Borgman, C. 2011. How Institutional Factors Influence the Creation of Scientific Metadata. Proceedings'11 iConference, s. 417-425. ACM: Nye York. DOI: 10.1145/1940761.1940818

Mayernik, M. 2019. Metadata accounts: Achieving data and evidence in scientific research. Social Studies of Science, 49(5), s. 732-757 Saatavilla: doi:10.1177/0306312719863494

McKiernan E. ym. 2016. How open science helps researchers succeed. eLife. 2016:5. Luettu 25.3. Saatavilla http://www.ncbi.nlm.nih.gov/pmc/articles/4973366

Metsämuuronen, J. 2002a. Tilastollisen kuvauksen perusteet. 2. uud. p. Helsinki: International Methelp.

Metsämuuronen, J. 2002b. Mittarin rakentaminen ja testiteorian perusteet. 2. uud. p. Helsinki: International Methelp.

Moen, W., Stewart, E. & McClure, C. 1998. Assessing metadata quality: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS). Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries, 246-255. Santa Barbara: CA, USA,

1998

doi: 10.1109/ADL.1998.670425

MOT IT-Ensyklopedia. 2019. Kielikone Oy. HarperCollins. Haettu hakusanalla ”mapping” 9.7.2019 Saatavilla: https://mot.kielikone.fi/finelib/netmot.shtml

Munafò, M., Nosek, B., Bishop, D., Button, K. Chambers, C., Percie du Sert, N. Simonsohn, U., Wagenmakers, E-J. Ware, J. & Ioannidis, J. 2017. A manifesto for reproducible science. Nature Human Behaviour, 1(0021). Luettu 24.3. Saatavilla: https://www.nature.com/articles/s41562-016-0021

Munthe, C. & Welin, S. 1996. The morality of scientific openness. Science and Engineering ethics 2(4), 411-428. https://doi.org/10.1007/BF02583928

National Information Standards Organization. 2007. A Framework of Guidance for Building Good Digital Collections. 3.p. Luettu 2.7. Saatavilla https://www.niso.org/sites/default/files/2017-08/framework3.pdf

National Institutes of Health. 2019. Big Data to Knowledge – Program Snapshot. National Institues of Health. Luettu 20.3. Saatavilla **https://commonfund.nih.gov/bd2k**

National Science Foundation. 2011. Digital Research Data Sharing and Management. Luettu 24.3. Saatavilla http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

Neumaier, S., Umbrich, J. & Polleres, A. 2016. Automated Quality Assessment of Metadata across Open Data Portals. ACM Journal of Data and Information Quality. (JDIQ), 8(1), s. 1-29. Luettu 9.7. Saatavilla https://aic.ai.wu.ac.at/~polleres/publications/neum-etal-2016JDIQ.pdf

Nosek, B. A. 2015. SCIENTIFIC STANDARDS. Promoting an open research culture. Science (New York, N.Y)., 348(6242), p. 1422. doi:10.1126/science.aab2374 Saatavilla: https://www.researchgate.net/publication/279302015_Promoting_an_Open_Research_Culture

Nummenmaa, L. & Kimpimäki, K. 2014. Tilastollisten menetelmien perusteet. 1. p. Helsinki: Sanoma Pro.

Nygård, A-J. 2018. Tutkijan henkilökohtainen ORCID-tunniste. Vastuullinen tiede – tutkimusetiikka ja tiedeviestintä Suomessa. Saatavilla: https://www.vastuullinen-tiede.fi/fi/julkaiseminen/tutkijan-henkil%C3%B6kohtainen-orcid-tunniste Viitattu 12.12.2019

Pasquetto, I., Randles, B. & Borgman, C. 2017. On the Reuse of Scientific Data. Data Science Journal, 16(1), . doi:10.5334/dsj-2017-008 Saatavilla: https://datascience.codata.org/articles/10.5334/dsj-2017-008/

Ochoa, X. & Duval, E. 2006. Quality Metrics for Learning Object Metadata. Teoksessa E. Pearson & P. Bohman (toim)., Proceedings of EdMedia + Innovate Learning 2006 (s. 1004-1011). Waynesville, NC: Association for the Advancement of Computing in Education (AACE).

Ochoa, X. & Duval, E. 2009a. Metadata Quality. Teoksessa Handbook of Metadata, Semantics and Ontologies (toim. Sicilia, M). Singapore ; Hackensack, N.J.: World Scientific Pub. Co.

Ochoa, X. & Duval, E. 2009b. Automatic evaluation of metadata quality in digital repositorities. International Journal on Digital Libraries 10(2-3), 67-91Luettu 31.3. Saatavilla: https://doi.org/10.1007/s00799-009-0054-4

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

Ogungbeni, J., Obiamalu, A., Ssemambo, S., & Bazibu, C. 2016. The roles of academic libraries in propagating open science: A qualitative literature review. Information Development, *34*(2), 113–121. Luettu 28.3. Saatavilla https://doi.org/10.1177/0266666916678444

Opetus- ja kulttuuriministeriö. 2016. Avointa tiedettä edistettävä kaikin keinoin. Luettu 23.3. Saatavilla: https://minedu.fi/artikkeli/-/asset_publisher/open-science-must-be-promoted-by-all-means-necessary

Opetus- ja kulttuuriministeriö. 2014. Avoimen tieteen ja tutkimuksen strategiaryhmä ja asiantuntijaryhmä. Luettu 22.3. Saatavilla https://minedu.fi/hanke?tunnus=OKM031:00/2014

Qin, J. Ball, A. & Greenberg, J. 2012. Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. Proceedings of DCMI International Conference on Dublin Core and Metadata Applications, Sep. 2012. Luettu 11.7. Saatavilla: http://dcpapers.dublincore.org/pubs/article/view/3660/1883

Qin, J. & Li, K. 2013. How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. Teoksessa (toim. M, Foulonneau & K, K. Eckert Proceedings of the DC-2013 Conference, s. 25-34, 2013.

Palavitsinis, N. 2013. Metadata Quality Issues in Learning Repositories. Väitöskirja, Alcala de Henares, Spain. Luettu 11.8. Saatavilla: https://core.ac.uk/download/pdf/58910780.pdf

Palavitsinis, N., Manouselis, Nikos & Sanchez-Alonso, S. 2014. Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process. Journal of the Association for Information Science and Technology, 65(6), pp. 1202-1216. doi:10.1002/asi.23045Park, J-R. 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly. Taylor & Francis.

Piwowar, H. Day, R. & Fridsma, D. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. & Chapman, W. 2010. Public sharing of research datasets: a pilot study of associations. Journal of informetrics, 4(2), 148–156. doi:10.1016/j.joi.2009.11.010

Piwowar, H. Priem, J. & Larivière, V. ym. 2018. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ — the Journal of Life and Environmental Sciences (6) e4375. Luettu 26.3. Saatavilla: doi:10.7717/peerj.4375

Pomerantz, J. 2015. Metadata. Cambridge, Massachusetts ; London, England: The MIT Press.

Ottaviani, J. 2016. The Post-Embargo Open Access Citation Advantage: It Exists (Probably), Its Modest (Usually), and the Rich Get Richer (of Course). PLoS One 11(8):e0159614. Luettu 26.3. Saatavilla doi:10.1371/journal.pone.0159614

Rantasaari, J., & Kanerva, P. 2017. ”A Change Is Gonna Come” – Avoimen tieteen palveluita rakentamassa Turun yliopiston kirjastossa. Signum, 49(3), 25–29. Luettu 25.3. Saatavilla: https://doi.org/10.25033/sig.68833

Research Data Alliance. 2017. Standards. Luettu 25.3 Saatavilla http://rd-alliance.github.io/metadata-directory/standards/

Rew, R. & Davis, G. 1990. NetCDF: An interface for scientific data access. IEEE Computer Graphics and Applications, 10(4), pp. 76-82. doi:10.1109/38.56302

Riley, J. 2017. Understandint metadata – what is metadata, and what is it for?. A Primer Publication of the National Information Standards Organization. Baltimore: National Information Standards Organization (NISO). Luettu 1.7. Saatavilla https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

Rockey S. 2012. Revised Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research. National Institutes of Health. Luettu 24.3. Saatavilla http://nexus.od.nih.gov/all/2012/11/16/improving-public-access-to-research-results/

Ross-Hellauer T. 2017. What is open peer review? A systematic review. F1000Res. 2017; 6: 588. Luettu 25.3. Saatavilla https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437951/

Rousidis, D., Garoufallou, E., Balatsouskas, P. & Sicilia, M-A. 2014a. Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. Information Services & Use 34 (2014) 279-286. Luettu 25.3.2019 Saatavilla https://content.iospress.com/articles/information-services-and-use/isu746

Rousidis, D., Garoufallou, E., Balatsouskas, P. & Sicilia, M-A. 2014b. Data Quality Issues and Content Analysis for Research Data Repositories : The Case of Dryad. ELPUB2014. Let's put data to use: digital scholarship for the next generation, 18th International Conference on Electronic Publishing June 19-20, 2014, Thessaloniki, Greece. Luettu 15.3. Saatavilla: https://www.semanticscholar.org/paper/Data-Quality-Issues-and-Content-Analysis-for-Data-%3A-Rousidis-Siclia/7fa750ca24886df018e455ffe515c001d8f5d402

Rousidis D., Garoufallou E., Balatsoukas P. & Sicilia M-A. 2015 Evaluation of Metadata in Research Data Repositories: The Case of the DC.Subject Element. In: Garoufallou E., Hartley R., Gaitanou P. (toim). Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science, vol 544. Springer, Cham.

Royal Society. 2012. Science as an open enterprise: open data for open science. 2012. Luettu 24.3. Saatavilla: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.

Ruotsalainen, A. 2016. Metadatajen Laatu Julkaisuarkistoissa. Informaatiotieteiden yksikkö, School of Information Sciences, and University of Tampere. Saatavilla: http://tampub.uta.fi/bitstream/10024/100222/1/GRADU-1480943391.pdf

Sarewitz, D. 2016. The pressure to publish pushes down quality. Nature, 533(5), 147. Luettu 24.3. Saatavilla: https://www.nature.com/news/the-pressure-to-publish-pushes-down-quality-1.19887

Science Digital Library Metadata Repository. Teoksessa CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation (toim. Moukdad). Luettu 24.3. Saatavilla: http://hdl.handle.net/1860/1600

SFS 5895. 2001. Dublin Core -metadataformaatin suomalainen versio. Suomen Standardisoimisliitto SFS.

Shreeves, S., Palmer, C., Stvilia, B. & Twidale, M. 2005. Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practives for Federated Collections. Teoksessa Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005 (toim. Thompson, H.). Minneapolis: Minnesota. Luettu 25.3. Saatavilla: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/pdf/shreeves05.pdf

Sicilia, M. 2014. Handbook of metadata, semantics and ontologies. Singapore ; Hackensack, N.J.: World Scientific Pub. Co.

Silva F., Amorim R.C., Castro J.A., da Silva J.R. & Ribeiro C. 2016 End-to-End Research Data Management Workflows. Teoksessa (Toim. Garoufallou, E., Subirats, I., Stellato, A. & Greenberg, J). Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham.

Smaldino, P. & McElreath, R. 2016. The natural selection of bad science. Royal Society Open Science. 3(9). Luettu 24.3. Saatavilla https://royalsocietypublishing.org/doi/10.1098/rsos.160384

Smith R. 2006. Peer review: a flawed process at the heart of science and journals. Journal of the Royal Society of Medicine, 99(4), 178-82. Luettu 25.3. Saatavilla https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1420798/

Steen, G. 2011. Retractions in the scientific literature: is the incidence of research fraud increasing. Journal of Medical Ethics, 37, 249-53. Luettu 24.3. Saatavilla: https://jme.bmj.com/content/37/4/249.long

Stvilia, B. Ym. 2004. Metadata Quality for Federated Collections. Teoksessa Proceedings of ICIQ04 - 9th International Conference on Information Quality (toim. Talburt, J. ym.). Cambridge, Ma. Luettu 23.3. Saatavilla http://hdl.handle.net/2142/721

Stvilia, B., Gasser, L., Twidale, M. & Smith, L. 2007. A Framework for information quality assessment. Journal of the American Society for Information Science and Technology. 58(12), 1720-1733. Luettu 26.3. Saatavilla https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20652

Suomen Akatemia. 2019a. Aineistonhallintasuunnitelma. Luettu 22.10. 2019. Saatavilla: https://www.aka.fi/fi/rahoitus/hae-rahoitusta/ohjehakemisto/aineistonhallinta/

Suomen Akatemia. 2019b. Avoin tiede: avoin saatavuus ja aineistojen avaaminen. Luettu 2.7. Saatavilla https://www.aka.fi/avoin_tiede

Swan, A. (2010) The Open Access citation advantage: Studies and results to date. Technical Report. School of Electronics & Computer Science. Luettu 31.1. Saatavilla: https://eprints.soton.ac.uk/268516/

Tampereen korkeakouluyhteisö. 2019. Tampereen yliopiston avoimen tieteen ja tutkimuksen toimenpideohjelma. Saatavilla https://www.tuni.fi/sites/default/files/2019-02/tau-avoimen-tieteen-toimenpideohjelma.pdf

Tani, A., Candela, L. & Castelli, D. 2013. Dealing with metadata quality: The legacy of digital library efforts. Information Processing and Management. Elsevier.

Tennant, J. ym. 2017a. A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research, 6. Luettu 26.3. Saatavilla https://f1000research.com/articles/6-1151/v3

Tennant, J., Waldner, F., Jacques, D. ym. 2017b. The academic, economic and societal impacts of Open Access: an evidence-based review [version 3; peer review: 4 approved, 1 approved with reservations]. F1000Research 2016, **5**:632 Luettu 26.3. Saatavilla https://doi.org/10.12688/f1000research.8460.3

Tenopir, C., Birch, B. & Allard, S. 2012. Academic Libraries and Research Data Services - Current Practices and Plans for the Future. Association of College & Research Libraries. Luettu 28.3. Saatavilla

Tenopir, C., Sandusky, R., Allard, S .& Birch, B. 2014. Research data management services in academic research libraries and perceptions of librarians. Library and Information Science Research 36(2), 84-90. Luettu 30.3. Saatavilla: https://www.researchgate.net/publication/262489867_Research_data_management_services_in_academic_research_libraries_and_perceptions_of_librarians

Tenopir, C., Hughes, D., Allard, S., Frame, M., Birch, B., Baird. & Lundeen, A. 2015. Research data services in academic libraries: Data intensive roles for the future? Journal of eScience Librarianship, 4(2). https://doi.org/10.7191/jeslib.2015.1085

Tenopir, C. ym. 2017. Research Data Services in European Academic Research Libraries. The Journal of the Association of European Research Libraries. Luettu 28.3. Saatavilla http://urn.fi/URN:NBN:fi:uta-201704031396

Tenopir, C., Christian, L., Allard, S. & Borycz, J. 2018. Research Data Sharing: Practives and attitudes of geophysicits. Eart and Space Science (5), 891-902. Luettu 30.3. Saatavilla: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018EA000461

Tietoarkisto. 2015. Aineiston kuvailu ja metadata. Aineistonhallinnan käsikirja. Luettu 12.3. Saatavilla **https://www.fsd.uta.fi/aineistonhallinta/fi/aineiston-kuvailu-ja-metadata.html**

Tietoarkisto. 2017. Miksi aineistonhallintaa ja jatkokäyttöä. Aineistonhallinna käsikirja. Luettu 26.3. Saatavilla: https://www.fsd.uta.fi/aineistonhallinta/fi/miksi-aineiston-hallintaa-ja-jatkokayttoa.html

Tutkimuksen tietoaineistot. 2013. TTA-minimimetadatamalli. Luettu 30.8. Saatavilla: https://www.tdata.fi/documents/10180/44526/TTA-minimimetadatamalli/

UniFi. 2018. Avoin tiede ja data – toimenpideohjelma suomalaiselle tiedeyhteisölle. Luettu 25.3. Saatavilla: http://urn.fi/URN:NBN:fi-fe2018052424593

Valli, R. 2015. Johdatus tilastolliseen tutkimukseen. 2. uudistettu painos. Jyväskylä: PS-kustannus.

Vardigan, M. 2013. DDI Timeline. IASSIT Quarterly 371(4), 51-54. Luettu 22.10. Saatavilla https://iassistdata.org/sites/default/files/iqvol371_4_vardigan2.pdf

Waaijers, L., & van der Graaf, M. 2011. Quality of research data, an operational approach. D-Lib Magazine, 17(1/2). Luettu 26.8. Saatavilla: http://www.dlib.org/dlib/january11/waaijers/01waaijers.print.html

Welle Donker, F. & Van Loenen, B. 2016. How to Assess the success of the open data ecosystem. International Journal of Digital Earth (10:3), 284-306. Taylor & Francis. Luettu 9.7. Saatavilla: https://www.tandfonline.com/doi/full/10.1080/17538947.2016.1224938

Vilar, P. & Zabukovec, V. 2019. Research data management and research data literacy in Slovenian science", Journal of documentation 75(1), 24-43. Luettu 30.3. Saatavilla: https://www.emeraldinsight.com/doi/abs/10.1108/JD-03-2018-0042

Whyte, A. & Tedds, J. 2011. Making the case for research data management. Edinburgh: Digital Curation Centre. Luettu 26.3. Saatavilla http://www.dcc.ac.uk/webfm_send/487.

Wilkinson, M.D. ym. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Nature - Scientific Data. Luettu 28.3. Saatavilla: https://www.nature.com/articles/sdata201618

Yoon, A. & Schultz, T. 2017. Research Data Management Services in Academic Libraries in the US: A Content Analysis of Libraries' Websites . College & Research Libraries. Luettu 28.3. Saatavilla https://crl.acrl.org/index.php/crl/article/view/16788/1834