Tampereen yliopisto

Vikke Niemi

YHTEISKUNTATIETEELLISTEN TUTKIMUSAINEISTOJEN JATKOKÄYTTÖ

Tilastollinen analyysi tutkimusaineistoihin kohdistuvasta mielenkiinnosta ja jatkokäytöstä

Informaatioteknologian ja viestinnän tiedekunta

Pro gradu -tutkielma

Lokakuu 2021

TIIVISTELMÄ

Vikke Niemi: Yhteiskuntatieteellisten tutkimusaineistojen jatkokäyttö: Tilastollinen analyysi tutkimusaineistoihin kohdistuvasta mielenkiinnosta ja jatkokäytöstä
Pro Gradu -tutkielma
Tampereen yliopisto
Informaatiotutkimuksen maisteriohjelma
Lokakuu 2021

Osana avoimen tieteen liikettä on ajatus ja konsepti tutkimusaineistojen avoimesta saatavuudesta, joka on viimeisen 10 vuoden aikana konkretisoitunut osaksi tieteen tekemisen kenttää. Tutkimusaineistoille on pystytetty lukuisia niin tieteenalakohtaisia kuin yleisiäkin sähköisiä arkistoja, joihin aineistoja voidaan tallentaa. Arkistoissa tutkimusaineistot ovat esillä ja saavutettavissa, ja aineistot ovat arkistosta riippuen laadultaan sellaisessa tilassa, joka sallii niiden jatkokäytön.

Tarkastelen tutkimuksessa yhteiskuntatieteellisten tutkimusaineistojen jatkokäytön astetta määrällisten menetelmien avulla. Näytteenä ovat Swedish National Data Service (SND) -tutkimusaineistoinfrastruktuurin ylläpitämät yhteiskuntatieteelliset tutkimusaineistot. Aineisto koostuu SND:n aineistojen kuvailutietojen metadatatietueista, aineistojen latausten lokitiedostosta vuosilta 2015-2021, Event Data -palvelun tapahtumadatasta sekä Google Scholarista ja Web of Sciencestä kerätyistä viittaustiedoista. Aineistojen saamien latausmäärien avulla tarkastelen aineistoihin kohdistuvaa mielenkiintoa, ja viittauksien avulla selvitän aineistojen jatkokäyttöä. Tarkastelen viittauksia myös niiden laadun ja lähteen näkökulmasta.

Vuosina 2015-2021 SND:stä ladattiin tutkimusaineistoja 23896 kertaa. Siinä missä rajatusti avointen tutkimusaineistojen latausjakauma oli suhteellisen tasainen, olivat täysin avointen tutkimusaineistojen latausmäärät kasvaneet merkittävästi viimeisen kolmen vuoden aikana. Viittaukset kohdistuivat lähes yksinomaan rajoitetusti avoimiin tutkimusaineistoihin, ja viittaustavat vaihtelivat epävirallisemmista maininnoista tarkkoihin lähdeluetteloon merkittyihin viittauksiin.

Tutkimuksen tulokset mukailevat rajoitteidensa puitteissa aiempaa tutkimusta yhteiskuntatieteellisten tutkimusaineistojen jatkokäytöstä. Tulokset osoittavat, että mielenkiinto avoimia tutkimusaineistoja kohtaan on kasvanut, ja näytteen osalta tutkimusaineistojen jatkokäyttö on kasvamassa, etenkin opinnäytteiden yhteydessä. Tulokset indikoivat myös muutoksesta tutkimusaineistojen viittauskäytänteissä.

Avainsanat: tutkimusaineisto, avoimet tutkimusaineistot, tutkimusaineistojen jatkokäyttö, tutkimusaineistojen uudelleenkäyttö, data reuse, avoin tiede

SISÄLLYSLUETTELO

1 JOHDANTO 1

2 TUTKIMUSAINEISTOT OSANA AVOINTA TIEDETTÄ 3

2.1 Tutkimusaineisto 3
2.2 Avoin tiede 4
2.3 Avoimet tutkimusaineistot 5
2.4 Tutkimusaineistojen tieteenalaeroista 9

3 TUTKIMUSAINEISTOJEN JULKAISEMINEN JA JAKAMINEN 11

3.1 Julkaisun ja jakamisen tapoja 11
3.2 Julkaisemisen ja jakamisen edellytykset 12
3.2.1 Metadata 13
3.2.2 Infrastruktuuri 14
3.2.3 Pysyväistunnisteet 16
3.3 Toimintaympäristön ja käytänteiden vaikutus 17

4 TUTKIMUSAINEISTOJEN JATKOKÄYTÖN MITTAAMINEN 22

4.1 Tutkimusaineistojen jatkokäytön määritelmä 22
4.2 Tutkimusaineistojen vaikuttavuus 23
4.3 Tutkimusaineistoihin viittaaminen 25
4.4 Tutkimusaineistojen lataaminen 26
4.5 Tutkimusaineistojen jatkokäyttö yhteiskuntatieteissä 26

5 TUTKIMUSASETELMA JA TOTEUTUS 29

5.1 Tutkimuskysymykset 29
5.2 Swedish National Data Service (SND) 31
5.3 DataCite (& Crossref) 32
5.4 Tutkimusaineisto 34
5.4.1 Aineiston keruu 35
5.4.2 Aineiston käsittely 37
5.5 Tutkimusmenetelmä 43

6 TULOKSET 44

7 DISKUSSIO 56

8 PÄÄTELMÄT 61

LÄHTEET 63

1 JOHDANTO

Tarkastelen tässä tutkimuksessa ruotsalaisen Swedish National Data Service (SND) -tutkimusaineistoinfrastruktuurin ylläpitämiin yhteiskuntatieteellisiin tutkimusaineistoihin kohdistuvaa mielenkiintoa ja jatkokäyttöä. Mielenkiintoa kartoitetaan aineistojen saamien latausmäärien perusteella aikavälillä 2015-2021, ja jatkokäyttö määrittyy aineistojen saamien viittausten perusteella. Tutkimuskysymykset ja -asetelma kokonaisuudessaan esitellään tarkemmin luvussa 5.

Tutkimusaineistojen julkaisemiselle, jakamiselle ja jatkokäytölle on keskeistä tutkimusaineistojen avoimuuden käsite. Tiedepoliittisella tasolla avointen tutkimusaineistojen voidaan katsoa saaneen alkunsa vuonna 2003 niin kutsutussa Berliinin julistuksessa, ”Berlin Declaration on Open Access to Knowledge in the Science and Humanities”. Aiemmin avoimen saatavuuden (open access) fokus on ollut tieteellisissä julkaisussa mutta Berliinin julistus huomioi myös tutkimusaineistot. (Pampel & Dallmeier-Tiessen 2014, 214.) Ajatus on edelleen pitkälti sama, kuin mistä se lähti liikkeellekin: julkisin varoin tuotetut tieteelliset tuotokset tulisi olla avoimia, kaiken kansan saatavilla. Tutkimusaineistojen avoimuuden voi nähdä nykyisellään konkretisoituneen suuremmin osaksi tieteen kenttää 2010-luvun tienoilla, kun merkittävät tutkimusrahoittajat ryhtyivät edellyttämään rahoitusten hakijoilta suunnitelmia heidän tulevien tutkimusten tutkimusaineistojen laadinnasta, ylläpidosta ja pitkäaikaissäilytyksestä. (Borg 2010; Burwell et al. 2013; Nuorteva 2008; OECD 2007.)

Nämä aineistonhallintasuunnitelmat ovat tätä nykyä arkipäivää, mutta siltikin aineistojen julkaisutavat eroavat aloittain, kuin myös ajatus niiden jatkokäytöstä. Tieteenalat eroavat toisistaan tutkimuksen tekemisen tavoiltaan, mikä vaikuttaa syntyvien tutkimusaineistojen formaattiin, tyyppiin ja kokoon. Tieteenalojen erot vaikuttavat myös tutkimusaineistojen synnyn nopeuteen, mikä vaikuttaa ylipäänsä alalla olevien aineistojen saatavuuden määrään. (Borgman 2015, 55-56,79.)

Niin ikään aineistojen jatkokäytön ja julkaisun yhteyteen liittyvät käytänteet, suositukset, politiikat ja esimerkiksi metadatastandardit ovat jatkuvassa kehityksessä. Tutkimusaineistojen pitkäaikaissäilytyksen ja jatkokäytön mahdollistavia sähköisiä arkistoja on lukuisia, ja aineistojen pirstaloituessa näiden välille hankaloituu myös niiden löytäminen ja tätä kautta myös jatkokäyttö (Wilkinson et al. 2016.). Tutkimusaineistojen jatkokäyttö on kehittyneintä luonnontieteellisillä aloilla, kun taas yhteiskuntatieteet sekä humanistiset alat ovat kokonaisuudessaan tutkimusaineistojen julkaisussa ja jatkokäytössä vasta pääsemässä vauhtiin (Borgman 2015, 281).

Tästä syystä onkin mielekästä tutkia, miltä näyttää yhteiskuntatieteellisten tutkimusaineistojen osakseen saama mielenkiinto sekä niiden jatkokäyttö nyt, SND:n rajaaman näytteen valossa.

Tarkastelen ensin luvussa 2 tutkimuksen aihepiirin peruskäsitteitä ja niiden kehityskaaria. Luvussa 3 perehdyn tutkimusaineistojen julkaisemiseen ja jakamiseen, ja luvussa 4 keskityn aineistojen jatkokäyttöön sen määrittelyn, mittaamisen ja aiemman tutkimuksen kautta. Luvussa 5 esittelen tutkimusasetelman ja tutkimuksen toteutuksen. Luku 6 kattaa tutkimustulokset, ja luvussa 7 syvennyn tuloksiin diskussion muodossa. Luku 8 kattaa loppupäätelmät.

2 TUTKIMUSAINEISTOT OSANA AVOINTA TIEDETTÄ

Tarkastelen tässä luvussa tutkimusaineistojen julkaisemisen, jakamisen ja jatkokäytön kannalta peruskäsitteistöä ja kehityskaaria, jotka auttavat ymmärtämään tutkielman viitekehystä. Luvussa 2.1 esittelen tutkimusaineiston käsitteen, luvussa 2.2 tarkastelen avointa tiedettä pääpiirteiltään, luvussa 2.3 avaan avointen tutkimusaineistojen syntyä ja kehitystä, ja viimeisessä luvussa 2.4 tarkastelen lyhyesti tutkimusaineistojen tieteenalaeroja.

2.1 Tutkimusaineisto

Suomeksi puhuttaessa tutkimuksen aikana syntyvästä tai tutkimusta varten käytettävästä aineistosta käytetään termejä kuten tu*tkimusaineisto, aineisto, tutkimusdata* ja *data* (ks. Turun yliopisto n.d.; Helsingin yliopisto n.d.). Tätä lukua lukuun ottamatta, käytän tutkielmassa enimmäkseen käsitteitä tutkimusaineisto ja aineisto selkeyden ja luettavuuden vuoksi. Aihepiirin kirjallisuudessa käytetään englanniksi usein vain monitulkintaista termiä *data*. Myös tarkempaa määritelmää *research data* on käytetty (ks. Tenopir 2015).

Termin data historia ulottuu 1600-luvulle teologian yhteyteen, jossa termiä käytettiin monikossa (Borgman 2015, 17). Myös Rosenbergin (2013) analyysi termin käyttöasteesta osoittaa selkeää kasvua 1600-luvulta eteenpäin. Termin aikaisimmat käyttötapaukset olivat latinaksi, ja ”data” tuli terminä osaksi englannin kieltä matematiikan ja teologian kautta. Tuolloin ei ollut yhtenäistä mielipidettä siitä, tulisiko sanaa käyttää yksikössä vai monikossa. Dataa käytettiin kuvaamaan joukkoa hyväksyttyjä periaatteita, jotka toimivat argumentin pohjana, sekä erityisesti sellaisina faktoina, jotka on otettu kirjoitetusta tekstistä (scripture). Vasta 1700-luvun loppupuolella datalla ryhdyttiin viittaamaan tieteellisiin faktoihin, jotka on saavutettu kokeilla, havainnoilla ja muilla tutkimuksilla. (Rosenberg 2013.)

Ainakaan englanninkielisessä maailmassa ei Borgmanin (2015, 28) mukaan ole vieläkään yhtä ja oikeaa vakiintunutta määritelmää datalle eli tutkimusaineistolle. Tyydyn käyttämään tämän tutkielman puitteissa Borgmanin (2015, 28) koostavaa määritelmää: "Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship." Harvoin voidaan määritellä mitään tiettyä hetkeä, jolloin asioista tulee tutkimusaineistoa. Prosessi, jossa asiat muuttuvat tutkimusaineistoksi, edellyttää tyypillisesti tutkijan, joka tunnistaa havainnon, objektin, luettelon tai jonkin muun entiteetin, jota voisi käyttää todisteena jostakin ilmiöstä. Sitten tutkija kerää, hankkii, esittää, analysoi ja tulkitsee näitä entiteettejä tutkimusaineistona. Entiteetit voivat olla oikeastaan mitä tahansa, esimerkiksi tekstiä, valokuvia tai vaikkapa täytettyjä digitaalisia kyselylomakkeita. (Borgman 2015, 62.)

Tutkimusaineistot ovat samankaltaisia arkistojen aineistojen kanssa: jokainen aineisto on uniikki ja vaatii oman metadatansa sekä tiedot sen alkuperästä. Aineistot voivat olla tutkijoiden itsensä keräämiä, ne voivat olla jo olemassa olevaa dataa jostain ilmiöstä mutta useimmiten tutkijat yhdistävät vanhaa ja uutta aineistoa luodakseen tarvitsemansa aineiston. Tutkimusaineistot voivat myös olla missä tahansa muodossa ja formaatissa, ja minkä tahansa kokoisia: pikseleistä fotoneihin ja aina kirjeistä useiden teratavujen datasetteihin. (Borgman 2015, 50, 64, 280.)

2.2 Avoin tiede

Avoin tiede (open science) on laaja kattokäsite, joka kuvaa tieteen ja tutkimuksen kokonaisvaltaista muutosta. Se on ennen kaikkea tiedepoliittinen liike, jonka ytimessä on pyrkimys mukauttaa tieteen perusolemukseen kuuluvat avoimuuden käytännöt osaksi digitaalista toimintaympäristöä. Siinä missä perinteisesti on keskitytty suurelta osin vain tutkimustuloksiin ja niiden julkaisuun tieteellisissä lehdissä, ovat avoimeen tieteen myötä huomionalaiseksi tulleet myös tutkimusprosessin muiden vaiheiden tuottama tieto ja niiden jakaminen sekä käyttäminen. Avoin tiede vaikuttaa koko tieteen tekemisen sykliin sekä kaikkiin relevantteihin sidosryhmiin kuten yliopistoihin ja julkaisijoihin. (European Commission 2016a; UNIFI 2018.)

Avoin tiede ilmeni käsitteenä ensimmäisen kerran vuonna 2003 (David 2003), mutta ei voida sanoa, että vieläkään olisi yhtä yhtenäistä määritelmää olemassa. Yhtenä hyvänä määritelmä voidaan pitää Euroopan komission Horizon 2020 -projektin (Horizon 2020, tässä Tennant et al. 2016) yhteydessä tuotettua määritelmää avoimelle tieteelle:

> "The transformation, opening up and democratization of science and research through ICT, with the objectives of making science more efficient, transparent and interdisciplinary, of changing the interaction between science and society, and of enabling broader societal impact and innovation."

Yksi avoimen tieteen päätehtävistä on siis taata tutkimustuotoksille niin laaja saatavuus kuin mahdollista. Tutkimustuotoksiin luetaan tutkimusjulkaisut (artikkelit, monografiat ja muut tuloksia esittelevät materiaalit), tutkimusaineistot, tutkimusmenetelmät sekä materiaalit, jotka tuovat tutkimustuotoksia suuremman yleisön saataville, esimerkiksi opetusmateriaalit. Avoimen tieteen perimmäisenä tarkoituksena on kuitenkin koko tutkimuskulttuurin muuttaminen avoimempaan suuntaan. Tämä voi siis tarkoittaa myös avoimempaa ja läpinäkyvämpää tutkimuksen evaluointia, tai tehokkaampaa tavallisten kansalaisten osallistamista tieteeseen kansalaistieteen (citizen science) kautta. (Avoin tiede n.d.; OECD 2015; UNIFI 2018.)

~~2.2~~2.3 Avoimet tutkimusaineistot

Tutkimusjulkaisujen avoimen saatavuuden (open access) liikkeen syntyyn vaikutti merkittävästi internetin ja teknologian kehitys 2000-luvun taitteessa, mutta yhtäältä nämä vaikuttivat myös tutkimusaineistojen kasvavaan volyymiin maailmalla, sekä tapoihin tehdä tutkimusta ylipäätänsä (Gold 2007; Hey & Trefethen 2003; Jankowski 2007; Whyte & Tedds 2011). Avoimen saatavuuden liike sekä alati kasvavat aineistomäärät käänsivät huomiota myös tutkimusaineistoihin, ja hiljalleen alettiinkin puhua *avoimista tutkimusaineistoista* (open data, open research data, open access to data), erityisesti julkisrahoitteisten hankkeiden yhteydessä. Vuonna 2010 julkaistuissa avointen tutkimusaineistojen *jatkokäytön* suosituksissa "Panton Principles" (Murray-Rust et al. 2010) avoimet tutkimusaineistot määriteltiin aineistoina, jotka ovat vapaasti saatavilla verkossa sallien niiden lataamisen, kopioinnin, analysoinnin ja *uudelleenkäyttämisen* ilman maksullisia, lail-

lisia tai teknisiä rajoitteita. Käytän tutkielmassa jatkokäyttöä ja uudelleenkäyttöä synonyymien tavoin, luettavuuden kannalta. Niillä tarkoitetaan tässä tutkielmassa sähköisestä arkistosta löytyvien tutkimusaineistojen käyttämistä johonkin muuhun kuin niiden alkuperäiseen tarkoitukseen. Avaan jatkokäytön käsitettä tarkemmin luvussa 4.1.

Lukuisat tieteenalat ovat vielä kaukana edellä mainittujen Panton Principles -suositusten tavoitteista, mutta tavoitteet ovat osaltaan myös mahdottomia joillain aloilla. Esimerkiksi biotieteissä ja yhteiskuntatieteissä on paljon sellaisia aineistoja, joihin vaikuttavat tietosuojalait ja ihmisoikeudet, jolloin tutkimusaineistojen avoin jakaminen ei ole aina mahdollista. (Pampel & Dallmeier-Tiessen 2013, 215.) Tästä huolimatta, avoimien tutkimusaineistojen keskeisinä tekijöinä voidaan pitää niiden avointa ja maksutonta saavutettavuutta sekä riittävän hyvää laatutasoa, mikä mahdollistaa niiden jatkokäytön (Boulton et al. 2012, 14; Open Data Commons 2013, tässä Borgman 2015, 44; Murray-Rust et al. 2010; Pampel & Dallmeier-Tiessen 2014, 215)

Keskeisimpiä argumentteja tutkimusaineistojen avoimelle saatavuudelle ovat mahdollisuus käyttää jo olemassa olevia aineistoja uutta tutkimusta varten (Pampel & Dallmeier-Tiessen 2013, 214), tutkimuksen todennettavuuden tehostuminen (Borgman 2015, 208; Open Science Collaboration 2015; Pampel & Dallmeier-Tiessen 2013, 214; Tennant et al. 2016), julkisin varoin tuotetun tiedon parempi saavutettavuus julkisen yleisön toimesta (Borgman 2015, 208; OECD 2007) sekä tutkimuksen ja innovaation kehittäminen esimerkiksi uusien tutkimusmenetelmien ja -kysymysten kautta (Borgman 2015, 208; Pampel & Dallmeier-Tiessen 2013, 214; Tennant et al. 2016).

Tutkijoilla voidaan nähdä olevan avointen tutkimusaineistojen suhteen kaksi roolia: tutkimusaineistojen käyttäjinä sekä tuottajina. Käyttäjinä heidän pitäisi kyetä löytämään aineistot, pääsemän niihin käsiksi sekä käyttämään niitä. Tutkimusaineistojen tuottajina heidän tulisi tehdä ja kuvailla aineistonsa riittävän hyvin, jotta muut voivat niitä ymmärtää ja käyttää. Tämän lisäksi aineistot tulisi tallentaa tutkimusaineistojen avoimuutta tukeviin luotettaviin sähköisiin arkistoihin (repository), josta ne voidaan saavuttaa ja ladata. (High Level Expert Group on Scientific Data 2010; Pampel & Dallmeier-Tiessen 2010, 214-215; Royal Society 2012.) Tarkastelen seuraavaksi tarkemmin, mistä ajatus

avoimista tutkimusaineistoista on lähtenyt liikkeelle, ja miten edellä mainitut argumentit ja tavoitteet on pyritty saavuttamaan.

Avointen tutkimusaineistojen juuret voidaan asettaa jo 50-luvulle, jolloin perustettiin World Data Center -järjestelmä, ja CODATA hieman myöhemmin vuonna 1966 (Korsmo 2010; Lide and Wood 2012). Yksi aikaisimpia pioneereja tutkimusaineistokäytäntöjen saralla on ollut Ison Britannian Ekonomisen ja valtiotieteellisen tutkimuksen neuvosto (ja sen edeltäjä) 1970-luvulla, joka edellytti kaikkia apurahan haltijoita tarjoamaan heidän määrälliset tutkimusaineistonsa virallisesti arkistoitavaksi (Bishop & Kuula-Luumi, 2017).

Aineistorikkaassa internetin ympäröimässä nykypäivässä avointen tutkimusaineistojen kannalta ensimmäinen merkittävä virstanpylväs oli vuoden 2003 Berliinin julistus, ”Berlin Declaration on Open Access to Knowledge in the Science and Humanities”. Julistus eroaa aiemmista avoimen saatavuuden julistuksista ja niiden määritelmistä siinä, että avoin saatavuus ei koskenut pelkästään tieteellisiä julkaisuja, vaan mukana olivat vaatimukset avoimesta saatavuudesta ja käytettävyydestä tutkimusaineistojen raakadataan, metadataan, lähdeaineistoihin, digitaalisiin esityksiin, graafisiin materiaaleihin sekä muihin multimedia materiaaleihin. (Pampel & Dallmeier-Tiessen 2014, 214.)

Vuonna 2004 myös Organisation for Economic Co-operation and Development (OECD) herätti keskustelua tutkimusaineistojen avoimuudesta esittelemällä suosituksensa, jossa avoimen saatavuuden periaatteet pätisivät julkisrahoitteisten tutkimistulosten lisäksi myös tutkimusaineistoihin. Suosituksen pohjalta julkaistiin tarkennetut linjaukset vuonna 2007 raportissa ”OECD Principles and Guidelines for Access to Research Data from Public Funding”. Raportti on yksi tarkimmista asiakirjoista kuvaamaan ja tunnistamaan monia eri huomioonotettavia tekijöitä tutkimusaineistojen julkaisemisessa, kuten tieteenalojen erot ja monimuotoisuuden, lailliset ongelmat (yksityisyys, patentit), pitkäaikaissäilytyksen ja kuratoinnin eli ylläpidon takaamisen, yhtenäisten dokumentointistandardien omaksumisen tutkimusaineistojen ja metadatan laadun takaamiseksi sekä keskeisimpänä infrastruktuuria koskevan kestävyyden: pysyvän pääsyn tarjoaminen tutkimusaineistoihin. (Borgman 2015, 207; OECD 2007; Pampel & Dallmeier-Tiessen 2014, 214; Regazzi 2015, 192-193.)

Asiakirjassa korostui myös tarve käsitellä tutkimusaineistojen jakamista koskevia järjestelyjä mielellään jo rahoitushakujen yhteydessä. Tutkimusrahoittajat ryhtyivätkin hiljalleen vaatimaan *aineistonhallintasuunnitelmia* (research data management plan) rahoitusta hakevilta hankkeilta. Keskeistä näissä suunnitelmissa oli määritellä koko tutkimusaineiston elinkaari, tarkoituksena taata sekä edesauttaa tutkimusaineistojen pitkäaikaissäilytystä ja potentiaalista jatkokäyttöä. (Borg 2010; Burwell et al. 2013; Holdren 2013; Nuorteva 2008; OECD 2007.)

2010-luvun alkupuolella on ollut selkeästi havaittavissa kirjastojen, sähköisten arkistojen ja muiden instituutioiden alati kasvava yhteistyö sekä pyrkimys kehittää informaatioinfrastruktuureja tukemaan tutkijoita tutkimusaineistojensa käsittelyssä mutta myös yleisellä tasoa tuomaan esille aineistojen jakamista (Osswald and Strathmann 2012; Pampel, Bertelmann & Hobohm 2010, tässä Pampel & Dallmeier-Tiessen 2014, 218; Reilly 2012).

Tutkimusaineistojen saatavuuteen liittyviä kysymyksiä käsiteltiin 2010-luvun alussa lukuisilla eri foorumeilla. Jo tuolloin monet kansainväliset tieteelliset kausijulkaisut eivät julkaisseet artikkelia, jonka argumenttien pohjana toimiva tutkimusaineisto ei ollut avoimesti saatavilla. (Forsman 2010.) Vuonna 2012 Euroopan komissio julkaisi suosituksen nimeltä “Commission Recommendation on Access to and Preservation of Scientific Information”, jossa EU:n jäsenmaita pyydettiin takaamaan julkisin varoin tuotettujen tutkimusten tutkimusaineistojen avoimuus ja jatkokäyttö digitaalisen infrastruktuurien kautta. (European Commission 2012; Pampel & Dallmeier-Tiessen 2014, 214.)

Vuoteen 2016 mennessä ei OECD:n jäsenmaissa ollut vielä yhtenäistä kansallista säädöstä, joka ohjaisi tutkimusaineistojen jakamista pakolliseksi toimenpiteeksi aineistonhallintasuunnitelmia vaativien tahojen kasvusta huolimatta. Kaikissa aiemmin mainituissa säädöksissä, politiikoissa, raporteissa ja toimintaperiaatteissa näkyy kuitenkin vahva tutkimusaineistojen jakamisen eetos, mikä vaikutti jo myös organisaatiotasolla. (Bishop & Kuula-Luumi 2017; Fecher, Friesike & Hebing 2015.)

Vaikka aineistonhallintasuunnitelmia edellytettiin enenevissä määrin julkisrahoitteisilta tutkimushankkeilta, niin ymmärrys siitä mikä loppujen lopuksi oli hyvää aineiston hallintaa, oli suurelta osin vielä määrittelemättä. Tutkimusaineistojen hyvän hallinnan takaaminen koko niiden elinkaaren ajan on avain asemassa tutkimusaineistojen löydettävyyden ja mahdollisen jatkokäytön yhteydessä. Tähän ja muihin ongelmiin tarjottiin ratkaisua ja ohjeistusta kansainvälisen organisaation Force11 toimesta vuonna 2016 julkaistuilla FAIR-periaatteilla (FAIR Guiding Principles), jotka edustavat tieteenkentällä viimeisintä suuremman skaalan ohjeistusta. (Wilkinson et al. 2016)

FAIR-periaatteita olivat löydettävyys (Findable), saavutettavuus (Accesible), yhteentoimivuus (Interoperable) ja uudelleenkäytettävyys (Reusable). Perimmäisenä ajatuksena oli parantaa tutkimusaineistojen (ja muiden digitaalisten tuotosten) löydettävyyttä, saavutettavuutta, analysoitavuutta ja uudelleenkäytettävyyttä tutkimusaineistojen monimuotoisuudesta ja niitä ylläpitävien sähköisten arkistojen runsaudesta huolimatta. Erityistä painoarvoa annettiin myös ohjelmallisten menetelmien ja koneluettavuuden parantamiselle avointen tutkimusaineistojen yhteydessä. G20-johtajat osoittivat tukensa FAIR-periaatteille jo samana vuonna, ja tätä nykyä periaatteet ovat jo laajalti hyväksyttyjä. (European Commission 2016b; Sustkova et al. 2020; Wilkinson et al. 2016.)

~~2.3~~2.4 Tutkimusaineistojen tieteenalaeroista

Tutkimusaineistojen monimuotoisuus ei johdu ainoastaan laajoista tavoista toteuttaa tutkimusta, vaan myös monista eri tavoista, joilla eri entiteetit voidaan esittää tutkimusaineistoina. Tieteenalat eroavat toisistaan näiden seikkojen lisäksi myös esimerkiksi tutkimushankkeiden tavoitteiden, aineistojen keruun, aineistojen analyysin sekä vaihtoehtojen suhteen siinä, mistä kaikkialta tutkimusaineistoja voidaan kerätä. Nämä kaikki vaikuttavat siihen kuinka paljon tieteenaloilla on esimerkiksi potentiaalisia tutkimusaineistoja saatavilla ja käytettävissä, minkä kokoisia ne ovat, minkälaisella tahdilla niitä julkaistaan, kuinka monimuotoisia ne ovat, ja ovatko ne esimerkiksi laadullisia vai määrällisiä ja missä formaatissa. Niin ikään eroavat myös itse kulttuurit tutkimusaineistojen jakamisen ja jatkokäytön suhteen, sekä kriteerit aineistojen omistajuudesta ja vastuusta. (Borgman 2015, 55-56,79.)

Mahdollisuus löytää ja saada käsiinsä tutkimusaineistoja on suurinta tieteenaloilla, jotka ovat investoineet sähköisiin arkistoihin. Alat, joilla aktiivisesti kerätään tutkimusaineistoja muiden saataville, omaavat kannustimia arkistojen ja muiden relevanttien palveluiden luomiselle. Tällaisia aloja ovat muun muassa astronomia, biologia, biolääketiede, geotieteet sekä -omiikka päätteiset biologian alat, kuten genomiikka. (Borgman 2015, 280; Pampel & Dallmeier-Tiessen 2014, 213; Shotton 2011.) Yhteiskuntatieteissä tätä voidaan rajata erityisesti kyselytutkimuksien ympärille, ja humanistissa tieteissä tekstipohjaisia korpuksia käsitteleville tieteenaloille. Kaikille näille on yhteistä mahdollisuus verrata ja yhdistellä tutkimusaineistoja; tutkijat mielellään osallistuvat aineistojen jakamiseen, jotta pääsevät myös käsiksi muiden aineistoihin. (Borgman 2015, 281.)

Yhteiskuntatieteellisissä tutkimusaineistoissa on keskeistä pyrkimys kuvata ihmisten käyttäytymistä niin runsaasti kuin mahdollista, samalla kunnioittaen tutkimuskohteiden oikeutta yksityisyyteen. Yksityisyyden suojaan liittyen, on tyypillistä asettaa tutkimusaineistoille rajoitteita avoimuuden asteen suhteen, ja tämän lisäksi myös embargoajat ovat tyypillisiä (Faniel, Kriesberg & Yakel 2015). Embargoajan aikana aineistoa ei saa julkaista vapaasti verkossa. Tutkimusaineistoja kerätään paljolti määrällisillä menetelmillä kuten kyselyillä, sekä laadullisesti etnografian ja haastattelujen avulla. (Borgman 2015; 126-127.)

3 TUTKIMUSAINEISTOJEN JULKAISEMINEN JA JAKAMINEN

Tarkastelen tässä luvussa tutkimusaineistojen julkaisemisen ja jakamisen tilannetta, hahmottaen tutkimusaineistojen jatkokäyttöön vaikuttavia tekijöitä. Luvussa 3.1 avaan niitä tapoja ja väyliä, joita aineistojen julkaisemiseen ja jakamiseen käytetään. Luvussa 3.2 avaan keskeisiä infrastruktuurillisia edellytyksiä tutkimusaineistojen julkaisemiselle ja jakamiselle. Luvussa 3.3 käyn läpi toimintaympäristön ja käytänteiden vaikutusta tutkimusaineistojen julkaisuun ja jakamiseen.

3.1 Julkaisun ja jakamisen tapoja

Tutkimusaineistojen julkaisemiselle voidaan koostaa neljä virallisempaa ja vähemmän virallisempaa väylää: (1) vähemmän virallisesti esimerkiksi julkaistuna erillisellä verkkosivulla tai niin, että tutkija antaa pääsyn aineistoon sitä kysyttäessä, (2) aineiston julkaiseminen niin kutsuttuna data-julkaisuna (data paper), (3) julkaisu artikkelien rikastajana (enriched publication) sekä (4) tutkimusaineiston julkaiseminen omana itsenäisenä tutkimustuotoksena hyväksyttyihin tiedeyhteisöjen tai instituutiokohtaisiin sähköisiin arkistoihin (Alsheikh-Ali et al. 2011; Wallis, Rolano and Borgman 2013; Borgman 2015, 12; Pampel & Dallmeier-Tiessen 2014, 219.)

Yksityinen aineistojen jakaminen sekä julkinen verkkoon jakaminen saattavat tyydyttää tutkimusaineistojen julkaisun vaatimukset mutta eivät juurikaan edesauta aineistojen löydettävyyttä, alkuperän ymmärtämistä, käytettävyyttä tai kestävyyttä. Nämä molemmat keinot ovat parhaimmillaan vain jakohetkenä, sillä tutkijoiden kyky tuntea ja ymmärtää aineistojaan huononee ajan myötä, ja verkkosivulle linkitetty aineisto ei myöskään päivity kehittyvän teknologian, formaattien ja tiedostomuotojen myötä. (Borgman 2015, 229.)

Toinen vaihtoehto data-julkaisu on saanut 2010-luvulla enenevissä määrin huomiota ja mielenkiintoa osakseen. Chavanin ja Penevin (2011) määrittämänä tämän kaltaisen julkaisun pääasiallinen tehtävä on kuvailla itse tutkimusaineistoa, sen sijaan että raportoi-

taisiin tutkimuksesta. Data-julkaisujen parissa on tehty kokeiluja muun muassa geotieteissä ja ekologiassa. Mallin käyttö on edennyt myös niin kutsuttuihin data-kausijulkaisuihin. Sekä itse tutkimusaineistoille että aineistoa kuvaileville julkaisuille rekisteröidään *pysyväistunnisteet* (persistent identifier), jotka mahdollistavat myös tutkimusaineistoihin viittaamisen. Pysyväistunnisteita tutkimusaineistojen osalta avataan tarkemmin luvussa 3.2.3. Tämä proseduuri tukee myös tutkimusaineistojen näkyvyyttä. Tutkimusaineistokeskeiset kausijulkaisut ovat kasvussa, mutta näiden perustaminen on mielekästä vain, kun tutkimusaineisto, sen metadata ja relevantit tekstijulkaisut ovat kaikki vapaasti saatavilla, koska vain ja ainoastaan silloin voi rajaton tutkimusaineiston ilmaiskäyttö olla mahdollista. (Pampel & Dallmeier-Tiessen 2014, 219-220.)

Kolmannessa rikastetun artikkelin vaihtoehdossa artikkelin ja datan linkitys toisiinsa on myös läsnä (Woutersen-Windhouwer & Vernooy-Gerritsen 2009). Tarkoituksena on rakentaa ja ylläpitää teknisiltä ominaisuuksiltaan sen kaltaista ympäristöä, joka yhdistää kaikki relevantit informaatio-objektit artikkelin ympäriltä. Näin syntyisi tiedon alue, jossa artikkelin perustana toimiva tutkimusaineisto voidaan saattaa vapaasti kaikille saatavaksi. (Pampel & Dallmeier-Tiessen 2014, 220.)

Neljäs ja keskeisin julkaisumuoto, eli tutkimusaineistojen julkaisu ja jakaminen omana itsenäisenä tieteellisenä tuotoksena sähköisessä arkistossa mahdollistaa kaksi jälkimmäistä julkaisumuotoa. Aineistojen julkaisu arkistoissa edellyttää niiden saattamista julkaisukelpoisiksi ja uudelleenkäytettäviksi kuvailutietojen laadun avulla. (Borgman 2015, 228; Lee, Dourish and Mark 2006; Mayernik 2016; Pampel & Dallmeier-Tiessen 2014, 220). Tarkastelen sähköisiä arkistoja tarkemmin alaluvussa 3.2.2.

3.2 Julkaisemisen ja jakamisen edellytykset

Käsittelen tässä luvussa tutkimusaineistojen julkaisemisen ja jakamisen perusedellytyksiä infrastruktuurillisesta näkökulmasta. Alaluvussa 3.2.1 esittelen tutkimusaineistojen metadatan käsitteen ja alaluvussa 3.2.2 tutkimusaineistojen ylläpidosta vastaavat infrastruktuurit. Alaluvussa 3.2.3 tarkastelen pysyväistunnisteiden merkitystä jatkokäytön kannalta.

3.2.1 Metadata

Tutkimusaineistojen dokumentointi on tärkeä, koko tutkimuksen läpi jatkuva prosessi, joka tekee aineistojen uudelleenkäytön mahdolliseksi. Tutkimusaineistoista pyritään kuvaamaan käytetyt välineet ja ohjelmistot niin analysoinnissa kuin hankinnassakin. Aineistojen tarkka dokumentointi on kuitenkin aina pois tutkijan muista tehtävistä, ja joillain aloilla dokumentointia hidastaa myös kuvailustandardien puute. (Pampel & Dallmeier-Tiessen 2014, 217-218.)

Tutkimusaineistojen kuvailutietoja kutsutaan *metadataksi*. Metadata määritellään yksinkertaisimmillaan tietona tiedosta, ja tutkimusaineistojen kontekstissa metadata kertoo jotain tutkimusaineistoista ja niihin liittyvistä suhteista. (Borgman 2015, 66.) NISO:n (National Information Standards Organization 2004, 1) yleisen määritelmän mukaan metadata on rakenteellista informaatiota, joka kuvaa, selittää, paikantaa tai muuten helpottaa informaatioresurssin käyttöä, hallintaa tai hakemista. Tutkimusaineistojen metadata ja muut aineistoja kuvailevat dokumentit kuten koodikirjat ovat tarpeellisia, jotta aineistoja voi löytää, tulkita ja käyttää (Borgman 2015, 275).

Jotta aineistojen metadatan dokumentointi olisi yhtenäisempää ja vakiintuneempaa, ehdottavat ja kehittävät yhteisöt *metadatastandardeja* (Metadata standard, metadata vocabulary standard, metadata schema), jotka ovat muodollisia dokumentteja, joissa määritetään yhteiset käytännöt aineiston kuvailuun (Lei Zeng & Qin 2016, 12). Standardien muodostamisen ja käyttöönoton jälkeen tutkijat voivat löytää, louhia sekä yhdistellä tutkimusaineistoja useissa eri lähteistä. Samasta syystä nämä mekanismit luovat kuitenkin myös kitkaa niiden tieteenalojen kesken, joilla ei ole yhteensopivia standardeja. (Borgman 2015, 68.) Standardeja ja skeemoja on lukuisia, niin tieteenala- kuin arkistokohtaisiakin, mutta yhtenä hyvänä lähtökohtana on DataCiten metadataskeema; se on tarkoitettu tieteenalojen rajat ylittäväksi, aineistojen löytämisen kannalta relevantiksi metadatastandardiksi (Ball & Duke 2015).

Borgmanin (2015, 70) mukaan tutkimusaineistojen metadatan kyky esittää tutkimusaineistojen alkuperää ja kehityskulkua tähän pisteeseen asti on yksi sen tärkeimmistä ominaisuuksista. Aineiston alkuperää katoaa vääjäämättä jo siinä tilanteessa, kun se erotetaan tutkimuksestaan, ja kehityskulkua ja muutosta aineistoissa voi olla esimerkiksi redusoitu yksinkertaisempaan, laajemmin tunnettuun formaattiin kuten Excel-tiedostoon (Borgman 2015, 219).

Tutkimusaineiston alkuperän esittäminen voi vaikuttaa siihen, pidetäänkö aineistosta johdettuja tuloksia luotettavina tai toistettavina, ja kenelle meriitti on merkitty (Buneman, Khanna, and Wang-Chiew 2001; Moreau et al. 2008). Argumentit tutkimusaineistoihin viittaamisen puolesta edellyttävät parempaa alkuperän tallentamista osaksi aineistojen metadataa (Uhlir 2012, tässä Borgman 2015, 70). Yhtä lailla argumentit tutkimuksen eri vaiheiden säilyttämiselle perustuvat tarpeeseen paremmasta metadatasta alkuperän suhteen (Gamble and Goble 2011).

Tutkimusaineistojen metadatan tärkeys korostuu siinä, että suurin osa työstä ja vaivasta aineistojen jatkokäytön yhteydessä koostuu aineistojen dokumentoinnin siivoamisesta, koodausten tulkinnasta sekä sisällön tarkistamisesta (Borgman 2015, 222). Huonosti kuvaillut tutkimusaineistot vaikuttavat negatiivisesti tutkijoiden päätöksiin aineistojen jatkokäytön suhteen (Boyle 2013; Eisen 2012; Kwa & Rector 2010).

3.2.2 Infrastruktuuri

Avointen tutkimusaineistojen reilu ja yhdenvertainen saavutettavuus sekä käyttömahdollisuus, jaettu vastuu aineistojen ylläpidosta, ja aineistojen optimaalinen tuottaminen, hallinnointi, jatkokäyttö sekä pitkäaikainen kestävyys ovat haasteita joihin vastaaminen edellyttää sähköisiä arkistoja. Vain muutamat tieteenalat ovat kyenneet tähän. Sähköiset arkistot ovat tieteellisten julkaisujen ja tutkimusaineistojen tallentamiseen ja avoimeen verkkojulkaisemiseen soveltuvia teknisiä järjestelmiä ja niiden ympärille rakennettuja palveluita. Näistä voidaan puhua myös (digitaalisina) arkistoina, kokoelmina, tietopankkeina, tietokantoina tai tietojärjestelminä. (Borgman 2015, 43)

Joillain aloilla tutkimusaineistojen yksityinen jakaminen tutkijoiden välillä riittää. Se voi tapahtua esimerkiksi kasvotusten, tutkijoiden omien verkkosivujen tai tutkijoiden tutkimushankkeiden verkkosivujen kautta. Osa tieteenaloista taas saattaa kääntyä tieteellisten kirjastojen puoleen. Sähköisten arkistojenkin kesken rahoituksen määrä ja rahoituksen kesto vaihtelee, aiheuttaen ongelmia pitkäaikaissäilytykselle. (Borgman 2015, 43.)

Tutkimusaineistoihin keskittyvät sähköiset arkistot ovat pääkomponentti tutkimusaineistojen julkaisemisessa, uudelleenkäyttämisessä ja eheän pitkäaikaissäilytyksen takaamisessa. Tutkimusaineistojen pysyväistunnisteita jakava DataCite määrittelee tutkimusaineistojen sähköiset arkistot tutkimusorganisaatioiden ylläpitäminä palveluina, joissa tutkimustuotokset säilytetään, ylläpidetään ja tehdään saavutettaviksi. (Ball & Duke 2015; Borgman 2015, 226; DataCite 2020.)

Arkistojen sitoutuminen aineistojen ylläpitoon vaihtelee paljon, ja arkistot määrittelevätkin itse omat laatu- ja formaattistandardinsa aineistoille, joita ottavat vastaan. Osalle arkistoista riittää, kunhan aineistot täyttävät vaadittavat tekniset standardit, jättäen suurimman osan aineiston tieteellisestä vahvistamisesta itse tutkijalle. Toiset taas tarkistavat arkistoon jätettävän aineiston laatustandardeja vasten, ennen kuin hyväksyvät aineiston osaksi kokoelmiansa. (Borgman 2015, 225-226.) Standardien määrittelyn tueksi on erilaisia sertifikaatteja, joita arkistot voivat hankkia osoittamaan varteenotettavuuttaan. Yksi tällainen sertifikaatti on CoreTrustSeal, jonka pyrkimys on parantaa sähköisten arkistojen kestävyyttä ja luotettavuutta (CoreTrustSeal 2021-a).

Esimerkiksi tämän tutkielman tarkastelun kohteena olevan Swedish National Data Service (SND) -tutkimusaineistoinfrastruktuurin sähköisessä arkistossa ohjeistetaan aineistojen jättämisessä ja tietojen syöttämisessä. Jätetty aineisto ja sen metadata käyvät läpi vielä arviointiprosessin, ennen kuin ne hyväksytään osaksi arkiston kokoelmaa. (SND 2021-h.) Yhteiskuntatieteisiin keskittyvillä sähköisillä arkistoilla on ylipäänsä suhteellisen pitkä historia tutkimusaineistojen jakamisen ja ylläpidon saralla (Daniels et al 2012; Yoon 2014).

Kun aineistot on jätetty sähköiseen arkistoon, siirtyy vastuu ylläpidosta ja saavutettavuudesta yleensä arkiston ylläpitäjälle. Myös sekin vaihtelee, kuinka pitkäksi aikaa arkistot sitoutuvat aineistoa säilyttämään (Consultative Committee for Space Data Systems 2012; CoreTrustSeal 2019; Jantz and Giarlo 2005) Tutkimusaineistojen varmuuskopiointi niiden luovutetussa muodossa voi olla kallista, mutta niiden ylläpito vaatii paljon suurempia investointeja, kun niitä täytyy siirtää uusiin teknologioihin ja formaatteihin sitä mukaa, kun niitä ilmenee. Kokonaisuudessaan sähköiset arkistot ovat huomattavia investointeja niitä rahoittavien yhteisöjen, yliopistojen ja valtion tahojen toimesta. (Borgman 2015, 225-226.)

Sähköisten arkistojen välillä on myös huomattavaa vaihtelua sen suhteen, kuinka paljon ne vahvistavat ylläpitämiään aineistojaan metadatalla ja alkuperää koskevalla dokumentaatiolla. Eroa on myös aineistoille lisäarvoa tuottavien palveluiden tarjonnassa. Arkistot voivat tuoda lisäarvoa ylläpitämiinsä tutkimusaineistoihin metadatan, alkuperän, luokittelun, datarakenne standardien sekä migraation avulla. Myös aineistojen löydettävyyden ja käytettävyyden parantaminen eri työkalujen ja palveluiden kautta parantaa aineistoista saatavia hyötyjä. (Borgman 2015, 225-226.) Tämä tutkielma ei välttämättä olisi sellaisenaan toteutunut, ellei SND yhteistyössä DataCiten kanssa tarjoaisi avointa ohjelmointirajapintaa, jonka kautta voi hankkia SND:n arkistossa olevien tutkimusaineistojen kuvailutietojen metadatatietueet. Borgman (2015, 226) huomioikin, että tutkimusaineistojen uudelleenkäyttöäkin vielä korkeampi tavoite on aineistojen valjastaminen niihin uusiin ja odottamattomiin tutkimuskysymyksiin ja tarpeisiin, joissa tutkimusaineistoja käytetään uudelleen eri tavalla kuin niiden alkuperäisessä kontekstissa.

3.2.3 Pysyväistunnisteet

Lukuisille tutkimusaineistoille on rekisteröity pysyviä ja uniikkeja tunnisteita, jotka auttavat viittaamisen yksiselitteisyydessä, sekä aineiston paikantamisessa muualla verkossa. Yksi tällainen tunnisteellinen skeema eli pysyväistunniste (Persisten Identifier, PID) on Digital Object Identifier (DOI), joita tutkimusaineistoille myöntää DataCite. Da-

taCitea voidaan pitää olennaisimpana tutkimusaineistojen pysyväistunnisteiden rekisteröintitoimistona. Pysyväistunnisteet ovat myös merkittävässä roolissa FAIR-periaatteiden toteuttamisessa. (Ball & Duke 2015; (SND 2021-f).)

Pysyväistunnisteet ovat olleet olemassa jo yli 20 vuotta. Ne kehitettiin alun perin ratkaisemaan niin kutsutun linkkimädän (link rot) ongelma, jonka keskiössä olivat verkko-osoitteiden pysymättömyys, mikä johti linkkien toimimattomuuteen (Dellavalle et al. 2003; Lawrence et al. 2001). Uniikkien pysyväistunnisteiden avulla verkossa oleva objekti ei ollut enää rajoitettu verkossa olevan sijaintinsa puolesta (ks. Lawrence et al. 2001). Alun perin pysyväistunnisteet luotiin parantamaan tutkimustulosten saavutettavuutta, läpinäkyvyyttä ja uudelleentuotettavuutta. (Klump & Huber 2017.)

Klump ja Huber (2017) tutkivat Registry of Research Data Repositories (re3data.org) -palvelun kautta rekisteröityjä tutkimusaineistoja ylläpitävien sähköisten arkistojen pysyväistunnisteiden käyttöä. 1381 arkistosta 475 rekisteröi pysyväistunnisteita aineistoilleen. Reilusti käytetyin pysyväistunniste oli DOI (76,6 %), ja toisiksi käytetyin oli Handle (noin 26,3 %). Osa arkistoista käytti useampaa kuin yhtä pysyväistunnistetta. Tämän tutkimuksen aineiston koostaville SND: yhteiskuntatieteellisille tutkimusaineistoille on rekisteröity DOI-pysyväistunnisteita. (Klump & Huber 2017; SND 2021-f.)

3.3 Toimintaympäristön ja käytänteiden vaikutus

Infrastruktuurillisten edellytysten (luku 3.2) lisäksi on havaittu myös muita tekijöitä, jotka voivat hankaloittaa tutkimusaineistojen julkaisemista ja jakamista. Näitä tekijöitä ovat esimerkiksi lailliset ongelmat, tutkijoiden kannustimien puute, tieteelliset käytännöt ja tutkijoiden osaaminen aineistonhallinnan suhteen. Lailliset ongelmat liittyvät tutkimusaineistoissa esiintyvien ihmisten yksityisyydensuojaan sekä aineistoja koskeviin omistajuuden kysymyksiin.

Ihmisiä käsittelevissä tutkimusaineistoissa tulee kiinnittää huomiota tutkittavien yksityisyydensuojaan, jota on pyritty ratkaisemaan anonymisoinnin avulla. Tutkimusaineistojen paraneva avoimuus, saatavuus ja löydettävyys tuovat esille ongelmia tässä yhtey-

dessä, kun aiemmin anonymisoidut yksilöt saatetaan jatkossa pystyä tunnistamaan rikastaessa tutkimusaineistoja muilla lähteillä ja aineistoilla. (Borgman 2015, 11, 77-78, 229.) Tutkimusaineistojen anonymisoinnin keinojen kehittäminen, niin ettei aineistojen hyödyllisyys tutkimukselle kärsi, on niin tekninen kuin poliittinenkin ongelma (Ohm 2009; Sweeney 2002). Yhtenä ratkaisuna on ihmisyksilöitä koskevien tutkimusaineistojen luovuttamisen valvonta: ainoastaan muodollisesti päteville tutkijoille, ja taaten ettei tutkimuksen yksilöitä yritetä tunnistaa (Borgman 2015, 229). Yhteiskuntatieteiden alalla tämänkaltaista valvontaa harjoittavat esimerkiksi Tietoarkisto Suomessa ja Swedish National Data Service Ruotsissa (SND 2021-i; Tietoarkisto n.d.).

Tutkimusaineistojen omistajuuteen liittyvät ongelmat vaihtelevat alueittain, käytännöittäin sekä lainkäyttöalueittain. Ongelmia aiheuttavat tekijät kuten hämmennys ja epäselvyys omistajuudesta, oikeudesta itse aineistoon, julkaisuoikeuksista, julkaisuvastuista sekä julkaisuehdoista. Mitä enemmän tutkimushankkeessa on tutkijoita eri maista ja lainkäyttöalueista (jurisdiction), sitä vaikeampaa on varmistaa oikeus julkaista tutkimusaineisto(t). (Arzberger et al. 2004; Hirtle 2011, tässä Borgman 2015, 218.) Nämä ovat kaikki merkittäviä rajoitteita tutkimusaineistojen saatavuudelle, niin tutkijoiden kuin sähköisten arkistojenkin kannalta. Mitä aikaisemmin tutkimushankkeissa päästään yhteisymmärrykseen näiden ongelmien osalta, sitä tehokkaammin aineistojen potentiaalinen jatkokäyttö voidaan taata. (Borgman 2015, 75, 229; Erdos 2013a, 2013b.)

Epäselvyydet tutkimusaineistojen omistajuudesta johtavat myös tutkijoiden huoleen oikeudellisesta vastuusta, sekä potentiaalisista riskitekijöistä heidän maineelleen, mikäli heidän aineistojaan käytetään tai tulkitaan väärin. Vaikka tutkijat ovatkin osaltaan jo tottuneet siihen, että heidän tuloksia tulkitaan väärin massamedian myötä, on tulosten valikoiva tulkitseminen, ihmisyksilöiden tunnistaminen sekä muut tahalliset tai tahattomat väärinkäytöt oikeutettuja huolia. (Borgman 2015, 218.)

Keskeisimpiä esteitä tutkimusaineistojen jakamiselle ovat kuitenkin kannustimien puute. Luvussa 3.1 läpikäydyt erilaiset julkaisutavat kuitenkin osoittavat, että tutkimusaineistojen jakamisella on selkeä mahdollisuus tulla osaksi tieteellisiä meritoitumiskäytäntöjä. (Pampel & Dallmeier-Tiessen 2014, 219, 221.)

Yksi pääargumentti tutkimusaineistojen jakamiselle on tutkijoiden saama tunnustus aineistojen saamien viittausten kautta, mutta Borgman (2015, 266) kuuluttaa perään, että tämä on testaamaton hypoteesi. Tutkimusaineistojen saamia viittauksia voidaan arvostaa, etenkin kun aineistot ovat laajalti käytettyjä, mutta tutkijan saama tunnustus perinteisten tutkimusjulkaisuihin kohdistuvien viittausten kautta omaa niin paljon enemmän painoarvoa, että jotkut tutkijat toivovat tutkimusaineistoihinsa kohdistuvien viittausten tulevan ennemmin välillisesti (as proxy) heidän tutkimusjulkaisuihinsa, joissa aineistojen olemassaolo käy ilmi. (Borgman 2015, 266.)

Se mikä motivoi tutkijoita julkaisemaan tutkimuksensa, ei välttämättä päde tutkimusaineistojen julkaisemiseen. Tutkimustulosten julkaiseminen on tutkijoiden pääasiallinen muoto tunnustuksen saamiselle, mikä taasen tuo uramahdollisuuksia, ylennyksiä ja muita palkintoja. On kuitenkin hyvin vähän näyttöä siitä, että tutkimusaineistoihin viittaaminen olisi kannustin tutkimusaineistojen julkaisemiselle ja jakamiselle. Tutkimusaineistojen avoimuuden puolesta on vaikea argumentoida, kun kannustimet ovat lähes olemattomat. (Borgman 2015, 48; Fecherin & Frisken 2014.) Vaikka tutkimusaineistoihin viittaaminen ei itsessään vielä välttämättä ole tunnistettu ja arvostettu mittari, on kuitenkin näyttöä siitä, että tutkimusaineistojen jakaminen vaikuttaa viittaustahtiin *artikkelien* osalta. Ne artikkelit, joiden tutkimusaineisto on jaettu, saavat keskimäärin useammin viittauksia kuin ne, joissa aineistoja ei ole jaettu. (Botstein 2010; Dorch 2012; Henneken & Accomazzi 2011; Pampel & Dallmeier-Tiessen 2014, 220; Piwowar et al. 2007; Sears 2011).

Tutkimusaineistojen avoimuuden kannalla olevat säädökset (ks. luku 2.3) vetoavat tutkijoiden parempaan luonteeseen historiallisten avoimuuden argumenttien kautta, kuten tiedon jakamisen eettisyys. Harvemmin kuitenkaan otetaan huomioon tutkijana olemisen kilpailullista luonnetta ja toimintaympäristöä, hyödykkeitä palkintojen tai kunnian muodossa, epäsuhdetta työmäärän ja hyötyjen yhteydessä, vaihtelevuutta toimintatavoissa tieteenalojen ja tutkijoiden välillä, resurssien erilaisuutta yhteisöjen välillä, vaikeuksia jaetun tutkimusaineistojen tulkinnassa ja sitä resurssien skaalaa, jota tarvitaan aineistojen jakamiselle ja ylläpitämiselle. (Borgman 2015, 207.)

Tutkimuksen tekemisen ja koko toimintaympäristön kilpailullisuuteen liittyen on useilla tieteenaloilla kannattavampaa luoda uutta aineistoa, kuin käyttää olemassa olevaa, sillä tutkimusapurahoja myönnetään herkemmin aineistoa tuottaville tutkimuksille. Kuitenkin esimerkiksi astronomiassa, sosiaalisen median sekä ilmastojen tutkimuksessa etsitään varta vasten uudelleen käytettäviä tutkimusaineistoja. Kysymällä uusia kysymyksiä uudella aineistolla ja kartoittamalla aiemmin kartoittamattomia alueita on yksi vakaimmista keinoista luoda ja kehittää tutkijan uraa. Kysymällä uusia kysymyksiä vanhoista aineistoista voi myös johtaa uusiin tuloksiin, mutta voi olla haastavaa vakuuttaa lehtien editorit ja vertaisarvioijat siitä, että uudelleenanalysoinnit ovat tarpeeksi arvokkaita kontribuutioita. (Borgman 2015, 10, 213.)

Borgman (2015, 207) korostaakin, että eri sidosryhmät tutkimusaineistojen avoimen saatavuuden ympärillä ovat keskittyneet enemmän niin sanotun ”kepin käyttämiseen, koska porkkanoita ei juuri ole”. Tämä ilmenee jatkokäyttöä edesauttavien aineistonhallintasuunnitelmien, aineistojen pitkäaikaissäilytyksen sekä muiden samankaltaisten toimenpiteiden edellyttämisenä. Säädöksissä ei juurikaan mainita odotettua jatkokäytön astetta, tai sitä aineistojen julkaisemiselle ja jatkokäytölle oleellista infrastruktuuria. (Borgman 2015, 207.)

Siinä missä tutkijoiden tutkimusaineistojen kannalta tarpeelliset, itse ohjelmoidut ohjelmistot ja muut työkalut voivat olla tärkeä kilpailuvaltti, voi sama päteä myös tutkimusaineistoihin. Tuotetut koodit eivät liity vain yksittäiseen tutkimukseen tai julkaisuun, vaan niitä voi käyttää hyödyksi myöhemminkin, ja yhtä lailla tutkimusaineistoilla voi potentiaalisesti taata oma tulevaisuuden tutkimus ja työllistyminen. Tieteenaloilla, joilla ei ole runsaasti tutkimusaineistoja, on jopa yleisesti hyväksytty käytäntö kieltäytyä jakamasta aineistoja. (Borgman 2015, 12, 215, 220-221; Sawyer 2008.)

Borgman (2015, 214) kokee, että tutkimusaineistojen julkaisemista edistävien toimintaperiaatteiden onnistuminen saattaa riippua paljon radikaaleimmista investoinneista tutkimusaineistojen avoimuuden kenttään, kuin mitä tutkimusrahoittajat, julkaisijat, kirjas-

tot, arkistot tai tutkijat käsittävätkään. Tutkimusaineistojen kohtelu julkaistavina tai jaettavina tuotteina vaatii muutoksia tutkimuksen tekemisen kentän metodeissa ja käytännöissä. (Borgman 2015, 214.)

Tutkijoiden osaaminen tutkimusaineistojen käsittelyn ja kuvailun suhteen on niin ikään ongelma aineistojen julkaisemiselle ja jakamiselle. Nämä liittyvät osaltaan infrastruktuurien edellyttämään riittävän kuvailtuun metadataan (luku 3.2.1) mutta myös tutkimusaineistojen hallintaan niiden koko elinkaaren ajan. Aineistojen uudelleenkäytettävyys (ja löydettävyys) riippuu aineistojen hallinnasta ja kuvailusta, ja tutkijat tarvitsevat tätä varten työkaluja, palveluita ja tukea, jotta he kykenevät arkistoimaan aineistonsa sähköisiin arkistoihin riittävän laadukkaina. (Borgman 2015, 282; Van der Graaf & Waaijers 2011.) Tutkimusaineistojen hallinnointi ja julkaisukelpoiseksi tekeminen on pois kaikesta muusta tutkimustyöstä, mutta tutkimusaineistojen julkaisu niin, että muutkin tutkijat voivat niitä tulkita, edellyttää aineistojen riittävän hyvää käsittelyä ja hallinnointia (Borgman 2015, 217).

Luvussa 2.3 mainitut aineistonhallintasuunnitelmat tulevat vastaan tässä kohtaan: niiden avulla pyritään varmistamaan tutkimusaineistojen riittävä dokumentointi. Aineistonhallinnassa sekä itse suunnitelmien laatimisessa auttavat niin tutkimusrahoittajat (ks. Suomen Akatemia n.d.), tieteelliset kirjastot (ks. Helsingin yliopisto n.d.; Tampereen yliopiston kirjasto 2021) kuin myös neuvoa antavat elimet (ks. CESSDA Training Team 2020). Myös sähköiset arkistot auttavat ja ohjaavat metadatan merkitsemisessä ja syöttämisessä osaksi niiden infrastruktuuria (SND 2021-k).

4 TUTKIMUSAINEISTOJEN JATKOKÄYTÖN MITTAAMINEN

Käyn tässä luvussa läpi tutkimusaineistojen jatkokäyttöä, sen mittaamista ja kuinka tätä on tutkittu. Luvussa 4.1 tarkastelen tutkimusaineistojen jatkokäytön määritelmää, luvussa 4.2 perehdyn tutkimusaineistojen vaikuttavuuteen. Luvussa 4.3 keskityn tutkimusaineistojen viittaamisen määrittelyyn, ja luvussa 4.4 avaan mitä aineistojen saamat lataukset tarkoittavat. Luvussa 4.5 esittelen aiempia tutkimuksia tutkimusaineistojen jatkokäytöstä yhteiskuntatieteissä.

4.1 Tutkimusaineistojen jatkokäytön määritelmä

Puhuttaessa tutkimusaineistojen jatkokäytöstä, tai uudelleenkäytöstä, on tehtävä ero aineistojen käytön ja jatkokäytön välille. Tutkimusaineistojen käyttö viittaa aineiston käyttöön sen alkuperäisessä kontekstissa, sen tutkimuksen yhteydessä, jota varten se on luotu tai kerätty. Kun tämä aineisto ladataan sähköisestä arkistosta, ja käytetään jossain uudessa kontekstissa, puhutaan jatkokäytöstä. (Pasquetto, Randles & Borgman 2017.)

Käytön ja jatkokäytön erottaminen toisistaan ei kuitenkaan aina ole helppoa. On vaihtelua sen suhteen, lasketaanko esimerkiksi tutkijan oman aineiston käyttö tutkijan toimesta uudessa kontekstissa käytöksi vai jatkokäytöksi. Entä jos tutkija lataa oman aineistonsa sähköisestä arkistosta ja käyttää sitä uudestaan uudessa kontekstissa? Käytännössä myös tiedeyhteisöjen koostamien tutkimusaineistojen käyttö missä tahansa kontekstissa olisi aina aineiston ensimmäinen käyttökerta. (Pasquetto, Randles & Borgman 2017.)

Tässä tutkielmassa jatkokäytöksi lasketaan tutkimusaineiston käyttö jossain muussa, kuin sen alkuperäisessä kontekstissa. Myös aineiston laatijan oman aineiston käyttö uudessa kontekstissa on laskettu jatkokäytöksi.

4.2 Tutkimusaineistojen vaikuttavuus

Tutkimusaineistojen jatkokäytön mittaaminen on osa tutkimuksen *vaikuttavuuden* mittaamista. Vaikuttavuus on kuvainnollisesti yhden toimijan, tapahtuman tai resurssin vaikutus toiseen. Vaikuttavuudella on samankaltaisia piirteitä kuin esimerkiksi huomiolla (kuinka moni ihminen tietää resurssista) ja levittäytyneisyydellä (kuinka laajalle resurssi on levinnyt). Siksi onkin tärkeää miettiä mitä tarkalleen mitataan, ja kuinka vahvaa on sen kautta saatava todiste tarkastelun kohteena olevan entiteetin vaikuttavuudesta (Hicks et al. 2015). Tutkimuksen vaikuttavuuden mittaaminen tarjoaa varteenotettavia todisteita tutkimuksen tuottamista hyödyistä, vastapainona tutkimuksen teettämisen hinnalle. Perinteinen tapa mitata tutkimuksen vaikuttavuutta on tarkastella tutkimusjulkaisuja bibliometristen, viittauksia tarkastelevien analyysien kautta. Tutkimuksen vaikuttavuutta on kuitenkin katsottava laajemmin, sillä ne voivat esimerkiksi vaikuttaa käytäntöjen ja politiikoiden luomiseen, vaurauden luontiin (generating wealth), teollisten innovaatioiden kehittymiseen sekä merkittävien yhteiskunnallisten kysymysten ja ongelmien ratkaisemiseen. (Ball & Duke 2015.)

Tutkijoiden ja tutkimuslaitosten intresseissä on siis seurata tuottamansa tutkimuksen vaikuttavuutta. Selkeä aloituspiste on tarkastella tutkimustuotosten, mukaan lukien tutkimusaineistojen, vaikuttavuutta. Määrälliset menetelmät vaikuttavuuden arvioinnissa eivät kuitenkaan ole ongelmattomia. Ongelmia ovat itse mittaamisen kulttuuri, jolla on negatiivinen vaikutus tutkijoiden hyvinvointiin (Kansa 2014). Mittaamisen suhteen ongelmia ovat esimerkiksi mitattavien kohteiden rajallisuus, tarve arvioida säännöllisesti vaikuttavuutta seuraavia työkaluja, mittauksien keskisen vertailun rajallisuus sekä tarve huomioida eri metriikoiden vahvuudet johdettua dataa käyttäessä. (Ball & Duke 2015.)

Näistä ongelmista huolimatta, valittuja metriikkoja voi käyttää mittarina osoittamaan vaikuttavuutta tapauskohtaisesti. Yhtä lailla seuraamalla tutkijoiden jakamien tutkimusaineistojen jatkokäyttöä, tutkijat voivat saada tietoa siitä minkälainen tutkimusaineiston valmistelu ja julkaisutapa toimii parhaiten. Näkemällä kuka heidän aineistojaan käyttää voi avata myös uusia yhteistyöväyliä. Voidaan myös havaita yhteisöjä, jotka eivät olleet

alkuperäinen kohdeyleisö, mutta joilla kuitenkin on mielenkiintoa tutkimusaineistoa ja sen aihetta kohtaan. (Ball & Duke 2015)

Myös instituutiot voivat hyötyä tutkimusaineistojen käytön seuraamisesta. Tutkimusaineistoille suunnattujen sähköisten arkistojen omistajat osaavat varautua arkistojen käytettävyyden määristä johtuviin järjestelmien kuormituksiin, voivat mainostaa ja juhlia tutkijoidensa menestystä aineistojen jakamisen ja jatkokäytön yhteydessä, luoda erityiskokoelmia suosituista aineistoista sekä vastata tutkimusrahoittajien vaatimuksiin aineistojen säilöntäaikoja koskien. (Ball & Duke 2015)

Kaikki nämä seikat ovat tärkeitä laajemman liikkeen kannalta parantaakseen tutkimusaineistojen jakamisen laatua, läpinäkyvyyttä, tehokkuutta ja akateemisen tutkimuksen mahdollisuuksia. Tutkimusaineistojen metriikkojen avulla voidaan asettaa kannustimia aineistojen jakamiselle urakehityksen ja meritoitumisen viitekehyksessä, jossa tutkimusaineistot tunnistetaan olennaisina tutkimustuotoksina. (Ball & Duke 2015, Costas et al. 2013)

Vaikka viittauksellinen metriikka ei ole täydellistä, on se silti hyvä välillinen vaikuttavuuden mitta. Tutkimusaineistojen kautta saatavaa tietoa niiden vaikuttavuudesta ovat esimerkiksi tieto tutkijoiden jakamien aineistojen jatkokäytön asteesta, kuinka sisällytettyjä tutkijoiden jakamat aineistot ovat muihin, isompiin data-aineistoihin ja kuinka laajalti tutkijoiden kehittämää koodia on otettu käyttöön. Tämänkaltaiset tiedot ovat myös tärkeitä tutkimusrahoittajille, jotka haluavat laajempia analyyseja tutkijoiden vaikuttavuudesta. (Ball & Duke 2015.)

Mikään yksi metriikka ei kykene esittämään koko kuvaa kunkin tutkimuksen vaikuttavuudesta, ja täten onkin mielekästä tutkia millä metriikoilla saadaan paras kokonaiskuva vaikuttavuudesta. Minkään tunnusluvun liiallinen tulkinta ei ole hyvästä, mutta metriikat ovat kätevä keino saada potentiaalisesti relevanttia näyttöä tarkastelun alla olevan tutkimuksen vaikuttavuudesta. (Ball & Duke 2015.)

4.3 Tutkimusaineistoihin viittaaminen

Kaikista kypsin malli tutkimusaineistojen vaikuttavuuden mittaamiselle on verrattavissa kirjallisuuden julkaisuun ja viittaamiseen (Costas et al. 2013). Tutkimusaineistojen julkaisu sähköisessä arkistossa parhaimmassa tapauksessa osoittaa, että aineistot on tarkastettu laatunsa puolesta, taaten sen sopivuuden jatkokäytön kannalta, tehden niistä esitettäviä ja löydettäviä, ja taaten myös aineistojen pitkäaikaissäilytyksen. Julkaistu tutkimusaineisto omaa vakaat bibliografiset tiedot, jotta siihen voidaan luotettavasti viitata muista tieteellisistä tuotoksista. (Ball & Duke 2015; Borgman 2015, 225-226.)

Tutkimusaineistoihin kohdistuvia ja niiden välisiä viittauksia ei kuitenkaan voida vielä tässä vaiheessa nähdä yleisinä, vaan ennemminkin toiveikkaana lopputuloksena. Jotkut tieteenalat ovat ottaneet lähestymistavakseen väylän, jossa tutkimusaineistojen sijasta viitataan datajulkaisuun (ks. luku 3.1). Viittaukset näihin julkaisuihin voidaan tulkita välillisinä viittauksina itse aineistoihin, kun pyritään selvittämään tutkimusaineistojen vaikuttavuutta. Tämän lisäksi aineistot saatetaan myös vain mainita lähteenä, viittamatta sen tarkemmin, ja joskus aineistoja käytetään myös viittaamatta niihin laisinkaan. (Ball & Duke 2015; Borgman 2015, 251-252.)

Monilla tieteenaloilla vallitsevin lähestymistapa on viitata sitä paperia, joka aineistoa ensimmäisen kerran käytti. Tällöin luotetaan siihen, että julkaisusta ilmenee jos ja kuinka käytetty tutkimusaineisto on asetettu jakoon. Näissä tapauksissa ei ole täysin selvää kuuluko vaikuttavuus julkaisun argumentille ja tuloksille, vai pohjana olevalle tutkimusaineistolle. Tällaisilla aloilla viittauskeskeisistä metriikoista on hyvin vähän apua tutkimusaineistojen vaikuttavuuden arvioimisessa, joten muita vaihtoehtoisia mittareita on etsittävä. (Ball & Duke 2015.)

Kaikesta huolimatta, tutkimusaineistoihin viittaaminen näyttää olevan kasvussa tietyillä aloilla. Se onko aineistojen todellinen jatkokäyttö kasvussa, vai onko tutkimusaineistoihin viittaaminen saavuttamassa suurempaa hyväksyntää hyvänä tieteellisen käytäntönä ei ole tiedossa, eikä sitä todennäköisesti voida tietää. (Borgman 2015, 223.)

4.4 Tutkimusaineistojen lataaminen

Tutkimusaineistojen lataaminen ei takaa sitä, että tutkimusaineistoja olisi käytetty (Late & Kekäläinen 2020). Lataukset kertovat kuitenkin jonkin asteisesta mielenkiinnosta aineistoja kohtaan.

Tutkimusaineistojen laskeutumissivuihin (landing page) kohdistuneet vierailut eli katselukerrat viittaavat aineistoon kohdistuvasta mielenkiinnosta sekä tietoisuudesta tutkimusaineistojen olemassaolosta. Aineistoihin kohdistuvat lataukset voidaan nähdä laskeutumissivujen saamia katselukertoja vahvempana mielenkiinnon osoittimena, sillä se viittaa halusta katsoa itse aineistoa. (Ball & Duke 2015.)

Vuonna 2014 ei ollut vielä kovinkaan montaa sähköistä arkistoa tutkimusaineistoille, jotka avoimesti jakoivat tietoa latauskerroista, mutta oli kuitenkin monia, jotka keräsivät lokitietoja latauksista sisäisesti (Costas et al. 2013). Esimerkiksi Swedish National Data Service kerää ainakin toistaiseksi vielä lataustietoja sisäisesti.

4.5 Tutkimusaineistojen jatkokäyttö yhteiskuntatieteissä

Yleisesti ottaen tutkimusaineistojen jatkokäyttöä on tutkittu erityisesti haastatteluilla (Borgman, Scharnhorst & Golshan 2019; Curty & Qin 2015; Faniel, Kriesberg & Yakel 2013; Roos, Kumpulainen & Järvelin 2008; Wallis, Rolando & Borgman 2013; Yoon 2016) sekä kyselyillä (Curty et al. 2017; Faniel, Kriesberg & Yakel 2015; Roos, Kumpulainen & Järvelin 2008; Tenopir et al. 2011; Tenopir et al. 2015; Yoon & Kim 2017) mutta myös etnografisilla menetelmillä (Wallis, Rolando & Borgman 2013) sekä määrällisillä menetelmillä (Bishop & Kuula-Luumi 2017; Late & Kekäläinen 2020).

Yhteiskuntatieteissä tutkimusaineistojen uudelleenkäyttöä on tutkittu paljon haastatteluin ja kyselyin, mutta viime vuosina on tehty myös tutkimuksia määrällisin menetelmin aineistoja koskevien lokitietojen avulla. Kysely- ja haastattelututkimuksissa on keskitytty paljolti tutkijoiden asenteisiin, oletuksiin, kokemuksiin ja käytänteisiin tutkimusaineistojen uudelleenkäytön suhteen (ks. Curty et al. 2017; Curty & Qin 2015; Faniel, Kriesberg & Yakel 2013; Faniel, Kriesberg & Yakel 2015; Yoon 2016; Yoon & Kim 2017). Määrälli-

sissä tutkimuksissa (ks. Bishop & Kuula-Luumi 2017; Late & Kekäkälinen 2020) on tarkasteltu niin määrällisten kuin laadullistenkin yhteiskuntatieteellisten tutkimusaineistojen vaikuttavuutta aineistojen latausmäärien ja niiden saamien viittausten perusteella. Tutkimuksissa kartoitettiin myös aineistoja ladanneiden käyttäjien demografioita sekä heidän syitä aineistojen lataamiselle.

Kysely- ja haastattelututkimuksissa keskeisimmät havainnot koskivat tutkijoiden omia intressejä ja asenteita tutkimusaineistojen jatkokäytön suhteen, mutta myös heidän kollegoiden ja ympäröivän tiedeyhteisön sekä tieteenalan asenteita uudelleenkäytöstä. Tutkijoille kohdistettu tuki tutkimusaineistojen uudelleenkäytön suhteen, sekä sähköisten arkistojen tärkeys tässä kokonaisuudessa olivat myös vahvasti esillä.

Tutkijoiden omat intressit ja asenteet vaikuttivat siihen, kuinka he näkivät ja kokivat tutkimusaineistojen uudelleenkäyttämisen. Tätäkin enemmän heidän näkemyksiinsä vaikuttivat kollegoiden, ympäröivän tiedeyhteisön ja oman tieteenalan suhtautuminen, kokemukset ja asenteet tutkimusaineistojen uudelleenkäytön, siihen suunnatun tuen ja infrastruktuurin kannalta. (Curty et al. 2017; Curty & Qin 2015; Yoon & Kim 2017.) Tutkimusaineistojen jatkokäyttöön suunnattu tuki, ja erityisesti tarve tälle tuelle koettiin tärkeänä tekijänä jatkokäytön kannalta. Tuki jatkokäytölle nähtiin suurelta osin sähköisten arkistojen tehtävänä, mutta myös esimerkiksi tieteelliset kirjastot ja muut tiedeyhteistön tahot voisivat tukea tarjota. (Faniel, Kriesberg & Yakel 2013; Faniel, Kriesberg & Yakel 2015; Yoon 2016; Yoon & Kim 2017.)

Tutkimusaineistojen dokumentaation (aineisto ja metadata) laadulla koettiin olevan myös tärkeä rooli tutkimusaineistojen uudelleenkäytön yhteydessä (Curty & Qin 2015; Faniel, Kriesberg & Yakel 2015). Sähköisten arkistojen osalta tärkeimpiä uudelleenkäytön mahdollistavia palveluita olivat erityisesti tutkimusaineistojen dokumentaation parantaminen ja ylläpito, ja tätä kautta käytettävyyden parantaminen (Curty & Qin 2015; Yoon 2016; Yoon & Kim 2017), mutta myös aineistojen saavutettavuuden parantaminen (Yoon 2016; Yoon & Kim 2017).

Määrällisissä tutkimuksissa Bishop ja Luumi-Kuula (2017) keskittyivät laadullisten tutkimusaineistojen uudelleenkäyttöön Isossa Britanniassa ja Suomessa. Tarkastelun kohteela olivat kaksi sähköistä arkistoa ja niiden aineistot: UK Data Service ja Tietoarkisto. Laten ja Kekäläisen (2020) tutkimus oli jatke Bishopin ja Luumi-Kuulan kontribuutiolle, eroten aiemmasta ottamalla tarkasteltavaksi uuden aineiston Tietoarkistosta, jossa mukana olivat myös määrälliset aineistot. Molemmista tutkimuksista nousi esille, että yhteiskuntatieteellisten tutkimusaineistojen jatkokäyttö on kasvussa, ja molemmissa aineistoissa suurin syy aineistojen käytölle oli opetustehtävät ja opinnäytetyöt (Bishop & Kuula-Luumi 2017; Late & Kekäläinen 2020).

Bishop ja Kuula-Luumi (2017) myös havainnoivat, että peräti 3/5 UK Data Servicen aineistoista sai osakseen käyttöä. He myös havaitsivat tutkimusaineistojen jatkokäytön ympärillä olevien käytäntöjen kehittyneen. Late ja Kekäläinen (2020) huomioivat aineistossaan, etteivät ladatuimmat tutkimusaineistot edustaneet viitatuimpia, mutta viitattujen aineistojen osalta määrälliset aineistot ovat edustetumpia kuin laadulliset. He havaitsivat myös, että niin viittausten laatu kuin käytänteetkin ovat kehittymässä.

Tutkielmani jatkaa osaltaan Bishopin ja Kuula-Luumin (2017) sekä Laten ja Kekäläisen (2020) tutkimusten jalanjäljissä: tarkastelen yhteiskuntatieteellisten tutkimusaineistojen jatkokäytön astetta määrällisestä näkökulmasta. Tutkimusta rajaavana sähköisenä arkistona toimii edellä mainituista poiketen Swedish National Data Service. Tutkimus eroaa myös siinä, etten tarkastele tutkimusaineistojen lataajia tai käyttäjiä.

5 TUTKIMUSASETELMA JA TOTEUTUS

Tämän Pro-gradu -tutkielman tutkimusosuus voidaan nähdä empiirisenä perustutkimuksena, joka on toteutettu määrällisesti (kvantitatiivisesti) havainnoituna, kuvailevana retrospektiivisenä pitkittäistutkimuksena. Tällaiselle tutkimukselle tyypillisiä piirteitä ovat suuret tutkimusaineistot numeerisine mittaustuloksineen, joista havaittuja ilmiöitä kuvataan numeerisesti pitemmältä aikaväliltä, takautuvasti. Kuvailevalle tilastotieteelle on myös tyypillistä esittää tiedot ja havainnot tiivistetysti taulukoina ja graafisina kuvioina. (Holopainen & Pulkkinen 2002, 18; Nummenmaa, Holopainen & Pulkkinen 2014, 15-17.)

Esittelen luvussa 5.1 tutkimuskysymykset, luvuissa 5.2 ja 5.3 esittelen tutkimuksen kannalta relevantit organisaatiot niiden perusperiaatteiltaan ja palveluiltaan. Luvussa 5.4 ja sen alaluvuissa käsittelen tutkielman tutkimusaineiston keruuta ja käsittelyä. Luvussa 5.5 esittelen tutkimusmenetelmäni.

5.1 Tutkimuskysymykset

Tutkielman teemana voidaan nähdä olevan kansainvälinen olettamus ja visio tutkimusaineistojen julkaisemisen, jakamisen ja jatkokäytön yleistymisestä sekä kasvusta. Siinä missä aineistojen jakaminen ja jatkokäyttö, sekä nämä mahdollistava infrastruktuuri käytäntöineen on vankemmalla pohjalla lukuisissa luonnontieteissä, ovat nämä seikat vasta lähivuosina yleistyneet ja kehittyneet laajemmassa mittakaavassa huomioimaan myös muita tieteenaloja, kuten humanistisia tieteitä ja yhteiskuntatieteitä. Tämän tutkielman tarkastelun kohteena ovatkin yhteiskuntatieteellisiin tutkimusaineistoihin kohdistuva mielenkiinto, sekä niiden potentiaalinen jatkokäyttö, mutta osaltaan myös tätä toimintaa ympäröivä ja yhtä lailla kehittyvä infrastruktuuri sekä käytännöt.

Tutkielman täsmällisempänä rajauksena on sen kohdistuminen Swedish National Data Service (SND) –tutkimusaineistoinfrastruktuuriin ja tämän ylläpitämiin yhteiskuntatieteellisiin tutkimusaineistoihin. SND:n yhteistyö tutkimusaineistoille pysyväistunnisteita myöntävän DataCiten kanssa antaa tutkielman kautta myös kuvaa siitä, minkälaisia avoimia menetelmiä ja mahdollisuuksia tutkimusaineistojen ja niitä ympäröivän toiminnan analysoinnille on, ja on kehitteillä.

SND valiutui tarkastelun kohteeksi sijaintinsa vuoksi varteenotettavana pohjoismaisena toimijana. Suomen näkökulmasta yhteiskuntatieteellisten tutkimusaineistojen jatkokäyttöä on jo tämän tutkielman tapaan tarkasteltu (ks. luku 4.5). Myös SND:n yhteistyö DataCiten kanssa puolsi tätä valintaa, mahdollistaen aineistojen osittaisen keruun ohjelmallisesti avointen rajapintojen kautta.

Arvioin SND:n ylläpitämiin yhteiskuntatieteellisiin tutkimusaineistoihin kohdistuvaa mielenkiintoa aineistojen saamien latausten kautta, ja aineistojen jatkokäyttöä arvioin niiden saamien viittausten perusteella. Viittauksia koskevat tiedot tulevat osaltaan DataCiten (ja Crossrefin) prototyyppivaiheen palvelusta Event Data. Pyrin Event Datan avulla selvittämään SND:n viitatuimmat tutkimusaineistot. Tämän lisäksi tarkastelen viittauksia myös Google Scholar ja Web of Science -palveluiden kautta. Hankin tutkielmassa käytetyt aineistot SND:n kautta sähköpostitse sekä DataCiten tuottamien palveluiden kautta ohjelmallisesti, sekä manuaalisesti Google Scholarin ja Web of Science -tietokannan kautta. Aineistojen avulla pyrin vastaamaan seuraaviin tutkimuskysymyksiin:

1. Kuinka paljon yhteiskuntatieteellisiä tutkimusaineistoja on ladattu SND:stä vuosina 2015-2021?
    - Onko mielenkiinto aineistoja kohtaan kasvanut vuosina 2015-2021?
2. Onko latausmäärissä havaittavissa selkeää eroa ladattujen aineistojen avoimuuden asteen suhteen?
3. Kuinka paljon 10 ladatuinta ja 10 Event Datan mukaan viitatuinta tutkimusaineistoa on saanut viittauksia?
    - Ovatko ladatuimmat aineistot myös viitatuimpia?
4. Miten SND:n ylläpitämiin yhteiskuntatieteellisiin tutkimusaineistoihin on viitattu?
    - Minkälaisista julkaisuista ja milloin?

5.2 Swedish National Data Service (SND)

SND on tutkimusaineistoinfrastruktuuri, jonka pääasiallinen tehtävä on tukea tutkimusaineistojen ja niihin liittyvien materiaalien saavutettavuutta, säilytystä ja uudelleenkäytettävyyttä. Mahdollistaakseen tämän, SND tarjoaa työkalut ja tuen yksityiskohtaisen metadatan tuottamiseen, sekä uniikkien pysyväistunnisteiden rekisteröintiin tutkimusaineistoille. SND tarjoaa myös sähköisen arkiston eli repositorion, joka takaa aineistojen säilyvyyden ja tavoitettavuuden. SND:n periaatteellisena tavoitteena on, että tutkimusaineistot olisivat niin avoimia kuin mahdollista ja vain niin suojattuja kuin tarpeellista. (SND 2021a; Swedish Research Council 2021.)

SND on ollut Swedish Research Councilin (SRC) toimeksiantona jo vuodesta 2008, ja on edelleen SRC:n rahoittama. SND muodostettiin Swedish Social Science Data Service -palvelun pohjalta. Vuodesta 2018 SND:n johtaminen muutettiin konsortiovetoiseksi. Vastuussa on tällä hetkellä yhdeksän ruotsalaista yliopistoa, joista Göteborgin yliopisto toimii isännöivänä tahona, ja SND:n pääkonttori sijaitsee myös Göteborgissa. Konsortiojäsenet toimivat myös neuvoa-antavina tahoina tutkijoille ja tiedeyhteisölle kokonaisuudessaan sekä SND:n muille jäsenille. (SND 2021a; SND 2021b; Swedish Research Council 2012; Swedish Research Council 2021.)

Konsortion lisäksi SND:hen kuuluu tällä hetkellä 35 verkostojäsentä, jotka koostuvat ruotsalaisista yliopistoista ja julkisista tutkimuslaitoksista. Kukin verkoston jäsen sitoutuu ylläpitämään paikallisia tutkimusaineiston hallintaan erikoistuvia yksiköitä. Niiden tehtävänä on auttaa tutkijoitaan tutkimusaineistojen saavutettavuudessa FAIR-periaatteiden mukaisesti. Tämä toiminta sisältää esimerkiksi koulutuksia ja neuvonantoa aineistonhallinnasta ja saavutettavuudesta, sekä aineistojen ja niiden metadatan kuvailusta ja tallennuksesta SND:n sähköiseen arkistoon. (SND 2021c)

SND:llä on myös CoreTrustSeal -sertifikaatti, promotoiden kestävää ja luotettavaa infrastruktuuria tutkimusaineistojen säilytykselle, saavutettavuudelle ja jatkokäytölle. SND

on ainoa sähköinen arkisto tutkimusaineistoille Ruotsissa, jolla on CoreTrustSeal. Suomessa sertifikaatin omaavat Tietoarkisto ja Kielipankki. (CoreTrustSeal 2021a; CoreTrustSeal 2021b; SND 2021d)

SND:n sähköisestä arkistosta on tällä hetkellä haettavissa 1718 tutkimusaineiston tiedot (SND 2021e). Tutkielman aineiston keruun ajankohtana (30.5.2021) SND ylläpiti 630 yhteiskuntatieteellistä tutkimusaineistoa, ja näistä 76 oli täysin avoimesti ladattavissa. Loput 554 on mahdollista saada käyttöönsä anomalla tätä SND:n kautta.

5.3 DataCite (& Crossref)

DataCite on vuonna 2009 perustettu johtava kansainvälinen yleishyödyllinen yhteisö, jossa on jäseniä yli 42 maasta. DataCiten pääasiallisena tehtävänä on tarjota pysyväistunniste DOI:ta (Digital Object Identifier) tutkimusaineistoille. DataCiten jäsenenä organisaatiot pystyvät rekisteröimään hallinnoimilleen aineistoillensa DOI:ta, samalla sitoutuen liittämään aineistoihinsa vähintään pakolliset metadataskeeman määrittämät kentät sekä asettamaan aineistojensa metadatan julkisesti avoimeksi. (DataCite 2012; DataCite 2021a; DataCite n.d.-b; SND 2021f)

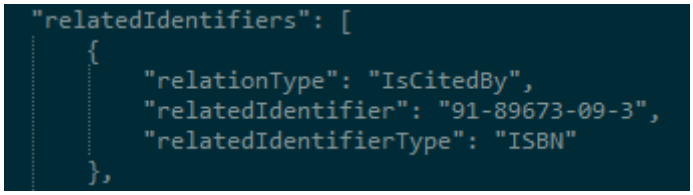
Tämän lisäksi DataCite myös kehittää ja tukee työkaluja sekä menetelmiä, jotka edesauttavat tutkimusaineistojen saavutettavuutta. DataCite pyrkii aktiivisesti kehittämään tutkimusaineistojen jakamista ja niihin viittaamista niin tiedeyhteisön sisällä kuin ulkopuolellakin. (DataCite n.d.-b; SND 2021f; SND 2021g.)

SND:n yhteistyö DataCiten kanssa antaa siis SND:n konsortion jäsenille mahdollisuuden rekisteröidä DOI-pysyväistunnisteita tutkimusaineistoillensa. Tämä edesauttaa aineistojen pysyvyyttä, löydettävyyttä, uudelleenkäytettävyyttä sekä niihin viittaamista. DataCiten jäsenten rekisteröidessä DOI:ta aineistoihinsa, on heidän myös tallennettava DOI ja relevantti metadata DataCiten metadatavarantoon (Metadata Storage). Tämän jälkeen metadata on avoimesti kaikkien ladattavissa (harvest) avointen ohjelmointirajapintojen kautta (SND 2021b).

DataCite on myös mukana kehittämässä prototyyppivaiheen palvelua nimeltä Event Data, jonka beta-versio julkaistiin vuonna 2017 (Crossref n.d.-g). Palvelua toteutetaan yhteistyössä Crossrefin kanssa, joka jakaa DOI-tunnisteita tutkimusjulkaisuille. Perinteisesti tieteellinen diskurssi on tapahtunut tieteellisessä sisällössä – esimerkiksi artikkeleiden välillä. Event Data -palvelussa on kiinnostuttu epätavanomaisemmasta kommunikaatiosta tieteellisen sisällön suhteen, esimerkiksi viittauksista artikkelien ja tutkimusaineistojen välillä sekä verkko-osoitteiden ja DOI-tunnisteiden välisistä, *altmetriikkaa* edustavista linkeistä. Altmetriikka on suhteellisen uusi bibliometriikan suuntaus, jossa tarkastellaan julkaisujen näkyvyyttä erityisesti sosiaalisessa mediassa (Forsman & Englud 2014). Näiden aktiviteettien välille muodostettuja linkkejä kutsutaan *(linkittäväksi) tapahtumiksi* (Linking Events). (DataCite n.d.-c; DataCite 2021d; Crossref 2020a; Crossref n.d.-b; Crossref n.d.-c.)

Tutkimusaineistoihin kohdistuvien viittausten ja altmetristen tietojen lisäksi DataCite tarjoaa jäsenilleen standardit ja ohjeet *käyttötapahtumien* (Usage Events) keruuseen ja ilmoittamiseen metadatan osaksi. Käyttötapahtumat kartoittavat jäsenien ylläpitämiin aineistoihin kohdistuvia lataus- ja katselutilastoja. (DataCite n.d.-d.)

Jokainen Event Datan tapahtuma (ei kata käyttötapahtumia) on niin kutsutun Agentin tuottama. Agentit ovat tietokoneohjelmia, jotka seulovat ja käyttävät erinäisten palveluiden rajapintoja, ja seuraavat löytämiään linkkejään arvioiden näiden sisältöä tapahtumien näkökulmasta. Agentit ovat joko DataCiten, CrossRefin tai niiden yhteistyökumppaneiden ylläpitämiä. DataCiten ja Crossrefin metadatasta kummunneet tapahtuvat ovat siis heidän jäsenten rekisteröimiä, joista Agentit sitten koostavat tapahtumat. (Crossref n.d.-b; Crossref n.d.-c; Crossref n.d.-d.)



KUVIO 1: DataCiten metadatassa olevat pakolliset kentät tapahtuman tuottamisen kannalta

DataCiten metadatassa on kolme metadatakenttää, josta tapahtumat johdetaan, mikäli kentät ovat olemassa. Crossrefin tapauksessa jäsenet listaavat tiedot rekisteröitävän artikkelin viittauksista, joista tapahtumat johdetaan. Kuviossa 1 on esimerkki DataCiten kentistä, jotka ovat relationType, relatedIdentifier ja relatedIdentifierType. RelationType kertoo tapahtuman suhteesta itse aineistoon, relatedIdentifier sisältää tunnisteen (esimerkiksi pysyväistunniste), ja relatedIdentifierType kertoo mikä tämä edellisen kentän tunniste on. Tapahtumadata on osana tätä tutkielmaa tarkasteltaessa tutkimusaineistoihin kohdistuvia viittauksia. (Crossref n.d.-e; DataCite 2021d.)

5.4 Tutkimusaineisto

Tutkielman aineistossa on kolme isompaa aineistoa: SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistojen kuvailutietojen metadatatietueet, SND:n yhteiskuntatieteellisten aineistojen latausmääriä koskevat lokitiedostot sekä SND:n yhteiskuntatieteellisten aineistojen tapahtumadata Event Data -palvelusta. Mukana on myös pienempi viittausten aineisto Google Scholarista ja Web of Sciencestä haettuna.

Taulukko 1 havainnollistaa tutkielman lopullisen aineiston määrää:

TAULUKKO 1: Tutkielman tutkimusaineisto

Tarkasteltava yksikkö	**Havaintoyksikköjä**	**Tiedosto**
Lataukset (avoimet, SND)	13874	Lataukset.CSV
Lataukset (suljetut, SND)	14194	Lataukset.CSV
Viittaukset (Event Data)	932	Viittaukset.CSV
Viittaukset (10 ladatuinta)	108	10ladatuinta.XLSX
Viittaukset (10 ”viitatuinta”)	55	10viitatuinta.XLSX

Taulukossa 1 Lataukset.CSV on koostettu SND:ltä saaduista latausten lokitiedostoista, Viittaukset.CSV sisältää viittaukselliset tapahtumat Event Data -palvelun kautta, ja 10ladatuinta.XLSX sekä 10viitatuinta.XLSX on koostettu manuaalisesti Google Scholar ja Web of Science -palveluista.

Avaan luvussa 5.4.1 tutkimusaineistojen keruuta ja luvussa 5.4.2 kerron aineiston käsittelystä.

5.4.1 Aineiston keruu

Tutkielman alkuperäinen ajatus oli hankkia suurin osa tutkimuskysymysten kannalta relevantista aineistosta ohjelmallisesti DataCiten avointen rajapintojen kautta, koska SND:n DataCite-jäsenyys sivusi tätä mahdollisuutta. Parhaimmassa tapauksessa rajapintojen kautta olisi voinut ladata tutkimusaineistojen metadatatietueet, lokitiedot latauksista ja katselukerroista (käyttötapahtumat, usage events) sekä lokitiedot aineistoihin kohdistuneista viittauksista tapahtumien (Event Data) muodossa.

Puolivälissä prosessia kuitenkin ilmeni, ettei SND ollut täysimittaisesti sisällyttänyt niitä tietoteknisiä mahdollisuuksia, joita DataCite tarjosi, osaksi toimintaansa. Latauksia ja katselukertoja koskevaa raportointia ja tiedonsiirtoa ei ollut ainakaan 30.5.2021 mennessä otettu käyttöön. SND:n yhteiskuntatieteellisiä aineistoja koskeva lokitiedosto latausten osalta saatiin SND:ltä sähköpostitse (SND 2021-j).

Event Data ei palveluna ollut vielä tutkielman toteutuksen aikana riittävä tuottamaan tarpeeksi vakaata ja kattavaa dataa, jolla olisi pystytty vastaamaan tämän tutkielman viittauskeskeisiin tutkimuskysymyksiin. Tästä johtuen ladatuimpia ja viitatuimpia tutkimusaineistoja tutkittiin tarkemmin Google Scholar ja Web of Science (WoS) -hakupalveluiden kautta.

SND itsessään ei tarjoa mitään rajapintoja palveluunsa, mutta yhteistyö DataCiten kanssa mahdollisti sen, että tutkimusaineistojen kuvailutiedot metadatatietueina olivat avoimesti ladattavissa DataCiten rajapintojen kautta. DataCite tarjoaa aineistojensa selaamiseen ja lataamiseen kolme erilaista avointa rajapintaa eri protokolliin ja arkkitehtuurimalleihin perustuen: OAI-PMH, GraphQL API ja REST API. DataCiten jäsenien tarpeisiin ovat myös rajapinnat EZ API ja MDS API. Rajapinnat mahdollistavat jäsenille helpomman, ohjelmallisen ja automaattisemman tavan rekisteröidä DOI-tunnisteita ja aineistojen metadataa sen sijaan, että he käyttäisivät verkkosovellus Fabrican käyttöliittymää. Event Datan tapahtumatietueita pystyi lataamaan Event Data API:n, REST API:n ja Graph QL API:n kautta. (DataCite n.d.-e; DataCite 2021f, DataCite 2021g, DataCite n.d.-h.)

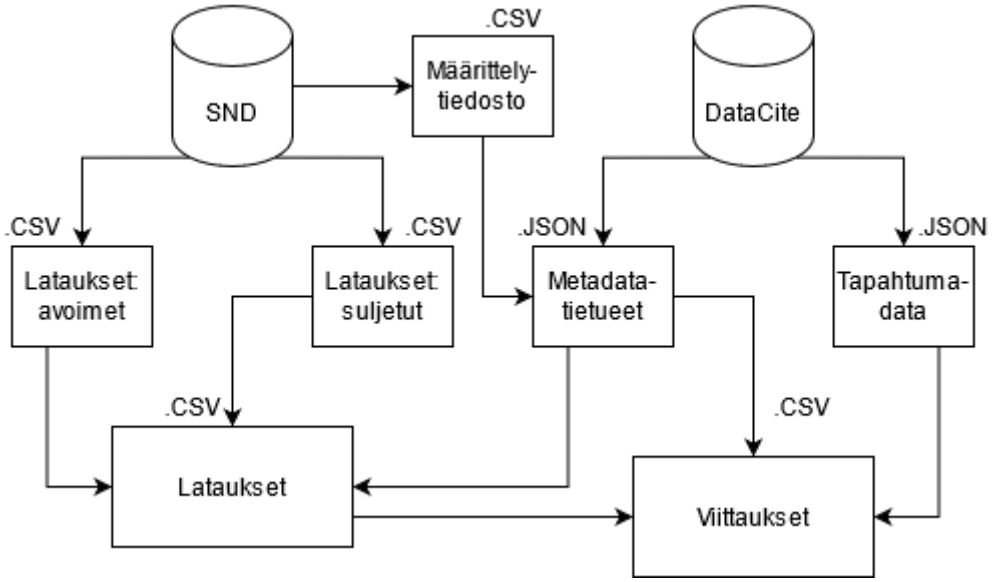
SND:n ylläpitämien tutkimusaineistojen metadatatietueet ladattiin DataCiten REST-rajapinnan kautta, ja tapahtumia koskeva data ladattiin Event Data API:n kautta. Lataukset toteutettiin Python (versio 3.7.4) -ohjelmointikielen avulla, käyttäen apuna Requests- ja JSON-kirjastoja. Kirjastot ovat jonkun muun toimesta toteutettua valmista koodia, jota muut voivat käyttää (Ryan 2020). Ladatut aineistot tulivat JSON muodossa. JSON ja CSV ovat yleisesti käytettyjä tiedon tallennus- ja siirtoformaatteja (P2PU n.d.). Metadatatietueet mukailevat DataCiten metadataskeemaa, ja tapahtumadata niin ikään omaa JSON-rakennettaan.

SND:ltä saatu latausten lokitiedosto oli sen sijaan CSV muotoisena. Se oli jaettu kolmeen osaan: avointen aineistojen lataukset, rajoitetusti saatavilla olevien (suljettujen) aineistojen lataukset sekä listaus niistä tieteenaloista ja asiasanoista, joiden avulla aineistojen yhteiskuntatieteellisyys SND:ssä määritetään.

Event Datan riittämättömyys tutkimusaineistojen saamien viittausten analysointiin johti siihen, että viittauksia tarkasteltiin manuaalisesti Google Scholarin ja WoS:n kautta. Google Scholar ja WoS valikoituvat kohdetietokannoiksi, sillä niitä on käytetty samaan tapaan, samankaltaisia aineistoja tutkivissa tutkimuksissa (ks. Bishop & Kuula-Luumi 2017; Late & Kekäläinen 2020). Tarkasteltavat tutkimusaineistot etsittiin Google Scholarista ja WoS:sta niiden englannin- sekä ruotsinkielisten nimien avulla fraasihakuina, joskin joissain tapauksissa mukana oli myös tutkimusaineiston tekijän nimi. Tämän lisäksi tuloksia haettiin myös tutkimusaineistojen DOI-tunnisteella. Jokainen hakutulos käytiin läpi, ja katsottiin, minkälaisesta julkaisusta oli kyse. Koska viittaus ei välttämättä kerro aineistojen käytöstä, jokaisesta julkaisusta tarkistettiin, käytettiinkö niissä oikeasti tutkimusaineistoa ja sen dataa. Tämän lisäksi tarkasteltiin, kuinka tutkimusaineistoihin viitattiin. Tutkimusaineistoihin viittaamista kategorisoitiin laadullisesti viittaustavan perusteella, ja sitä avataan lyhyesti tulosten yhteydessä luvussa 6. Myös viittaavien julkaisujen julkaisuvuosi otettiin ylös. Tiedot koostettiin Excel-tiedostoon.

5.4.2 Aineiston käsittely

SND:n ja DataCiten kautta hankituista aineistoista rikastettiin kaksi havaintomatriisia, joista toinen koski Event Datan tapahtumista suodatettuja viittauksellisia tapahtumia ja toinen latauksia SND:n yhteiskuntatieteellisten tutkimusaineistojen osalta. Käsittely tapahtui Python -ohjelmointikielellä, ja apuna oli myös aikaisemmin mainittujen kirjastojen lisäksi CSV-kirjasto. Python versio myös sama kuin aineistojen keruussa. Kuvio 2 hahmottaa aineiston keruun sekä käsittelyn prosessia:



KUVIO 2: Prosessikaavio SND:n ja DataCiten aineistojen koostamisesta

DataCitestä ladatut SND:n tutkimusaineistojen metadatatietueet kattavat kaikkien SND:n aineistojen tiedot. SND:ltä saadut latausten lokitiedot sisälsivät myös ”määrittelytiedoston” (kuvio 3), joka sisälsi listauksen yhteiskuntatieteiksi luokitelluista tieteenaloista sekä SND:n käyttämistä asiasanoista, joilla he kuvaavat yhteiskuntatieteellisiä tutkimusaineistoja. Määrittelytiedoston termistö pohjautuu “Standard för svensk indelning- av forskningsämnen 2011” -standardiin. Määrittelytiedoston avulla metadatatietueista suodatettiin pois kaikki paitsi yhteiskuntatieteellisten aineistojen tietueet. Lopuksi metadatatietueet rajattiin vielä ainoastaan niihin, joita SND ylläpitää ja hallinnoi. Lopputulema oli 630 yhteiskuntatieteellisen tutkimusaineiston metadatatietueet.

SND:ltä saatu latausten lokitiedosto koski ajanjaksoa 1.1.2015 – 25.5.2021. Yhteiskuntatieteellisten tutkimusaineistojen monimuotoisuus, verrattuna esimerkiksi astronomian puhtaasti kvantitatiivisiin ja erittäin homogeenisiin aineistoihin, vaikuttaa yhteiskuntatieteellisten tutkimusaineistojen avoimuuden asteeseen. Esimerkiksi luvuissa 2.4 ja 3.3. esille tuodut anonymisoinnin ongelmat saattavat olla läsnä, kuten myös tutkijoiden omat huolet, tarpeet ja edut. Muun muassa näistä syistä johtuen osa SND:n aineistoista on täysin avoimia, osa suljettuja. Suljettuihin aineistoihin käsiksi pääsy sekä mahdollinen jatkokäyttö vaatii erillisen anomuksen täyttöä SND:n palvelussa (SND 2021-

i). Tästä johtuen myös lataukset tulivat kahdessa osassa, koskien erikseen avoimiin aineistoihin kohdistuneita latauksia sekä suljettuihin aineistoihin kohdistuneita latauksia.

Molemmat lataustiedot täsmäytettiin metadatatietueiden dataan, ja lopuksi näistä aineistoista johdettiin yksi CSV muotoinen aineisto, jossa havaintoyksikkönä on yksi lataus. Latausten lokitiedot itsessään olivat melko suppeita sisällöltään (kuviot 3 ja 4), sisältäen lähinnä tutkimusaineiston SND-kohtaisen tunnuksen, ladatun datasetin numeron, mikäli kyseessä monesta datasetistä koostuva sarja sekä latausajankohdan.

Studie	Filnamn	Dataset	Version	Datum
snd0956	SND 0956-001-v1_1.zip	1	1.1	1.1.2015 4:53
snd0960	SND 0960-001-v2_1.zip	1	2.1	1.1.2015 9:25
snd0955	SND 0955-001-v1_1.zip	1	1.1	2.1.2015 4:06

KUVIO 3: SND:n avoimet lataukset

Avoimissa aineistoissa (kuvio 4) mukana oli myös tieto ladatusta tiedostosta ja versionumerosta, sillä käyttäjä valitsee itse mitä tiedostoja aineistosta haluaa ladata, mutta usein mukana on kuitenkin zip-tiedosto, joka sisältää kaiken aineiston. Suljetuissa aineistoissa (kuvio 5) näitä tietoja ei ole, ja voidaan olettaa, että jokaisen suljetun aineiston pyynnön yhteydessä on käyttäjälle annettu kaikki relevantti materiaali aineiston ymmärtämistä ja käyttöä varten.

Studie	Dataset	Datum
snd0822	1	6.1.2015
snd0905	1	9.1.2015
snd0843	1	13.1.2015

KUVIO 4: SND:n suljetut lataukset

Lataustiedostoissa oli myös poikkeamia. 102 avointen latausten havaintoyksikössä data oli korruptoitunutta, joten nämä jätettiin tarkastelun ulkopuolelle. Myös 636 avointa latausta ei täsmäytynyt metadatatietueisiin SND-kohtaisen tunnuksen *ja* versionumeron puolesta. Mutta koska tunnuksella täsmäytys onnistui, päädyttiin täsmäyttämään kyseiset lataukset relevanttien metadatatietueiden viimeisimpään versioon. 46 aineiston la-

taustiedot ilmenivät sekä avointen että suljettujen latausten lokitiedostossa. Syynä voivat olla muutokset aineistojen avoimuuden asteessa, tai virheet lokitiedostoissa. Näitä aineistoja koskevat tiedot otettiin kuitenkin mukaan tarkasteluun.

Lopulliset määrät analysoitaville latauksille olivat 13874 avointen aineistojen latausta sekä 14194 latausta suljetuille aineistoille aikavälillä 1.1.2015 – 25.5.2021. Avointen aineistojen latauksista reilusti suurin osa (n = 12001) kohdistui kokoaineistoihin, eli ziptiedostoihin. Aineistojen avoimuuden asteen aiheuttamista eroista lokitiedostojen tallennus ja/tai esittämistavoista johtuen, ei aineistoihin kohdistuneita latauksia voida suoraan verrata toisiinsa, mutta suuntaa antavasti kylläkin.

DataCiten Event Datan tapahtumien datasta suodatettiin tarkasteltavaksi ainoastaan viittauksiksi luokitellut tapahtumat. Metadatatietueiden avulla tapahtumat rajattiin myös koskemaan ainoastaan SND:n ylläpitämiä yhteiskuntatieteellisiä aineistoja. Näitä tietoja rikastettiin myös latauksista muodostetulla tiedostolla. Kuviossa 5 voidaan nähdä, miltä yksittäinen tapahtuma näyttää JSON-formaatissa.


```
"460": {
    "id": "864babc8-ed5c-4649-9040-94957ed26bd9",
    "type": "events",
    "attributes": {
        "subj-id": "https://doi.org/10.5878/000417",
        "obj-id": "http://books.google.com/books?id=fgn90dslg8qc&dq=isbn%3A3531145746&hl=&source=gbs_api",
        "source-id": "datacite-url",
        "relation-type-id": "is-cited-by",
        "total": 1,
        "message-action": "create",
        "source-token": "44fcac73-2945-4d04-9fcd-2ad679d7b695",
        "license": "https://creativecommons.org/publicdomain/zero/1.0/",
        "occurred-at": "2019-10-31T20:02:27.000Z",
        "timestamp": "2020-07-09T11:42:11.415Z"
    },
    "relationships": {
        "subj": {
            "data": {
                "id": "https://doi.org/10.5878/000417",
                "type": "objects"
            }
        },
        "obj": {
            "data": {
                "id": "http://books.google.com/books?id=fgn90dslg8qc&dq=isbn%3A3531145746&hl=&source=gbs_api",
                "type": "objects"
            }
        }
    }
},
```


KUVIO 5: Tapahtuman metadatan JSON-rakenne

Viittaustiedot tulevat Event Datan tapahtumiin tällä hetkellä DataCiten rekisteröityjen tutkimusaineistojen metadatasta, sekä Crossrefin rekisteröityjen tutkimusjulkaisujen

metadatasta. Event Datan Agentit muodostavat DataCiten ja Crossrefin metadatan avulla tapahtumaa kuvaavan tietueen (Kuvio 5), ja sisällyttää siihen tiedot muun muassa siitä, minkälainen tapahtuma on kyseessä ja milloin se on tapahtunut. (DataCite 2021d; DataCite 2020i.)

SND:n yhteiskuntatieteellisille tutkimusaineistoille oli kaiken kaikkiaan kertynyt 945 tapahtumaa, jotka voidaan palvelun kuvauksen mukaan lukea viittauksiksi. Jostain syystä kuitenkin 13 viittaavan tapahtuman kohdalla ei metadatatietueista löytynyt relatedIdentifiers -tietoja (ks. luku 5.3), joiden pohjalta tapahtuman olisi pitänyt syntyä, joten näitä 13 ei otettu mukaan tarkastelun. Jäljelle jäi 932 viittaavaa tapahtumaa, joista muodostettiin yksi CSV-muotoinen tiedosto, jossa yksi havaintoyksikkö on viittauksellinen tapahtuma SND:n ylläpitämään yhteiskuntatieteelliseen tutkimusaineistoon.

TAULUKKO 2: Kooste tutkimusaineiston isoimmista aineistoista

Tarkasteltava yksikkö	**Määrä**	**Tiedosto**
Avoimet lataukset	13874	Lataukset.CSV
Suljetut lataukset	14194	Lataukset.CSV
Viittaukset	932	Viittaukset.CSV

Taulukossa 2 on koostettuna tutkimuksessa käytetyt suurimmat tutkimusaineistot. Nämä aineistot ovat koneellisesti käsiteltyjä, ja osittain myös koneellisesti hankittuja. Edellä mainittujen aineistojen käsittelyn ja analysoinnin lomassa havaittiin tutkimuksen kannalta tarve vielä kahdelle aineistolle viittausten suhteen.

Edellisessä alaluvussa 3.2.3 mainittiin Event Datan sisältäneen heikkouksia tämän tutkielman rajatun näytteen suhteen, minkä takia aineistojen saamien viittausten tarkasteluun otettiin avuksi Google Scholar ja Web of Science. Event Datan viittauksellisia tapahtumia käsiteltäessä ja analysoitaessa seuraavat heikkoudet tulivat ilmi: viittauksien lähteet, viittauksien ajankohdan tunnistaminen sekä viittausmäärien paikkansapitävyys.

Kaikkien viittauksellisten tapahtumien lähteenä oli SND:n jäsenten rekisteröimä metadata, eli mikäli aineistoa ei olisi merkitty osaksi metadataa, tai päivitetty osaksi sitä, ei viittausta ilmenisi. Myös 97 % (n = 932) näistä viittauksellisista tapahtumista johdettiin muista pysyväistunnisteista, kuin DataCiten tai Crossrefin (+ yhteistyökumppaneiden) DOI-tunnisteista, mistä johtuen Event Datan Agentit eivät pystyneet tuottamaan ajankohtia näille viittauksille.

Viimeisin havaittu heikkous koski viittausten validiutta määrällisesti. En epäile viittausten validiutta aitouden puolesta, sillä 100 % viittauksellisista tapahtumista tuli kuitenkin SND:n jäsenten rekisteröimästä metadatasta, ja SND on varteenotettava ja luotettava taho. Ongelmia ilmeni kuitenkin Event Datan viittauksellisia tapahtumia analysoitaessa, jotka Agentit ovat muodostaneet edellä mainitusta metadatasta. Event Datan perusteella saatiin selville ne 10 tutkimusaineistoja, joilla oli potentiaalisesti eniten viittauksellisia tapahtumia, mutta tarkempi analyysi osoitti, että näihin kohdistuneet 165 viittauksellista tapahtumaa eivät olleet ongelmattomia. Peräti 30,3 % näistä tapahtumista johti toimimattomaan URL-osoitteeseen, ja 28,5 % tapahtumista oli duplikaatteja. Suurin osa toimimattoman URL-osoitteen omaavista tapahtumista vaikutti URL-osoitteen perusteella olevan myös duplikaatteja.

Event Datan avulla saatiin kuitenkin siis suuntaa antavaa dataa sen suhteen, mitkä SND:n ylläpitämistä yhteiskuntatieteellisistä tutkimusaineistoista olivat *potentiaalisesti* saaneet eniten viittauksia. Näiden 10 aineiston saamia viittauksia tutkittiin tarkemmin Google Scholar ja Web of Science -palveluiden kautta, joista kerättyjä viittauksia havainnollistetaan taulukossa 3. 10 potentiaalisesti viitatuimman aineiston saamista viittauksista koostettiin 55 havaintoyksikön Excel-tiedosto. Myös 10 ladatuimman tutkimusaineiston saamia viittauksia tutkittiin samaan tapaan, ja muodostettiin toinen 108 havaintoyksikön Excel-aineisto.

TAULUKKO 3: Google Scholarista ja Web of Sciencestä kerätyt viittaukset

Tarkasteltava yksikkö	**Havaintoyksikköjä**	**Tiedosto**
Viittaukset (10 ladatuinta)	108	10ladatuinta.XLSX
Viittaukset (10 ”viitatuinta”)	55	10viitatuinta.XLSX

5.5 Tutkimusmenetelmä

Tutkimuksen kohdepopulaationa eli perusjoukkona ovat yhteiskuntatieteelliset tutkimusaineistot, ja niiden tarkastelu tapahtuu aineistojen metadatatietueiden kautta. Maantieteellisen sijainnin myötä olen halunnut tarkastella pohjoismaista toimijaa, jolla on myös yhteistyökumppanuus DataCite -palvelun kanssa, ja tarkastelun kohteeksi päätyi ruotsalainen Swedish National Data Service (SND) -tutkimusaineistoinfrastruktuuri. Koska SND:n valinta oli harkinnanvarainen, on tutkimuksen perusjoukon osajoukko siis näyte, eikä puhdas otos. Otantayksikön muodostavat SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistot (ja niiden metadatatietueet), joista on yhdessä aineistojen latauksia koskevien lokitietojen sekä DataCiten Event Datan tuottaman viittausdatan avulla johdettu kaksi havaintomatriisia. Viittauksia tarkastellaan tämän lisäksi kahden pienemmän Excel-tiedoston avulla (ks. luku 5.4).

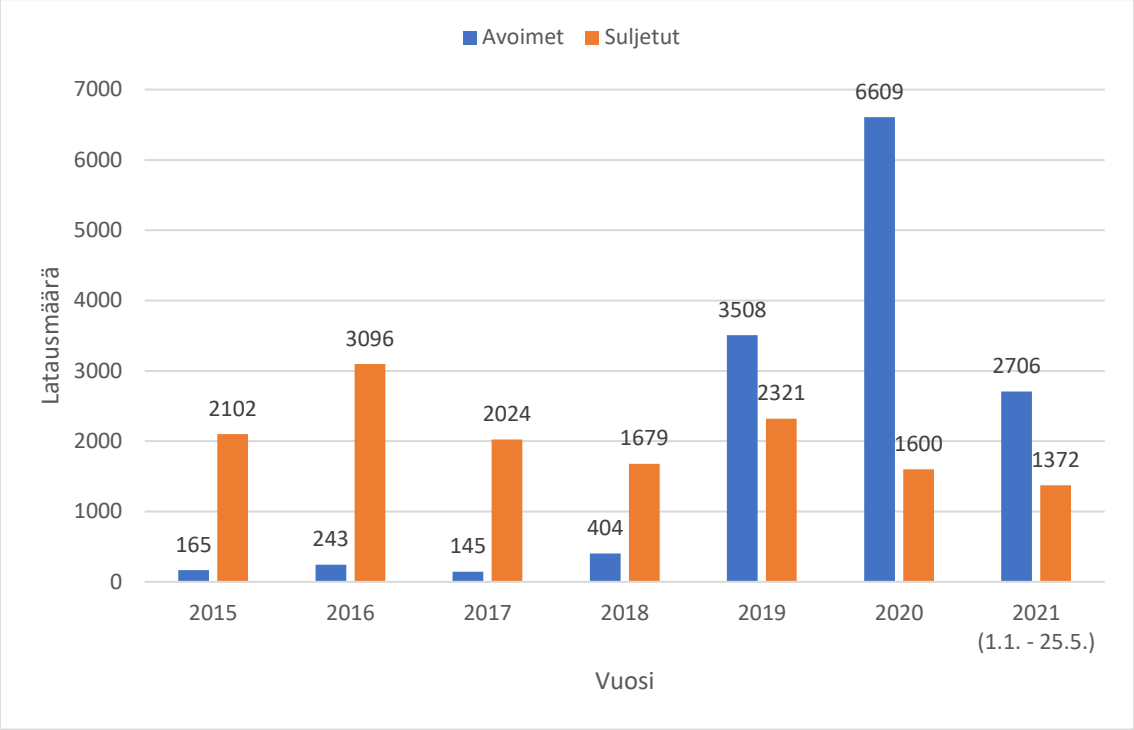
Tutkimuskysymysten, aineistojen ominaisuuksien sekä tarkasteltavien yksiköiden myötä analysoinnin toteutukseen koettiin riittävän tilastotieteelliset perusmenetelmät, kuten frekvenssitaulukot ja frekvenssijakaumat sekä ristiintaulukoidut yhteisjakaumat, joita on esitetty pylväskuvaajien muodossa. Tunnusluvuista on käytetty aritmeettista keskiarvoa sekä keskihajontaa apuna aineiston hahmottamisessa ja analysoinnissa. Aineistojen analysoinnissa ja visualisoinnissa oli apuna Pythonin (versio 3.7.4) lisäksi myös Excel (versio 2104, Build 13929.20386) sekä siihen asennettu Power Pivot -lisäosa, joka mahdollistaa kuvaajien ja taulukoiden muodostamisen useista eri lähteistä.

6 TULOKSET

Tarkastelen tuloksissa ensin latauksia, ja sitten viittauksia. Latauksissa olen kiinnostunut etenkin niiden määrällisestä muutoksesta vuosien 2015-2021 välillä, sekä siitä kuinka lataukset aineistoihin ovat jakautuneet. Tarkastelen näitä seikkoja tutkimusaineistojen avoimuuden asteen kautta, vertaillen tuloksia suuntaa antavasti täysin avoimesti saatavilla olevien ja rajatusti saatavilla olevien (suljettujen) aineistojen välillä. Viittauksia tarkastelen sen sijaan kehittyvien ja hiljalleen vakiintuvien käytäntöjen kontekstissa, ja niin ikään kehittyvässä infrastruktuurissa Event Data -palvelun kautta, mutta myös Google Scholarin ja Web of Sciencen avulla. Tarkastelen sekä latauksia että viittauksia tarkemmin 10 edustetuimman otantayksikön osalta.

SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistojen latausmääriä tarkastelen niiden vaikuttavuuden kannalta, kertoen suhteellisen vahvasta (ks. luku 4.4) mielenkiinnosta itse aineistoja kohtaan. Lataukset eivät kuitenkaan kerro sitä, minkälaista käyttöä aineistot saavat osakseen, tai saavatko ollenkaan. SND kerää tutkimusaineistoistaan latausten lokitietoja toistaiseksi sisäisesti, mutta niitä voi pyytää heiltä sähköpostitse.

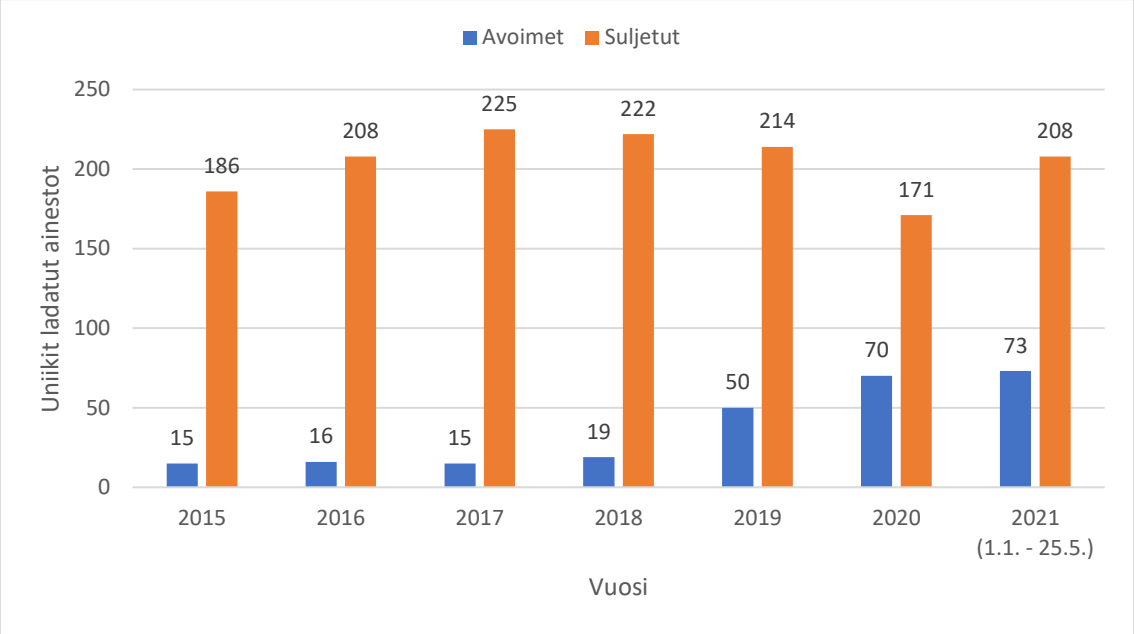
Näytteen otantayksikön muodostivat SND:n ylläpitämät yhteiskuntatieteellisen tutkimusaineiston metadatatietueet (N = 630), joista täysin avoimesti ladattavia on 76 (12,1 %). Näihin aineistoihin kohdistuneet lataukset ilmenivät latausten lokitiedostosta, ja se kattoi lataukset aikavälillä 1.1.2015 – 25.5.2021. Vajaa vuosi 2021 on mukana osassa kuvioita, antaen hieman kuvaa miltä kuluva vuosi vaikuttaa. Kokonaisina latausten tarkasteluvuosina (2015-2021) tutkimusaineistoja ladattiin 23896 kertaa; avoimia tutkimusaineistoja 11074 kertaa ja suljettuja 12822 kertaa. 393 (62,4 %) tutkimusaineistoa sai vähintään yhden latauksen. Kuvio 6 kuvastaa aineistoihin kohdistuneita latauksia ajanjaksolta aineistojen avoimuuden asteen suhteen.



KUVIO 6: SND:n ylläpitämien yhteiskuntatieteellisten aineistojen lataukset välillä 1.1.2015 - 25.5.2021

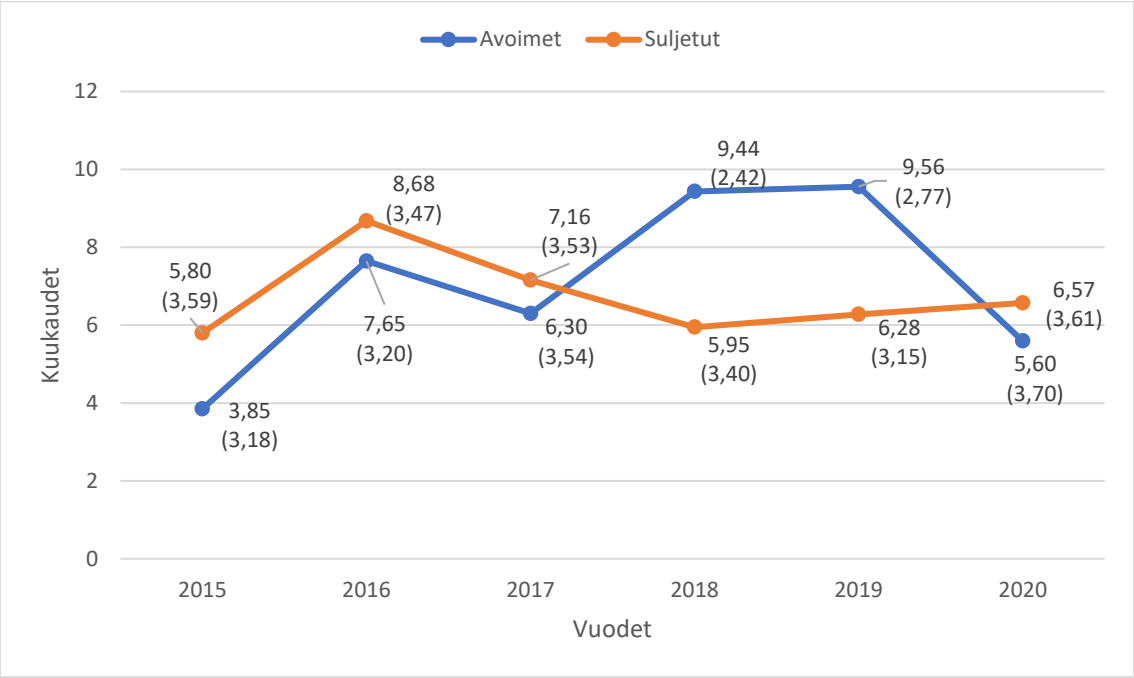
Kuviosta 6 nähdään, että vuosina 2015-2021 suljettujen aineistojen vaihtelu ei ole ollut kovinkaan suurta, voimme nähdä jakauman olevan melko suora. Vuosi 2021 näyttää ainakin toistaiseksi seuraavan samaa trendiä. Avointen aineistojen kohdalla on aineistojen latauksessa tapahtunut merkittävä nousu vuodesta 2018 eteenpäin, nousun jatkuen vielä vuonna 2019, joka oli vilkkain vuosi avointen aineistojen kohdalla 6609 latauksella. Vuonna 2019 avoimet lataukset nousivat 88,5 % verrattuna edelliseen vuoteen, ja vuonna 2020 ne nousivat vielä 46,9 %.

Vuosina 2015-2019 lataukset kohdistuivat suurelta osin vielä vain suljettuihin aineistoihin: 9858 latausta, joista 91,2 % (n = 8901) kohdistui suljettuihin tutkimusaineistoihin. Vuosi 2016 oli vilkkain vuosi suljettujen aineistojen osalta, 3096 latauksella.



KUVIO 7: Uniikit ladatut tutkimusaineistot

Kuvio 7 tuo latausmääriin perspektiiviä sen suhteen, kuinka moneen eri tutkimusaineistoon lataukset lopulta kohdistuivat. Suljettujen aineistojen osalta lataukset ovat melko tasaisesti kohdistuneet kokonaisina latausvuosina noin 200 aineistoon per vuosi. Avoimissa tutkimusaineistoissa kokonaisvaltaisesti aineistoihin kohdistuva mielenkiinto on kasvanut huomattavasti vasta vuonna 2019, noussen edellisestä vuodesta 62 %. Kuviosta 6 voi kuitenkin huomata, että latausmäärällisesti avoimet aineistot ovat viime vuosina (2019, 2020) menneet suljettujen aineistojen ohi, vaikka lataukset kohdistuvatkin pienempään määrään eri aineistoja.



KUVIO 8: Vilkkaimmat latauskuukaudet vuosilta 2015-2021.

Kuvio 8 havainnollistaa minä kuukausina (ja täten vuoden aikana) latauksia on tapahtunut eniten. Kuvaajissa suluissa oleva arvo on kunkin vuoden latauskuukausien keskihajonta, joka osoittaa kuinka paljon keskiarvon ympärille lataukset hajautuvat keskimääräisesti. Esimerkiksi avointen aineistojen ladatumpina vuosina 2019 ja 2020, on keskiarvollisesti eniten latauksia tapahtunut syksyllä ja keväällä, ja keskihajontakin on suhteellisen alhaista. Tosin 2020 vuoden syksyn hiljaisuutta latausten suhteen voi selittää kiihtynyt COVID-19 pandemia (Wikipedia 2021). Myös vuonna 2018, jolloin avoimet lataukset kasvoivat suhteellisesti melko paljon (64,1 %, kuvio 6), syksy oli vilkkainta aikaa latauksille. Suljetuissa tutkimusaineistoissa lataukset ovat pitkälti kohdistuneet kesäkuukausille.

Taulukko 4 havainnollistaa avoimiin ja suljettuihin tutkimusaineistoihin kohdistuneita latauksia ja latausmääriä vuosilta 2015-2021. Taulukoissa ensimmäisessä sarakkeessa on luokiteltu aineistojen saamia latausmääriä, ja toinen sarake osoittaa kuinka monta tutkimusaineistoa kuhunkin luokkaan kuuluu. Sarake ”Latauksia yhteensä” kertoo montako latausta tämän kyseisen luokan aineistot ovat saaneet yhteensä. Esimerkiksi aineistoija, joita ladattiin alle 30 (<30) kertaa oli 272 kappaletta, ja yhteensä ne saivat 1849 latausta.

TAULUKKO 4: Avointen ja suljettujen aineistojen latauskerrat

Latauksia	Ladattuja aineistoja		Latauksia yhteensä	
<30	272	69,21 %	1849	7,74 %
30-90	69	17,56 %	3866	16,18 %
>90	52	13,23 %	18181	76,08 %
Yhteensä	**393**	**100,00 %**	**23896**	**100,00 %**

Aineistot, joita ladattiin alle 30 kertaa, edustivat reilusti suurinta osaa (69,21%) ladatuista tutkimusaineistoista. Siltikin kokonaista latausmäärää tarkastellessa, edustivat nämä aineistot vain 7,74% (N = 23896) kaikista latauksista. Aineistot, jotka saivat osakseen yli 90 latausta edustivat vähemmistöä, ollen 13,23% kaikista vähintään yhden latauksen saaneista aineistoista (N = 393). Nämä aineistot olivat latausmääriltään kuitenkin niin sanottuja valaita, saaden 18181 latausta (76,08%) kaikista 23896 latauksesta aikaväliltä 2015-2021. Suurin osa latausmääristä kohdistuu siis pieneen osaan tutkimusaineistoista.

Taulukko 5 havainnollistaa ajanjaksolta 2015-2021 10 ladatuimman avoimen tutkimusaineiston ominaisuuksia. Ladatuin tutkimusaineisto “Inequality measures based on election data 1871 and 1892 for Swedish municipalities” kattaa noin 55,5 % (n = 4374) 10 ladatuimman aineiston latauksista. Avoimille aineistoille ei ole kertynyt juurikaan viittauksia Google Scholarin ja WoS -palveluiden kautta.

TAULUKKO 5: SND:n yhteiskuntatieteellisten avointen tutkimusaineistojen 10 ladatuinta aineistoa

Aineiston nimi	Julkaisuvuosi	Viittaukset	Lataukset
Inequality measures based on election data 1871 and 1892 for Swedish municipalities	**2019**	0	4374
The DREAM Dataset: Behavioural data from robot enhanced therapies for children with autism spectrum disorder	**2020**	0	826
Institutional Trust 2013	**2014**	1	530
Swedish Contextual Database for The Swedish Generations and Gender Survey and The International Generations and Gender Programme	**2018**	0	415
Violation and Satisfaction. A Sociology of Law Study of Non-Pecuniary Damages to Victims of Crime	**2016**	0	317
ESS 5 - European Social Survey 2010, Sweden	**2014**	1	317
ISSP 2015 - Work orientations IV: Sweden	**2017**	1	291
ISSP 2017 - Social Networks and Social Resources	**2018**	0	289
ESS 7 - European Social Survey 2014, Sweden	**2015**	0	265
European University Library Survey 2017 - User Surveys and other Methods for Library Development	**2020**	1	252
Yhteensä		**4**	**7876**

Taulukossa 6 tarkastellaan sen sijaan suljettujen tutkimusaineistojen 10 ladatuinta aineistoa. Suljettujen aineistojen ladatuin aineisto ” HUS - Household market and nonmarket activities” kattoi 43,9 % (N = 7602) 10 ladatuimman aineiston latauksista. Yhtenä selkeänä erona avoimiin aineistoihin on se, että 2 ladatuinta aineistoa ovat suhteellisen vanhoja, vuosilta 1993 ja 1994. Avoimissa aineistoissa vanhin ladattu aineisto 10 ladatuimman joukossa on julkaistu vuonna 2014. Aineistojen julkaisuvuosi on mukana kuitenkin vain suuntaa antavana määreenä, sillä aineisto voi olla tuotettu paljonkin aikaisemmin, ennen kuin se on julkaistu osaksi SND:n (tai sen edeltäjän) tietokantaa. Myös kumuloituvat aineistot tuottavat hälyä, esimerkiksi The National SOM Survey Cumulative Dataset -aineistoa on kerätty jo vuodesta 1986, mutta siltikin se koostetaan ja päivitetään osaksi tätä yhtä ja samaa tietuetta, päivittäen myös julkaisuvuoden. Siitä on kuitenkin mahdollista ladata yksittäisiäkin tarkasteluvuosia, joilla on omat pysyväistunnisteet.

Suljettujen aineistojen ladatuimpien osalta voimme havaita jo varsin suurta määrää viittauksia, yhteensä 104. Voidaan havaita, että vain 2 ladatuimmista tutkimusaineistoista jäi viittauksetta, ainakin tämän tutkielman menetelmien puitteissa. Göteborgin yliopiston SOM-instituutin (Society, Opinion and Media) kumulatiivinen SOM-kyselytutkimus vuodelta 2014 sai reilusti eniten viittauksia, 51 (49 %).

TAULUKKO 6: SND:n yhteiskuntatieteellisten suljettujen tutkimusaineistojen 10 ladatuinta aineistoa

Aineiston nimi	Julkaisuvuosi	Viittaukset	Lataukset
HUS - Household market and nonmarket activities	**1994**	7	3335
Popular movement archive 1881-1950	**1993**	8	1800
Swedish electoral data: General elections 1973-2006	**2016**	0	456
Swedish political party programs	**2006**	9	440
Swedish election manifestos	**2006**	6	374
The National SOM Survey Cumulative Dataset	**2020**	5	363
Swedish electoral data: Local elections 1976-2006	**2015**	0	325
Swedish election study 1998	**2002**	7	191
The National SOM Survey 2014	**2016**	51	160
Swedish election study 2006	**2012**	11	158
Yhteensä		**104**	**7602**

Niin avoimissa kuin suljetuissakin aineistoissa 10 ladatuinta tutkimusaineistoa kattaa latausmääriltään hyvin yli puolet kaikista latauksista: avoimissa 71,1 % (N = 11074, n = 7876), suljetuissa 59,3 % (N = 12822, n = 7602). Molemmilla aineistoilla on myös 1 aineisto, joka on saanut huomattavasti enemmän latauksia kuin muut. Tarkastellaan seuraavaksi ladatuimpien aineistojen saamia viittauksia koostetusti aineistojen avoimuuden tasosta riippumatta.

Taulukko 7 havainnollistaa edellä mainittujen ladatuimpien aineistojen saamien viittausten ominaisuuksia: minkälaisista julkaisuista tutkimusaineistoihin viitattiin, millä tavalla niihin viitattiin ja keskiarvollisesti minä vuonna? Taulukon sarake ”Julkaisut” kattaa kaikki muut julkaisut paitsi opinnäytteet, jotka ovat omana kategorianaan. Tässä tutkimuksessa opinnäytteisiin lasketaan kandidaatin tutkinnot, maisterivaiheen työt sekä väitöskirjat.

TAULUKKO 7: Ladatuimpien tutkimusaineistojen saamat viittaukset

Viittavan aineiston tyyppi	Viittaustapa	Määrä	
Julkaisu		**63**	**100 %**
	Tarkka viittaus	15	23,8 %
	Viittaus julkaisuun	1	1,6 %
	Virheellinen viittaus	1	1,6 %
	Maininta lähteenä	44	69,8 %
	Alaviitemaininta	2	3,2 %
Opinnäyte		**45**	**100 %**
	Tarkka viittaus	34	75,6 %
	Viittaus julkaisuun	2	4,4 %
	Virheellinen viittaus	2	4,4 %
	Maininta lähteenä	7	15,6 %
Yhteensä		**108**	

Avaan viittaustapojen kategorioita taulukossa 8. Viittaustapojen kategoriat ovat laadullinen analyysi Google Scholarin ja Web of Sciencen kautta havaituista tutkimusaineistoihin kohdistuneista viittauksista. Viittaustapojen selityksessä on myös mukana tieto siitä, tuleeko näissä tavoissa tutkimusaineistojen tekijät esille. Tätä tietoa käytettiin avuksi vain viittaustapojen kategorisoinnissa, eikä sitä analysoida tämän tutkielman puitteissa.

Ladatuimmat aineistot ovat saaneet yhteensä 108 viittausta, joista 63 (58,3 %) tulee tieteellisistä julkaisuista. Tarkkaa viittausta voidaan pitää parhaimpana osoituksena tutkimusaineistojen käytöstä, analogiana tieteellisten julkaisujen puolelta (ks. luku 4.3). 108 viittauksesta 49 (45,4 %) oli tarkkoja viittauksia. Julkaisujen kohdella tarkkojen viittausten (n = 15) ajankohtana oli keskimääräisesti 2016, ja opinnäytteiden kohdalla (n = 34) 2018.

TAULUKKO 8: Viittaustapojen selitykset

Viittaustapa	**Selitys**
Tarkka viittaus	Viittaus suoraan itse tutkimusaineistoon, eli lähdeluettelossa oleva täydellinen viittaus (tekijät tulevat esille)
Viittaus julkaisuun	Viitataan siihen julkaisuun, jossa aineiston sisältö on esitelty ja käytetty ensimmäisen kerran (tekijät tulevat esille)
Virheellinen viittaus	Viitataan tutkimusaineistoon, mutta puuttuu esimerkiksi tekijä (tekijät eivät tule esille)
Maininta lähteenä	Maininta siitä, mikä on käytetyn aineiston lähde, mutta ei lähdeluettelossa (tekijät eivät tule esille)
Alaviitemaininta	Erillinen maininta käytetystä tutkimusaineistosta alaviitteessä, mutta ei lähdeluettelossa (tekijät tulevat esille)

Viittaukset tutkimusaineistoja esitteleviin ja läpikäyviin julkaisuihin oli vähäisiä, kuten myös alaviitemaininnat sekä virheelliset viittaukset. Sen sijaan tutkimusaineistojen maininnat lähteenä esimerkiksi kuvaajien ja taulukkojen yhteydessä olivat suurin viittaustapa, edustaen 51 (47,2 %) havaintoyksikköä. Suurin osa näistä tuli julkaisujen osalta (n = 44), ja julkaisujen julkaisuvuodet olivat tässä viittaustavassa keskimäärin 2016. Opinnäytteet edustivat siis vain 7 tapausta tässä, ja niiden julkaisuvuosi keskimäärin 2014.

SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistojen 10 viitatuimman aineiston selvittämiseen käytettiin suuntaa antavasti apuna Event Data -palvelua, tiedostaen sen heikkoudet (ks. luku 5.4). Näiden 10 potentiaalisesti viitatuimman aineiston saamia

viittauksia ja mainintoja haettiin ladatuimpien tutkimusaineistojen tapaan Google Scholarista ja WoS:sta, ja ne ovat havainnoituna taulukossa 9.

TAULUKKO 9: Event Datan kautta havaitut 10 potentiaalisesti viitatuinta tutkimusaineistoa

Aineiston nimi	Avoimuus	Julkaisuvuosi	Lataukset	Viittaukset
The Comparative Parliamentary Democracy Data Archive	Avoin	2011	239	13
The National SOM Survey 1995	Suljettu	1997	34	13
Swedish journalist 2005	Suljettu	2010	0	8
The Regional Western Sweden SOM Survey 1998	Suljettu	2001	6	6
The National SOM Survey Cumulative Dataset	Suljettu	2013	363	5
Swedish journalist 2011	Suljettu	2013	2	3
The Regional Western Sweden SOM Survey 1999	Suljettu	2001	8	3
The Regional Western Sweden SOM Survey 2000	Suljettu	2002	7	2
A kinder, gentler democracy?	Suljettu	2015	1	2
Local elections 1979	Suljettu	1982	4	0
Yhteensä			**664**	**55**

Taulukosta 9 voidaan jo suoraltaan nähdä, etteivät Event Datan ehdottamat aineistot mitä todennäköisimmin edusta viitatuimpia aineistoja, eivät ainakaan viitatuinta, sillä suljettujen tutkimusaineistojen ladatuin sai vähintään 51 viittausta. Tästä huolimatta, voimme tarkastella näitä aineistoja ja niiden saamia viittauksia samaan tapaan kuin ladatuimpien aineistojen viittauksia. Suurempi määrä havaintoyksikköjä auttaa johtopäätösten teossa tutkimusaineistoihin kohdistuvien viittauskäytänteiden ja niiden kehityksen osalta.

Event Datan kautta koostettujen tutkimusaineistojen saamat viittaukset ovat taulukossa 10 esitetty ladatuimpien aineistojen tapaan.

TAULUKKO 10: Google Scholarista ja WoS:sta poimittujen viittausten ominaisuuksia

Viittaavan aineiston tyyppi	Viittaustapa	Määrä	
Julkaisu		**39**	**100 %**
	Tarkka viittaus	1	2,6 %
	Viittaus julkaisuun	13	33,3 %
	Maininta lähteenä	24	61,5 %
	Alaviitemaininta	1	2,6 %
Opinnäyte		**16**	**100 %**
	Tarkka viittaus	5	31,3 %
	Viittaus julkaisuun	8	50 %
	Maininta lähteenä	3	18,7 %
Yhteensä		**55**	

Viittaukset olivat suurelta osin viittauksia tutkimusaineistoja käsitteleviin julkaisuihin (20, 36,4 %) sekä mainintoja lähteenä (27, 49,1 %). Tämän kaltaisia viittauksia on tapahtunut eniten vuosina 2013-2015. Selkeimmät tapaukset, eli tarkat viittaukset itse tutkimusaineistoihin edustivat vajaata 11 % (n = 6). Tarkat viittaukset ovat tapahtuneet vuosina 2019 ja 2020, ja 5 näistä on opinnäytetöistä.

Hahmottaakseen tätä viittausten kokonaisuutta paremmin, tarkastelen edellä esiteltyä kolmea viittausten taulukkoa yhtenä kokonaisuutena taulukossa 11. Opinnäytteiden osalta voidaan havaita viittaustapojen hieman tasaantuneen, mutta tarkat viittaukset ovat edelleen selkeästi suurin edustaja 61,4 %.

TAULUKKO 11: Ladatuimpien ja Event Datan kautta koostettujen tutkimusaineistojen saamat viittaukset

Viittaivan julkaisun tyyppi	Viittaustapa	Määrä	
Julkaisu		**101**	**100 %**
	Tarkka viittaus	15	14,9 %
	Viittaus julkaisuun	14	13,9 %
	Virheellinen viittaus	1	1 %
	Maininta lähteenä	68	67,3 %
	Alaviitemaininta	3	3 %
Opinnäyte		**57**	**100 %**
	Tarkka viittaus	35	61,4 %
	Viittaus julkaisuun	10	17,5 %
	Virheellinen viittaus	2	3,5 %
	Maininta lähteenä	10	17,5 %
Yhteensä		**158**	

Julkaisujen kohdalla maininta tutkimusaineiston lähteenä olosta esimerkiksi taulukoiden yhteydessä on edelleen ylivoimaisesti suurin viittaustapa 67,3 %:lla. Tarkat viittaukset

tutkimusaineistoihin eivät kasvaneet Event Datan osoittamien aineistojen viittausten analyysien mukaan ottamisella, mutta sen sijaan viittauksen julkaisuun kasvoivat 13 yksiköllä.

KUVIO 9: Tarkat viittaukset tutkimusaineistoihin vuosittain

Tarkasteltaessa kaikkia tutkimusaineistoihin suoraan kohdistuneita viittauksia vuosittain, voidaan jakaumasta (kuvio 9) havaita selkeää kasvua vuodesta 2014 eteenpäin. Käyn seuraavassa luvussa läpi tarkemmin tulosten merkitystä ja niiden luotettavuutta.

7 DISKUSSIO

Tutkielman tarkoituksena oli osaltaan selvittää yhteiskuntatieteellisten tutkimusaineistojen vaikuttavuutta, mutta myös ympäröivien infrastruktuurien ja käytäntöjen kehitystä. Vaikuttavuutta arvioitiin aineistojen saamien lataus- ja viittausmäärien perusteella. Siinä missä lataus kertoo mielenkiinnosta aineistoa kohtaan, enemmän kuin esimerkiksi katselukerta, on tutkimusaineiston saama viittaus todiste siitä, että aineistoa on käytetty (Ball & Duke 2015).

Tutkielma rajattiin koskemaan Swedish National Data Service (SND) tutkimusaineistoinfrastruktuurin ylläpitämiä yhteiskuntatieteellisiä tutkimusaineistoja, sekä näihin kohdistuvia latauksia ja viittauksia. Tarkastelin viittauksia ensisijaisesti Google Scholarin ja Web of Sciencen (WoS) kautta, mutta myös DataCiten ja Crossrefin kehitteillä olevan Event Data -palvelun kautta. Tutkimusaineistot koostuivat SND:n koostamista aineistojen saamien latausmäärien lokitiedostoista, DataCiten kautta koneellisesti hankituista SND:n tutkimusaineistojen kuvailutietojen metadatatietueista sekä Event Data -palvelun kautta koneellisesti ladatuista viittauksia vastaavista tapahtumista. Viittauksia havainnoitiin tarkemmin koostamalla niistä tietoja Exceliin Google Scholarista ja WoS -palvelusta.

SND:n ylläpitämiä yhteiskuntatieteellisiä tutkimusaineistoja ladattiin vuosina 2015-2021 yhteensä 23896 kertaa, ja latausmäärät jakaantuivat tasaisesti avoimiin aineistoihin ja suljettuihin aineistoihin (11074/12822). Aineistosta (630 kpl) kuitenkin vain 76 oli avoimesti ladattavissa, ja loput vaativat käyttöluvan anomista SND:ltä. Molemmissa avoimuuden asteissa suurin osa latauksista kohdistui pieneen määrään aineistoja: 76,1 % prosenttia kaikista latauksista kohdistui 13,3 % prosenttiin ladatuista tutkimusaineistoista. 10 ladatuimman tutkimusaineiston tarkastelussa paljastui, että avoimilla ja suljetuilla aineistoilla oli molemmilla 1-2 niin sanottua valasta, jotka saivat osakseen suuren määrän latauksia.

Vaikka avointen ja suljettujen tutkimusaineistojen latausmäärät olivat lähellä toisiaan, olivat latausten jakaumat kuitenkin hyvin erilaisia. Siinä missä suljettujen aineistojen latausten jakauma oli hyvin tasainen kauttaaltaan, oli avointen aineistojen latauksissa havaittavissa merkittävää nousua vuosina 2019 ja 2020. Tarkastelun ulkopuolelle jäänyt vajaa vuosi 2021 näytti pysyvän suurin piirtein vuoden 2020 tasolla latausten suhteen. Kaiken kaikkiaan SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistojen latausmäärät ovat kasvaneet, joskin enemmän täysin avointen aineistojen osalta.

Koska lataustietojen yhteydessä ei ole tietoa siitä ketkä aineistoja ovat ladanneet, on vaikea arvioida esimerkiksi syytä avointen aineistojen suosion kasvulle viimeisen kolmen vuoden aikana. Avointen tutkimusaineistojen vilkkaimpina vuosina lataukset ovat keskiarvollisesti keskittyneet kevääseen ja syksyyn. Vuoden 2020 latausten painottuminen kevääseen voi olla seurausta COVID-19 pandemiasta. Nämä vuodenajat voivat viitata opiskelijoihin: opiskelua kesätauolla tai syksyn aloitus. Late ja Kekäläinen (2020, luku 4.5) epäilivät, että kasvu latausmäärissä voisi myös yleisemmin kertoa kasvavasta tietoisuudesta avointen tutkimusaineistojen olemassaolon suhteen, sekä kehittyvistä käytänteistä tutkimusaineistojen jakamisen ja jatkokäytön suhteen, sekä niiden tuomista hyödyistä.

SND:n ladatuimpien yhteiskuntatieteellisten tutkimusaineistojen yhteydessä tarkasteltiin myös niiden saamia viittauksia. Aineistojen saamien viittausten perusteella pyrittiin tarkastelemaan niiden uudelleenkäyttöä. Ladatuimpien aineistojen suhteen reilusti suurin osa kohdistui suljettuihin aineistoihin. Tämä tulos on kuitenkin vain suuntaa antavaa havaintoa aineistojen avoimuuden asteen tuomista eroista, tutkielman menetelmien puitteissa. Yksi syy voi kuitenkin olla se, että suljettujen aineistojen ladatuimmat koostuvat lukuisista pitkän aikavälin kyselytutkimuksista.

Tutkimusaineistojen saamien viittauksien avulla tarkasteltavaa aineistojen jatkokäyttöä on myös pidettävä suuntaa antavana havaintona tutkielman näytteenkin sisällä. Syynä on se, että viittausten yhteydessä ei olla aina voitu varmuudella sanoa, onko tutkimusaineistoon viittaava julkaisu se julkaisu, jossa aineistoa on ensimmäisen kerran käytetty. Tällöin ei olisi kyse jatkokäytöstä, vaan tavallisesta käytöstä (ks. Pasquetto, Bernadette

& Borgman 2017). Tutkijan oman aineiston jatkokäyttö aineiston alkuperäisen käyttötarkoituksen ulkopuolella on kuitenkin tässä tutkielmassa katsottu jatkokäytöksi.

Viitatuimpia aineistoja, sekä viittauksia ylipäänsä, oli tarkoitus selvittää DataCiten ja Crossrefin prototyyppivaiheen palvelun ”Event Data” avulla. Valitettavasti Event Datan tuottaman datan analysointi paljasti, ettei se vielä tässä vaiheessa elinkaartaan ole tähän kykenevä. Event Dataa käytettiin kuitenkin suuntaa antavasti osoittamaan mitkä tutkimusaineistot potentiaalisesti voisivat olla SND:n viitatuimpia aineistoja. Viittausten syvempi tarkastelu niin Event Datan tuottamien tapahtumien metadatan kuin Google Scholarin ja WoS:n kautta selvensi kuitenkin, etteivät nämä aineistot ainakaan kokonaisuudessaan edustaneet SND:n viitatuimpia yhteiskuntatieteellisiä tutkimusaineistoja. Tutkimuskysymyksiä ajatellen ei siis voida sanoa, olivatko ladatuimmat myös viitatuimpia.

10 ladatuimman tutkimusaineiston sekä Event Datan osoittamien aineistojen saamat viittaukset yhdistämällä saatiin 158 havaintoyksikön otos viittauksista. Tämän otoksen puitteissa voidaan tarjota vastauksia tutkielman näytteen raameissa viimeiseen tutkimuskysymykseen, eli miten SND:n tutkimusaineistoihin viitattiin, minkälaisista julkaisuista ja keskimäärin milloin?

Suurin osa viittauksista (101 kpl) tuli julkaisuista kuten tieteellisistä artikkeleista ja erilaisista raporteista. Opinnäytteet kattoivat loput 57 viittausta. Siinä missä julkaisujen kohdalla yli puolet (67,3 %) viittauksista tuli tyypiltään ”mainintana lähteenä”, tarkoittaen ettei tutkimusaineistoon viitattu lähdeluettelossa, tulivat viittaukset opinnäytteistä laajalti (61,4 %) ”tarkkoina viittauksina”, joissa aineistoihin viitattiin suoraan lähdeluettelossa.

Julkaisujen kohdalla ”maininnat lähteenä” voidaan ajoittaa keskimäärin vuosiin 2015 ja 2016, kun taas julkaisuista tulleet ”tarkat viittaukset” (15 kpl) tulivat keskimäärin vuonna 2016. Opinnäytteistä tulleet ”tarkat viittaukset” tapahtuivat keskimääräisesti vuonna 2018, kun taas eivät niin tarkat viittaukset tulivat keskimäärin vuosina 2013 ja 2014. Tä-

män näytteen puitteissa tämä voi kertoa kehityksestä tutkimusaineistojen viittauskäytänteiden yhteydessä. Tämä vastaisi Bishopin ja Kuula-Luumin (2017) sekä Laten ja Kekäläisen (2020) havaintoja yhteiskuntatieteellisten tutkimusaineistojen jatkokäytön sekä jatkokäyttöä ja viittauskäytänteitä koskevan kehityksen kasvusta. Vaikuttavia tekijöitä voivat olla tutkimusaineistojen avoimuuden ja kreditoinnin ympärillä käytävä keskustelu ja kehitys, sekä esimerkiksi pysyväistunnisteiden aktiivisempi rekisteröinti tutkimusaineistoille, joka edesauttaa niihin viittaamista. Tarkastelun kohteena olevan SND:n varteenotettavuus ja luotettavuus tutkimusaineistoinfrastruktuurina voi myös osaltaan vaikuttaa tähän, sillä ne tarjoavat jokaisen hallinnoimansa tutkimusaineiston yhteydessä ohjeen siihen, kuinka aineistoon tulisi viitata.

Vaikka tutkielman rajoitteista ja heikkouksista on jo ollut mainintoja, on ne hyvä koostaa uudelleen tähän. Rajoitteita ovat muun muassa sen yleistettävyys niin tutkielman otoksellisen luonteen, kuin myös tarkasteltavien viitattujen tutkimusaineistojen määrän myötä. Koska tutkielman perusjoukon, yhteiskuntatieteellisten tutkimusaineistojen, osajoukko on harkinnanvaraisesti valittu näyte (puhtaan otoksen sijasta) rajoittuen SND:n ylläpitämiin yhteiskuntatieteellisiin tutkimusaineistoihin, on tutkimustulosten yleistäminenkin rajoitettua.

Myös mittausvälineiden luotettavuus on osaltaan rajoite. Event Data -palvelu oli tietoisesti mukana tutkielmassa sen avointen ohjelmointirajapintojen myötä, mahdollistaen suhteellisen automaattiset ja ohjelmalliset keinot sen tuottaman datan analysoimiseen. Event Data ei riittänyt kaikkien niiden tutkimuskysymysten selvittämiseen, joita tutkielmassa sen avulla pyrittiin avaamaan. Täten Google Scholaria ja WoS:ia käytettiin viimeisten analyysien toteuttamiseen viittausten osalta. Tämäkin osana rajoitetta, sillä ei voida taata, että Google Scholarista ja WoS:sta saatiin esille kaikki mahdolliset aineistoihin kohdistuneet viittaukset tarpeellisten haunrajausten myötä (ks. luku 5.4.1). Ei voida myöskään varmuudella sanoa, että kaikki edellä mainituista palveluista haetut viittaukset olisivat olleet uudelleenkäyttöä. Osa saattaa hyvinkin olla vain perinteistä käyttöä, jossa aineistoa on käytetty sen alkuperäisessä kontekstissa. Sekaannusta tässä aiheuttaa

esimerkiksi se, että tutkimusaineistojen tekijäksi on saatettu merkitä vain pelkkä yliopisto, tai alkuperäinen julkaisu ei käy selkeästi ilmi.

Rajoitteisiin lukeutuvat myös tutkielmassa käytettyjen tutkimusaineistojen sisältämät virheet ja omituisuudet, kuten korruptoituneet havaintoyksiköt ja puuttuneet kentät. Nämä pyrittiin parhaan mukaan kuitenkin ottamaan huomioon, ja poistamaan tarkasteluista, keskittyen ainoastaan eheään dataan.

8 PÄÄTELMÄT

Pyrin pro gradu -tutkielmassani tarkastelemaan yhteiskuntatieteellisiin tutkimusaineistoihin kohdistuvaa mielenkiintoa sekä niiden saamaa jatkokäyttöä. Jatkokäytön yhteydessä tarkasteltiin myös sitä, millä tavalla tutkimusaineistoihin viitattiin, karkeasti minkälaisista julkaisuista ja milloin. Asetelmaa rajattiin entisestään kohdistaen tutkimus Swedish National Data Service (SND) -tutkimusaineistoinfrastruktuurin ylläpitämiin yhteiskuntatieteellisiin aineistoihin.

Aloitin taustojen kartoituksen aiheen suhteen ehkä liiankin kaukaa, mutta kuitenkin luoden perustaa sille mitä tutkin, ja mistä on kyse. Luvut 1-3 tuovat esille tieteenkentän kehitystä ja perusedellytyksiä, pohjustaen kuinka nykyiseen tilanteeseen tutkimusaineistojen saavutettavuuden ja jatkokäytön osalta on päädytty. Luvussa 4 keskityn aineistojen lataamisen ja viittaamisen mittaamiseen ja merkitykseen, tarkastellen myös aiempaa tutkimusta aiheesta.

Tässä tutkielmassa yhteiskuntatieteellisten tutkimusaineistojen osakseen saamaa mielenkiintoa ja jatkokäyttöä tutkittiin määrällisin menetelmin latausten lokitiedoilla, Event Data -palvelun tapahtumadatalla sekä manuaalisesti kerätyillä viittaustiedoilla Web of Sciencen ja Google Scholarin kautta.

Vaikka aineiston rajaus, käytettyjen aineistojen heikkoudet ja käytetyt menetelmät rajoittavat tutkimustulosten tarkkuutta ja yleistettävyyttä, osoittavat tulokset kuitenkin, että mielenkiinto tutkimusaineistoja kohtaan on Ruotsissa SND:n kohdalla kasvanut merkittävästi viime vuosina. Tulokset viittaavat hieman siihen suuntaan, että latausmäärien kasvu voisi johtua suurelta osin korkeakouluopiskelijoista. Yhtä lailla myös jatkokäytössä oli tapahtunut kasvua, etenkin tarkkojen, lähdeluetteloon merkittyjen viittausten yhteydessä. Tarkat viittaukset olivat kasvussa ja selkeästi käytössä erityisesti opinnäytetöiden kohdalla, mukaan lukien väitöskirjat.

Tulokset ja niistä tehdyt johtopäätökset vastaavat esimerkiksi Bishopin ja Kuula-Luumin (2017) sekä Laten ja Kekäläisen (2020) suhteellisen uusia ja samankaltaisia, mutta laajempia tutkimuksia (ks. luku 4.5). Tutkimuksissa havaittiin avointen tutkimusaineistojen latausmäärien ja jatkokäytön olevan kasvussa yhteiskuntatieteissä, etenkin opetuksen ja opinnäytteiden yhteydessä. Tämä vahvistaa ajatustani opiskelijoiden osuudesta aineistojen kasvaneisiin latausmääriin, sekä opiskelijoiden osuutta aineistojen tarkempien viittauskäytänteiden yleistämisessä.

Tästä voidaan päätellä, että diskurssi, kehitys ja päätökset tutkimusaineistojen avoimuuden, saavutettavuuden, käytettävyyden ja jatkokäytön ympärillä ovat saavuttaneet niin korkeakoulut kuin sähköiset arkistotkin, ainakin opiskelijoiden, yhteiskuntatieteiden ja SND:n rajoissa. Tämä on tärkeää, koska se voi kertoa SND:n ylläpitämien yhteiskuntatieteellisten tutkimusaineistojen olevan laadultaan riittävän hyviä mahdollistaakseen jatkokäytön ja niihin viittaamisen. Tärkeää on myös tämä opiskelijoiden osuus tutkimusaineistojen jatkokäytössä, koska tämänhetkiset opiskelijat vievät tiedot ja taidot avointen tutkimusaineistojen hyödyistä ja niitä ympäröivistä käytänteistä eteenpäin uusille (ja vanhoille) sukupolville. Tämä kasvattaa yhteiskunnassa tietoisuutta tutkimusaineistojen avoimuudesta ja niiden tuomista hyödyistä.

Jotta aineistoihin kohdistuvista latauksista ja viittauksista saataisiin tarkempaa analyysiä ja laajempaa yleistystä, tulisi toteuttaa kattavampi tutkimus useamman kuin yhden sähköisen tutkimusaineistoarkiston aineistoista. Myös tarkemman ja tehokkaamman viittausten keruumenetelmän kehittäminen osaksi tutkimusta voisi parantaa tulosten vaikuttavuutta. Niin ikään analyysit suurta käyttöä osakseen saavista tutkimusaineistoista voisi auttaa ymmärtämään, mikä tekee osasta aineistoja muita suositumman.

LÄHTEET

Alex Ball, & Monica Duke (2015). 'How to Track the Impact of Research Data with Metrics'. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/track-data-impact-metrics (käytetty 1.9.2021)

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PloS one*, *6*(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., ... & Wouters, P. (2004). An international framework to promote access to data. https://doi.org/10.1126/science.1095958

Avoin tiede. (n.d.). What is open science? | Open Science. https://avointiede.fi/en/what-open-science (käytetty 13.7.2021)

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*, *7*(1), 215824401668513. https://doi.org/10.1177/2158244016685136

Borg, S. 2010. Keskeiset tutkimusrahoittajat tukevat tutkimusdatan avoimuutta. Tietoarkisto 30 (2/2010). http://www.fsd.uta.fi/tietoarkistolehti/30/paakirjoitus.html (käytetty 25.6.2021).

Borgman, C. L. (2010). Research Data: Who Will Share What, with Whom, When, and Why? *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.1714427

Borgman, C. L. (2015). Big data, little data, no data. *Scholarship in the networked world*.

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888-904. https://doi.org/10.1002/asi.24172

Botstein, D. (2010). It's the data! *Molecular Biology of the Cell, 21*(1), 4–6. https://doi.org/10.1091/mbc.e09-07-0575

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O. & et al. (2012). "Science as an open Enterprise." The Royal Society. https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf (käytetty 1.9.2021)

Boyle, J. (2013). Biology must develop its own big-data systems. *Nature News*, *499*(7456), 7. https://doi.org/10.1038/499007a

Brase, J., & Farquhar, A. (2011). Access to research data. *D-Lib Magazine*, *17*(1/2). https://doi.org/10.1045/january2011-brase

Brase, J. (2014). Making data citeable: DataCite. In *Opening Science* (pp. 327–329). Springer, Cham. https://doi.org/10.1007/978-3-319-00026-8_26

Buneman, P., Khanna, S., & Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. In *International conference on database theory* (pp. 316-330). Springer, Berlin, Heidelberg.

Burwell, S. M., Vanroekel S., Park T., & Mancini D. J. (2013). "Open Data Policy-Managing Information as an Asset." Executive Office of the President, Office of Management and Budget. https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf (käytetty 25.6.2021)

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(Suppl 15), S2. 10.1186/1471-2105-12-S15-S2

CESSDA Training Team. (2020). CESSDA Data Management Expert Guide. CESSDA ERIC. https://doi.org/10.5281/zenodo.3820473

Consultative Committee for Space Data Systems. (2012). "Reference Model for an Open Archival Information System (OAIS)." Issue 2. Consultative Committee for Space Data Systems.

CoreTrustSeal. (2019). Core Trustworthy Data Repositories Requirements 2020–2022 Extended Guidance. CoreTrustSeal Extended Guidance v2.0. https://www.coretrustseal.org/wp-content/uploads/2019/11/2019-10-CoreTrustSeal-Extended-Guidance-v2_0.pdf (käytetty 27.8.2021)

CoreTrustSeal. (2021-a). About – CoreTrustSeal. https://www.coretrustseal.org/about/ (käytetty 27.8.2021)

CoreTrustSeal. (2021-b). Core Certified Repositories – CoreTrustSeal. https://www.coretrustseal.org/why-certification/certified-repositories/ (käytetty 27.8.2021)

Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013). The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report.

Crossref. (2020-a). Eventdata – Crossref. https://www.crossref.org/services/event-data/ (käytetty 27.8.2021)

Crossref. (n.d.-b). About the Data – Event Data User Guide. https://www.event-data.crossref.org/guide/data/about-the-data/ (käytetty 27.8.2021)

Crossref. (n.d.-c). Introduction – Event Data User Guide. https://www.eventdata.crossref.org/guide/introduction/ (käytetty 27.8.2021)

Crossref. (n.d.-d). Crossref Metadata – Event Data User Guide. https://www.event-data.crossref.org/guide/sources/crossref/ (käytetty 27.8.2021)

Crossref. (n.d.-e). How Data Contributors, Sources and Agents Work – Event Data User Guide. https://www.eventdata.crossref.org/guide/sources/how-agents-work/ (käytetty 27.8.2021)

Crossref. (n.d.-f). Time – Event Data User Guide. https://www.eventdata.crossref.org/guide/data/time/ (käytetty 27.8.2021)

Crossref. (n.d.-g). Release history – Event Data User Guide. https://www.event-data.crossref.org/guide/history/ (käytetty 27.8.2021)

Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1-4. https://doi.org/10.1002/meet.14505101072

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PloS one*, *12*(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Daniels, M., Faniel, I., Fear, K., & Yakel, E. (2012). Managing fixity and fluidity in data repositories. In *Proceedings of the 2012 iConference* (pp. 279-286). https://doi.org/10.1145/2132176.2132212

DataCite. (2012). DataCite – International Data Citation: Business Model Principles. https://doi.org/10.5438/0007

DataCite. (2021.-a). DataCite Metadata Schema 4.4. https://support.datacite.org/docs/datacite-metadata-schema-44 (käytetty 29.8.2021)

DataCite. (n.d.-b). DataCite's Value. https://datacite.org/value.html (käytetty 29.8.2021)

DataCite. (n.d.-c) DataCite – EventData. https://datacite.org/eventdata.html (käytetty 29.8.2021)

DataCite. (2021-d). DataCite Event Data. https://support.datacite.org/docs/eventdata-guide (käytetty 29.8.2021)

DataCite. (n.d.-e). DataCite – APIS. https://datacite.org/integratorapis.html (käytetty 29.8.2021)

DataCite. (2021-f). Consuming Citations. https://support.datacite.org/docs/consuming-citations (käytetty 29.8.2021)

DataCite. (2021-g). Getting Started. https://support.datacite.org/docs (käytetty 29.8.2021)

DataCite. (n.d.-h). DataCite Fabrica. https://doi.datacite.org/ (käytetty 29.8.2021)

DataCite. (2020-i). RelationType for Citations and References. https://support.datacite.org/docs/relationtype_for_citation (käytetty 29.8.2021)

David, P. A. (2003). The economic logic of "open science" and the balance between private property rights and the public domain in scientific data and information: a primer. *The role of the public domain in scientific and technical data and information*, 19-34.

Dellavalle, R. P., Hester, E. J., Heilig, L. F., Drake, A. L., Kuntzman, J. W., Graber, M. & Schilling, L. M. (2003). Going, Going, Gone: Lost Internet References. Science 302(5646): 787–88. https://doi.org/10.1126/science.1088234

Dorch, B. (2012). On the citation advantage of linking to data: Astrophysics. ⟨hprints-00714715v2⟩

Eisen, M. (10.10.2012). "Blinded by Big Science: The Lesson I Learned From ENCODE Is That Projects like ENCODE Are Not a Good Idea." It is NOT junk. https://www.michaeleisen.org/blog/?p=1179 (käytetty 15.9.2021)

Erdos, D. (2013a). Freedom of Expression Turned on its Head? Academic Social Research and Journalism in the European Union's Privacy Framework. *Pre-print of article published in Public Law*, 52-73. http://dx.doi.org/10.2139/ssrn.1928177

Erdos, D. (14.2.2013-b). "Mustn't Ask, Mustn't Tell: Could New EU Data Laws Ban Historical and Legal Research?" UK Constitutional Law Group https://ukconstitutionallaw.org/2013/02/14/david-erdos-mustnt-ask-mustnt-tell-could-new-eu-data-laws-ban-historical-and-legal-research/ (käytetty 15.9.2021)

European Commission. (2012). Commission recommendation on access to and preservation of scientific information. C(2012) 4890 final. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf (käytetty 16.9.2021)

European Commission. (2016a). Open Innovation, Open Science, Open to the World: A Vision for Europe. https://op.europa.eu/en/publication-detail/-/publication/3213b335-1cbc-11e6-ba9a-01aa75ed71a1 (käytetty 15.9.2021)

European Commission. (2016b). G20 Leaders' Communique Hangzhou Summit. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967 (käytetty 27.7.2021)

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1-10. https://doi.org/10.1002/meet.14504901068

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404-1416. https://doi.org/10.1002/asi.23480

Farquhar, A., & Brase, J. (2014). Data Identification and Citation—The Key to Unlocking the Promise of Data Sharing and Reuse. *D-Lib Magazine*, *20*(1/2). 10.1045/january2014-farquhar

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing?. *PloS one*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Forsman, M. (2010). Tutkijapalvelut haastavat kirjastoammattilaisia. *Signum*, (3).

Forsman, M., & Englund, J. (2014). Altmetriikka – bibliometriikan uusi suuntaus. *Signum*, (6).

Gamble, M., & Goble, C. (2011). Quality, trust, and utility of scientific data on the web: Towards a joint model. In *Proceedings of the 3rd international web science conference* (pp. 1-8). https://doi.org/10.1145/2527031.2527048

Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 1: Cyberinfrastructure Pri-mer for Librarians. *D-Lib magazine*, 13 (9/10).

Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib magazine*, 13(9/10).

Helsingin yliopisto. (n.d.) Helsingin yliopiston datatuki. https://helsinki.fi/fi/tutkimus/palvelut-tutkijoille/datatuki (käytetty 26.6.2021)

Henneken, E. A., & Accomazzi, A. (2011). Linking to data-effect on citation rates in astronomy. *arXiv preprint arXiv:1111.3618*. http://arxiv.org/abs/1111.3618v1

Hey, A. J., & Trefethen, A. E. 2003. The data deluge: An e-science perspective. <https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf> (käytetty 12.12.2020).

Hicks, D., Wouters, P., Waltman, L. et al. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature* 520, 429–431. https://doi.org/10.1038/520429a

High Level Expert Group on Scientific Data. (2010). Riding the wave, How Europe can gain from the rising tide of scientific data. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (käytetty 2.9.2021)

Hirtle, P. B. (2011). Introduction to intellectual property rights in data management. *Cornell University Research Data Management Service Group*.

Holdren, J. P. (2013). "Memorandum for the Heads of Executive Departments and Agencies." Executive Office of the President, Office of Science and Technology Policy. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (käytetty 4.9.2021)

Holopainen, M., & Pulkkinen, P. (2002). *Tilastolliset menetelmät* (2. p. 2003.). WSOY.

Jantz, R., & Giarlo, M. J. (2005). Digital preservation: Architecture and technology for trusted digital repositories. *D-Lib Magazine*. Volume 11, Number 6.

Kansa, E. (27.1.2014). It's the neoliberalism, stupid: Why instrumentalist arguments for Open Access, Open Data, and Open Science are not enough. http://blogs.lse.ac.uk/impactofsocialsciences/2014/01/27/its-the-neoliberalism-stupid-kansa/ (käytetty 5.10.2021)

FL, Korsmo. (2010). The origins and principles of the world data center system. *Data Science Journal*, *8*, IGY55-IGY65. https://doi.org/10.2481/dsj.SS_IGY-011

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., ... & Wächter, J. (2006). Data publication in the open access initiative. *Data Science Journal*, *5*, 79-83. https://doi.org/10.2481/dsj.5.79

Klump, J. (2012). Offener Zugang zu Forschungsdaten. In *Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft* (pp. 45–53). universaar. http://hdl.handle.net/10760/17213

Klump, J., & Huber, R. (2017). 20 Years of persistent identifiers–Which systems are here to stay?. *Data Science Journal*, *16*. https://doi.org/10.5334/dsj-2017-009

Kwa, C., & Rector, R. (2010). A data bias in interdisciplinary cooperation in the sciences: Ecology in climate change research. *Collaboration in the New Life Sciences. Farnham: Ashgate*, 161-176.

Late, E., & Kekäläinen, J. (2020). Use and users of a social science research data archive. *PloS one*, *15*(8), e0233455. https://doi.org/10.1371/journal.pone.0233455

Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., ... & Giles, C. L. (2001). Persistence of web references in scientific research. *Computer*, *34*(2), 26-31. https://doi.org/10.1109/2.901164

Lee, C. P., Dourish, P., & Mark, G. (2006, November). The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 483-492). https://doi.org/10.1145/1180875.1180950

Lide, D. R., & Gordon, H. (2012). CODATA @ 45 Years: 1966 to 2010. The Story of the ICSU Committee on Data for Science and Technology (CODATA) from 1966 to 2010. Paris: CODATA. https://www.codata.org/uploads/CODATA@45years.pdf (käytetty 19.5.2021)

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973-993. https://doi.org/10.1002/asi.23425

Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., ... & Varga, L. (2008). The provenance of electronic data. *Communications of the ACM*, *51*(4), 52-58. https://doi.org/10.1145/1330311.1330323

Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010). Panton Principles, Principles for open data in science. *Panton Principles*. https://pantonprinciples.org/ (käytetty 7.4.2021)

National Information Standards Organization. (2004). Understanding Metadata. Bethesda, MD.: NISO Press. https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf (käytetty 14.7.2021)

Nummenmaa, L., Holopainen, M., & Pulkkinen, P. (2014). *Tilastollisten menetelmien perusteet* (1. p.). Sanoma Pro.

Nuorteva, J. (2008). Tiedonhallintasuunnitelma tehostaa tutkimusdatan käyttöä. Tieteessä tapahtuu, 26(8), 36-40.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *Ucla L. Rev.*, *57*, 1701.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716

Organisation for Economic Co-operation and Development OECD. (2007). "OECD Principles and Guidelines for Access to Research Data from Public Funding." https://www.oecd.org/sti/inno/38500813.pdf (käytetty 18.6.2021)

Organisation for Economic Co-operation and Development OECD. (2015). “Making Open Science a Reality”, OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

Osswald, A., & Strathmann, S. (2012). The role of libraries in curation and preservation of research data in Germany: Findings of a survey. In *78th IFLA General Conference and Assembly*.

P2PU. (n.d.). P2PU | Datan avaaminen: Johdanto | Erilaiset datan muodot. https://courses.p2pu.org/en/courses/3271/content/7371/ (käytetty 20.10.2021)

Pampel, H., Bertelmann, R., & Hobohm, H.-C. (2010). *‘Data Librarianship’ – Rollen, Aufgaben, Kompetenzen* (RatSWD Working Paper No. 144). German Data Forum (RatSWD). Retrieved from German Data Forum (RatSWD) website: https://econpapers.repec.org/paper/rswrswwps/rswwps144.htm (4.11.2020)

Pampel, H., & Dallmeier-Tiessen, S. (2014). Open research data: From vision to practice. In *Opening science* (pp. 213–224). Springer, Cham.

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. Data Science Journal, 16, 8. https://doi.org/10.5334/dsj-2017-008

Piwowar, H.A., Day, R.S., & Fridsma, D.B. (2007). Sharing detailed research data is associated with increased citation rate J. Ioannidis, ed. *PLoS ONE*, *2*(3), 308. doi: 10.1371/journal.pone.0000308.

Regazzi, J. J. (2015). Scholarly communications: A history from content as king to content as kingmaker. Rowman & Littlefield

Reilly, S. (2012). The role of libraries in supporting data exchange. In *78th IFLA General Conference and Assembly*. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/116-reilly-en.pdf.

Roos, A., Kumpulainen, S., Järvelin, K., & Hedlund, T. (2008). The information environment of researchers in molecular medicine. *Information Research*, *13*(3), 13-3.

Rosenberg, D. (2013). “Data before the Fact.” In “Raw Data” is an Oxymoron, ed. Lisa Gitelman, 15-40. Cambridge, MA: MIT Press.

The Royal Society. (2012). Science as an open enterprise. The Royal Sociey Science Policy Centre report. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf. (käytetty 25.4.2021)

Ryan. (11.9.2020). What are libraries in programming | Coding Definition & Examples. https://www.idtech.com/blog/what-are-libraries-in-coding (käytetty 21.10.2021)

Sawyer, Steve. 2008. "Data Wealth, Data Poverty, Science and Cyberinfrastructure." Prometheus 26 (4):355– 371. https://doi.org/10.1080/08109020802459348

Sears, J.R., (2011). Data sharing effect on article citation rate in paleoceanography. IN53B1628. In *AGU Fall Meeting 2011*. http://static.core-apps.net/agu2011/html/IN53B-1628.html

Shotton, D. (2011). "Why Researchers Don't Publish Data." OpenCitations blog. https://opencitations.wordpress.com/2011/08/04/why-researchers-don%E2%80%99t-publish-data/ (käytetty 7.6.2021)

SND. (2021-a). SND – a national collaboration | Swedish National Data Service. https://snd.gu.se/en/about-us (käytetty 12.7.2021)

SND. (2021-b). The SND Consortium | Swedish National Data Service. SND.https://snd.gu.se/en/about-us/snd-consortium (käytetty 12.7.2021)

SND. (2021-c). The SND Network | Swedish National Data Service. https://snd.gu.se/en/about-us/snd-network (käytetty 12.7.2021)

SND. (2021-d). Describe and share data | Swedish National Data Service. https://snd.gu.se/en/describe-and-share-data (käytetty 12.7.2021)

SND. (2021-e). Search and find | Swedish National Data Service. https://snd.gu.se/en/catalogue/search (käytetty 12.7.2021)

SND. (2021-f). DataCite | Swedish National Data Service. https://snd.gu.se/en/about-us/international-collaborations/datacite (käytetty 12.7.2021)

SND. (2021-g). International Collaborations | Swedish National Data Service. https://snd.gu.se/en/about-us/international-collaborations (käytetty 12.7.2021)

SND. (2021-h). Share Data: Step by Step | Swedish National Data Service. https://snd.gu.se/en/describe-and-share-data (käytetty 12.7.2021)

SND. (2021-i). Accessibility levels at SND | Swedish National Data Service. https://snd.gu.se/en/find-data/research-data-catalogue/accessibility-levels-snd (käytetty 12.7.2021)

SND. (2021-j). FAQ | Swedish National Data Service. https://snd.gu.se/en/describe-and-deposit-data/faq (käytetty 12.7.2021)

SND. (2021-k). Manage Data | Swedish National Data Service. https://snd.gu.se/en/manage-data (käytetty 12.7.2021)

Suomen Akatemia. (n.d.). Aineistonhallintasuunnitelma | Suomen Akatemia. https://www.aka.fi/tutkimusrahoitus/hae-rahoitusta/nain-haet-rahoitusta/ohje-hakemisto/aineistonhallinta/aineistonhallintasuunnitelma/ (käytetty 20.10.2021)

Sustkova, H. P., Hettne, K. M., Wittenburg, P., Jacobsen, A., Kuhn, T., Pergl, R., ... & Schultes, E. (2020). FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. *Data Intelligence*, *2*(1-2), 158-170. https://doi.org/10.1162/dint_a_00038

Swedish Research Council. (2021). SND – Swedish National Data Service – Vetenskaps-rådet. https://www.vr.se/english/mandates/research-infrastructure/find-research-infrastructure/list/2018-10-18-snd---swedish-national-data-service.html (käytetty 11.7.2021)

Swedish Research Council. (2012). International evaluation of the Swedish National Data Service, SND – Vetenskapsrådet. https://www.vr.se/english/analysis/reports/our-reports/2012-06-11-international-evaluation-of-the-swedish-national-data-service-snd.html (käytetty 11.7.2021)

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570. https://doi.org/10.1142/S0218488502001648

Tampereen yliopiston kirjasto. (2021). Suunnittele – Tutkimusaineistojen hallinta – Oppaat. Tampereen yliopiston kirjasto. https://libguides.tuni.fi/tutkimusaineistojen-hallinta/suunnittele (käytetty 5.10.2021)

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research*, *5*, 632. https://doi.org/10.12688/f1000research.8460.3

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one*, *10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Tenopir, C., Hughes, D., Allard, S., Frame, M., Birch, B., Baird, L., ... & Lundeen, A. (2015). Research Data Services in Academic Libraries: Data Intensive Roles for the Future? Journal of eScience Librarianship 4(2): e1085. http://dx.doi.org/10.7191/jeslib.2015.1085

Tietoarkisto. (n.d.) Ohjeet – Aila. https://services.fsd.tuni.fi/help (käytetty 3.10.2021)

Turun yliopisto. (n.d.). Avoin tiede Turun yliopistossa – Tutkimusaineistot ja datapolitiikka. https://utu.fi/fi/tutkimus/avoin-tiede/aineistot (käytetty 28.6.2021)

Uhlir, P. F. (2012). *For Attribution--: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: National Academies Press. http://hdl.voced.edu.au/10707/232353

UNIFI. (2018). Avoin tiede ja data. Toimenpideohjelma suomalaiselle tiedeyhteisölle. Suomen yliopistojen rehtorineuvosto UNIFI ry. https://urn.fi/URN:NBN:fi-fe2018052424593

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS one*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Van der Graaf, M. and Waaijers, L. (2011). A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report.

Whyte, A., & Tedds, J. 2011. Making the Case for Research Data Management. DCC Briefing Papers. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm (käytetty 20.1.2021)

Wikipedia. (2021). COVID-19 pandemic – Wikipedia. https://en.wikipedia.org/wiki/COVID-19_pandemic (21.9.2021)

Wilkinson M. D. et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data. 3:160018 https://doi.org/10.1038/sdata.2016.18

Woutersen-Windhouwer, S. & Vernooy-Gerritsen, M. (2009). *Enhanced publications: Linking publications and research data in digital repositories*. Amsterdam University Press.

Yoon, A. (2014). "Making a square fit into a circle": Researchers' experiences reusing qualitative data. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1-4. https://doi.org/10.1002/meet.2014.14505101140

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1-6. https://doi.org/10.1002/pra2.2016.14505301126

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, *39*(3), 224-233. https://doi.org/10.1016/j.lisr.2017.07.008