TURUN YLIOPISTO

Datavuodot näkyviksi

Datavuototutkimus vuosina 2010–2022

Saimi Jukkara
Pro gradu –tutkielma
Median, musiikin ja taiteen tutkimuksen tutkinto-ohjelma, mediatutkimus
Historian, kulttuurin ja taiteiden tutkimuksen laitos
Humanistinen tiedekunta
Turun yliopisto
Marraskuu 2023

Pro gradu -tutkielma

Median, musiikin ja taiteen tutkimuksen tutkinto-ohjelma, mediatutkimus
Saimi Jukkara
Datavuodot näkyviksi: Datavuototutkimus vuosina 2010–2022
Sivumäärät: 80 sivua, liitteet 18 sivua

Nyky-yhteiskunta pyörii ihmisistä kerättävän datan ympärillä. Internetin sisältö kuratoidaan jokaiselle käyttäjälle heistä kerättyjen tietojen perusteella mahdollisimman kiinnostavaksi ja tiedot kerätään usein ilman käyttäjien tietoista suostumusta. Yksityisten ja julkisten tahojen palvelimien välillä kulkiessa henkilökohtainen data vuotaa myös kolmansille osapuolille. Tässä tutkielmassa perehdyn vuosina 2010–2022 tehtyyn datavuototutkimukseen systemaattisuuteen pyrkivän kartoittavan kirjallisuuskatsauksen avulla, kiinnittäen erityistä huomiota käytettyihin aineistoihin ja tutkimusmenetelmiin.

Datavuototutkimukseen liittyy monia erityispiirteitä aineistojen saavutettavuuden, tutkimusetiikan ja soveltuvien menetelmien osalta. Henkilökohtainen data, eli datavuotojen rakennusaine, on yksityisyydensuojan piirissä. Datayhtiöt eivät voi tai halua luovuttaa tietoja tutkimuskäyttöön. Aineistoja olisi saatavilla myös vapaasti verkossa tietomurtojen jäljiltä, mutta tällaiset aineistot ovat rikollista alkuperää ja siten tutkimuseettisesti kestämättömiä. Tutkimusten olisi oltava toistettavissa, mutta aineistot sisältävät yksityisiä tietoja.

Määrällisesti datavuototutkimus on keskittynyt kyselyihin, tapaustutkimuksiin ja datakaavintoihin, joista osa on eettisesti kestämättömällä pohjalla. Tutkimuksiin osallistuneet henkilöt ovat stereotyyppisesti nuoria korkeakouluopiskelijoita, ja tutkimuksen ulkopuolelle ovat jääneet haavoittuvassa asemassa olevat ihmisryhmät. Kuitenkin erilaisia aineistolähteitä yhdistelemällä, monitieteisillä tutkimusryhmillä ja ennakkoluulottomilla tutkimusotteilla voidaan päästä käsiksi myös uusiin aineistoihin, joilla pystytään kartoittamaan aiemmin tietotason ulkopuolelle jääneitä datavuotoja.

Avainsanat: mediatutkimus, data, datatalous, datafikaatio, dataistuminen, kirjallisuuskatsaus

Sisällysluettelo

1 Johdanto

Aamulla avaan puhelimen. Vilkaisen saapuneet sähköpostit ja uutiset. Selaan Instagramin tapahtumavirtaa, sen sekaan sijoitettuja mainoksia. Pelaan muutaman erän Bubble Witch 2 Saga -peliä[1], jossa kerään muutaman lisäpallon katsomalla lisää palveluntarjoajan esittämiä mainoksia. Viidessätoista minuutissa käyttämäni mobiilisovellukset ovat lähettäneet tietojani ympäri maapalloa. Tiedot siitä, mihin aikaan avasin sovelluksen, kuinka kauan käytin sitä, mitä tein, minkä mainoksen katsoin, minkä yli hyppäsin. Katsoinko vaatemainosta tänä aamuna sekunnin pidempään kuin eilen, googlasinko pelin lomassa tietoa lähialueen ravintoloista tai selasinko sosiaalisen median alustalla lyhytvideoita kissoista. Tiedot tallentuvat palvelimille niin Euroopassa kuin sen ulkopuolellakin. Kaiken tämän olen hyväksynyt käyttöehtosopimuksissa, joita en ole edes lukenut.

Internetin sivuja selatessamme meistä tallentuu jälki jokaisella hiirenklikkauksella. Jälki selaimen välimuistiin, jälki verkkosivun evästeisiin, jälki palveluntarjoajan tietokantoihin ja palvelimiin. Nämä jäljet kulkevat verkossa eri instituutioiden välillä, eli uppoavat dataa keräävien ja hyödyntävien tahojen verkostoon luoden virtoja eri organisaatioiden palveluiden välillä, tallentuen samalla lukemattomille eri palvelimille. Yksilötasolla jokaisesta verkon käyttäjästä vuotaa verkkoon henkilökohtaista dataa, jonka kulkua ei pystytä seuraamaan ja joka yhdistyy ihmismassojen datavirtaan. Nämä jäljet yksittäisten käyttäjien liikkeistä verkossa kulkevat erityyppisten yksityisten ja julkisten palveluiden välillä. Datavuotojen alkuperän, verkon infrastruktuurin ja sitä ohjaavan sääntelyn vuoksi niiden tutkimukseen liittyy uniikkeja eettisiä haasteita esimerkiksi aineistojen saatavuuden osalta. Lisäksi datavuototutkimukseen erityisiä piirteitä tuo esimerkiksi dataa keräävien tahojen haluttomuus osallistua tai edes mahdollistaa tutkimusta.

Tutkimusalana verkkovälitteisen datan tutkimus on nuori ja vakiintumaton. Tutkimuskentälle on tyypillistä, että tutkimuksilla saavutettu tieto ehtii vanheta jo ennen tutkimuksen julkaisuhetkeä. Tutkimuskenttää määrittää myös jatkuva muutos; datankäsittelyä koskeva lainsäädäntö, suuryritysten datan käyttöön liittyvät teknologiset ratkaisut ja yksilön

[1] King.com Ltd.-yrityksen julkaiseman pelin käyttäminen edellyttää tietosuojakäytännön hyväksymistä. Yrityksen mainontakumppanien listassa oli huhtikuussa 2022 yhteensä 105 yritystä, joista osa on datan jälleenmyymiseen keskittyviä. Pelaajista kerättävät tiedot sisältävät mm. sovelluksen käyttötiedot, pelaajan iän, maan tai alueen, sukupuolen ja ”muut tiedot, jotka yhtiö saa markkinointikumppaneilta tai muilta yhtiöiltä, joille olet antanut luvan, tai jotka ovat muuten oikeutettuja jakamaan näitä tietoja yhtiön kanssa”. (lähde: king.com/fi/privacyPolicy viitattu 2.5.2022).

mahdollisuudet vaikuttaa omiin datavirtoihinsa muuttuvat jatkuvasti. Datan kerääminen ja käyttö linkittyvät läheisesti yritystoimintaan ja markkinointiin, joten sitä koskevat käytänteet muuttuvat taukoamatta muodostaen kissa ja hiiri -leikin, jossa lakien ja asetusten säätäjät yrittävät keksiä keinoja hallita muutaman monikansallisen suuryrityksen toimintaa.

Julkisessa keskustelussa ihmisistä kerättävään dataa verrataan öljyyn, peltoihin tai muihin luonnossa esiintyviin raakamateriaaleihin, joiden kerääminen on yhteiskunnan toimimisen kannalta välttämätöntä. Lainsäädännöllisesti tietojen keräämistä, käsittelyä ja hyväksikäyttöä yritetään suitsia, mutta tulokset ovat tähän mennessä olleet ”hyväksyn evästeet”-ikkunoita verkkosivuilla, joita klikkaamme kiinni yhtä ponnekkaasti kuin ponnahdusikkunoita internetin alkuaikoina. Vuosituhannen vaihteen jälkeen mediatutkimuksessa on herätty dataistuneeseen yhteiskuntaan, jossa päätöksentekoamme ohjataan tiedoilla, joita eri tahot keräävät kuluttaessamme mediatuotteita. Datavuototutkimukseen liittyy kuitenkin monia uniikkeja haasteita, joita käsittelen tässä tutkielmassa.

1.1 Tutkimuskysymykset

- Millä menetelmin ja millaisilla aineistoilla datavuotoja on tutkittu humanistisyhteiskunnallisissa tieteissä? Mitä vahvuuksia ja heikkouksia eri metodeilla on?
- Miten autoetnografista tutkimusta voidaan hyödyntää datavirtoihin liittyvässä tutkimuksessa?
- Millaisia haasteita datavirtojen tutkimiseen liittyy ja miten niitä voitaisiin ratkaista?

Tässä tutkielmassa perehdyn yksityishenkilöistä verkkoon vuotaneen datan tutkimuksen aineistoihin ja menetelmiin vuosina 2010–2022, hyödyntäen systemaattisuuteen pyrkivää kartoittavaa kirjallisuuskatsausta. Kartoittavan kirjallisuuskatsaukseni aineisto koostuu 27 vertaisarvioidusta artikkelista, jotka keräsin neljää eri verkkohakemistoa hyödyntämällä ja käymällä 122 artikkelia läpi otsikko- ja abstraktitasolla[2]. Vastaavaa kirjallisuuskatsausta ei ole aiemmin ainakaan Suomessa tehty.

Aineistoni artikkeleissa perehdyin erityisesti niissä käytettyihin menetelmiin ja aineistoihin. Menetelmien osalta käyn läpi niissä esiinnousseita haasteita ja aineistoissa nostan esiin keräystapoja ja ihmisryhmiä, jotka ovat jääneet kirjallisuuskatsaukseni artikkelien ulkopuolelle. Aineistoni artikkeleissa toistui ongelma sopivien aineistojen saatavuudessa,

[2] Tarkemmin hakuprosessin etenemistä käyn läpi tämän tutkielman luvussa 3, erityisesti alaluvussa 3.1.

koska datavuotojen sisältö kuuluu yleisesti yksityisyydensuojan piiriin. Lisäksi henkilökohtaisen datan käyttöön liittyy erityisiä eettisiä haasteita. Tästä näkökulmasta pohdin tutkielmassani myös autoetnografian mahdollisuuksia ja haasteita tulevaisuuden datavuototutkimuksessa.

1.2 Tutkimuksen tavoitteet ja rakenne

Tässä tutkielmassa tarkastelen kirjallisuuskatsauksen avulla datavuototutkimusta ensin yleisellä tasolla, jonka jälkeen kartoitan aiemmassa tutkimuksessa käytettyjä menetelmiä ja aineistoja. Kirjallisuuskatsauksen tuloksia teemoittelemalla kartoitan aiemmassa tietotasossa olevia aukkoja ja sitä kautta uuden tutkimuksen tarpeellisuutta systemaattisesti. Tutkimukseni keskeisenä tavoitteena on ehdottaa uusia menetelmällisiä ja aineistoja koskevia ratkaisuja, joiden avulla datavuototutkimukseen voidaan tuottaa lisätietoa aikaisemmin syrjään jääneistä näkökulmista, aineistoista ja aineistojen lähteistä eli ihmisistä.

Ymmärtääksemme datavuotoja ja niiden akateemista tutkimusta tulee meidän ensin ymmärtää datavuotoihin liittyviä verkon toimintalogiikoita ja terminologiaa. Luvussa 2 taustoitan henkilökohtaisen datan tutkimusta käymällä läpi dataa ympäröivää infrastruktuuria ja määrittelemällä tutkielmani kannalta keskeiset termit. Termien määrittely on erityisen keskeistä nuoren tutkimusalan kohdalla, koska iso osa terminologiasta ei ole vakiintunutta ja osalle keskeistä sanastoa (esimerkiksi dataistuminen tai datafikaatio) ei ole vakiintuneita suomennoksia.

Kirjallisuuskatsaukseni prosessin kuvailen luvussa 3. Prosessin yksityiskohtainen kuvailu on erityisen tärkeää tutkimuksen toistettavuuden kannalta. Verkkohakemistot tuottavat eri tuloksia eri hakukerroilla ja hakijoilla, riippuen esimerkiksi hakuun käytetyn verkkoselaimen evästeistä. Lisäksi käsittelen alaluvussa 3.6 datayhtiö Alphabetin (Google) hallinnoiman Google Scholar -hakukoneen käyttämistä akateemisessa tutkimuksessa huomioiden erityisesti tutkimukseni aiheen. Eri verkkohakemistojen lisäksi hyödynsin kirjallisuuskatsauksessani lumipallo- ja helmihakutekniikoita. Lumipallotekniikassa tutustutaan muiden menetelmien avulla löydetyn artikkelin lähteisiin ja etsitään niistä tutkimuskysymyksiin sopivia artikkeleja. Vastaavasti helmitekniikan avulla pyritään paikantamaan esimerkiksi viittausmäärien perusteella avainartikkeleja, joiden avulla voidaan esimerkiksi määritellä tarkemmin tutkimusaiheen terminologiaa. Kirjallisuuskatsaukseeni päätyi näiden menetelmien avulla yhteensä 7 artikkelia, kun hakemistohakujen avulla artikkeleja valikoitui 20 (122 otsikko- ja abstraktitasolla läpikäydystä artikkelista).

Kirjallisuuskatsaukseni tulokset avaan luvussa 4. Aineistostani esiin nousseiden, datavuototutkimuksessa käytettyjen, aineistojen ja menetelmien lisäksi käyn erityisesti läpi aiemmassa tutkimuksessa ilmenneitä eettisiä haasteita. Näiden tulosten perusteella olen koostanut yhteen muutamia näkökulmia, teemoja, menetelmiä ja mahdollisia aineistoja, joita ei käsitelty kattavasti kirjallisuuskatsaukseni artikkeleissa (luku 5), pohtien erityisesti autoetnografian mahdollisuuksia ja haasteita datavuototutkimuksessa. Iso osa tutkimusten aineistoista perustuu tutkimuksen kohteiden haastatteluihin ja kyselyihin, jolloin haasteeksi on osoittautunut saada tutkimuskäyttöön dataa, joka ei perustu tutkittavien tahojen omiin subjektiivisiin näkemyksiin. Autoetnografia tarjoaa tähän haasteeseen näennäisesti yksinkertaisen ratkaisun, koska esimerkiksi GDPR:n avulla tutkijoiden on mahdollista saada käyttöönsä itseään koskevaa dataa, jota datayhtiöt keräävät kaikista ihmisistä luovuttamatta sitä tutkimuskäyttöön. Tämän lähestymistavan hyödyntäminen ei kuitenkaan ole ongelmatonta, kuten totean alaluvussa 5.3.

2 Datavuotojen taustat

Tässä luvussa käsittelen datavuotoihin läheisesti liittyvää terminologiaa humanistisyhteiskuntatieteellisestä näkökulmasta siinä määrin kuin se tämän tutkielman kannalta on välttämätöntä. Lopuksi käyn pintapuolisesti läpi datan kulkuun liittyvää infrastruktuuria, osapuolia ja sääntelyä. Koska tutkielmani tarkoitus on perehtyä nimenomaisesti humanistisyhteiskuntatieteelliseen tutkimukseen, en tässä luvussa käsittele datavuotoja teknologisesta näkökulmasta.

2.1 Henkilökohtainen data ja sen kerääminen

Etymologisesti latinan sana *data* tarkoittaa elementtejä, jotka voidaan irrottaa ilmiöstä. Tässä tutkielmassa käsittelen verkkodataa prosessissa muodostuvana yksikkönä. Prosessissa materiaali koostetaan ja irrotetaan uuteen muotoon, joka ei alkutilanteessa ollut nähtävissä. (Mejias & Couldry 2019, 2.) Data on symbolinen raakamateriaali, josta kootaan, lajitellaan ja tulkitaan tietoa[3], jota tarvitaan datayhtiöiden tuotantoprosesseihin (Couldry & Hepp 2016, 123; Zuboff 2019, 65). Käytännössä tämä tarkoittaa, että yksilöistä irrotetut tiedonmuruset yhdistetään algoritmeja hyödyntäen kokonaisuudeksi, jota voidaan käyttää esimerkiksi profilointiin, markkinointiin tai myydä eteenpäin kolmansille osapuolille.

Henkilökohtainen data sisältää informaatiota siitä mitä omistamme, teemme, keitä me olemme ja mitä ajattelemme, mihin uskomme, ja mitä tiedämme (Doss 2020, 11). Dataa kerätessä ihmislähtöiset elementit, eli kaikki puhelimeen kirjatuista kauppalistoista pikaviesteihin ja sosiaalisen median tykkäyksiin, päätyvät niiden keräämiseen ja prosessointiin tarkoitettuihin rakenteisiin (Couldry & Mejias 2019, xiii), joista valtaosa kuuluu nykyään yhteiskunnallisten toimielinten sijaan niin sanotuille datayhtiöille (esim. Meta ja Alphabet eli Google). Datan kerääminen on näin ollen sen analysoinnin ja hyödyntämisen lähtöpiste (Couldry & Mejias 2019, xiii). Yritykset keräävät dataa parantaakseen palvelujaan ja tarjotakseen lisää sopivaa sisältöä käyttäjilleen, mutta myös saadakseen lisätietoja käyttäjistään ja näyttääkseen heille personoitua mainontaa (Schufrin 2021, 1). Yksityiskohtainen tieto käyttäjistä tuottaa yrityksille merkittävää lisäarvoa, kun

[3] Suomen kielessä data ja tieto (engl. information) käsitetään usein synonyymeina. Esimerkiksi englanninkielinen termi database kääntyy suomeksi tietokannaksi ja data breach tietomurroksi. Tässä tutkielmassa lähestyn dataa ja tietoa eri yksikköinä: yksi data sisältää tietoa niin mitättömän määrän, että se on merkityksetön ilman, että siihen yhdistetään muita datayksiköitä. Esimerkiksi paikannustietojen osalta yksittäinen sijainti (data) on merkityksetön, ellei siihen muita tietoja (esim. mitä paikassa tapahtuu tai mikä laite on ollut kyseisellä paikalla ja milloin).

asiakasorganisaatioiden viestintää voidaan kohdistaa käyttäjiin tarkkojen muuttujien perusteella.

Datan kerääminen ympäristöä tarkkailemalla ja seuraamalla ei ole uusi ilmiö, eikä edes rajautunut vain ihmisiin, koska useat eri eläinlajit perustavat metsästystekniikkansa tarkkailulle (König et al. 2020, 1947). Ihmislajin keskuudessa dataa on tallennettu jo ainakin muinaisen Egyptin ja antiikin Kreikan aikoina (Mariani et al. 2021, 1; König et al. 2020, 1947). Yhteiskunnan monimutkaistuessa ja lopulta Internetin edesauttamana kerättävän ja tallennettavan datan määrä on kasvanut ja muuttanut muotoaan. Julkishallinnollisten (kirkko, verojen kerääminen, asuinpaikat ja tonttien rajat) instanssien sijaan tietoja keräävät nykyään yritykset, ja yhteisöllisesti merkittävien tietojen keräämisestä on laajennettu arkipäiväisten, vain henkilökohtaisesti merkittävien tietojen keräämiseen. Ilmiön ongelmien hahmottamista hankaloittaa, että datan keräämiseen, hallinnointiin ja tallentamiseen liittyvien infrastruktuurien ja teknologioiden toimintaa eivät hahmota kunnolla edes tutkijat, julkisesta keskustelusta puhumattakaan (Helles et al. 2020, 1958).

Datan keräämisessä, analysoinnissa ja hyödyntämisessä ei siis ole mitään uutta. Sen kaupallinen hyödyntäminen, usein ilman lähteen tietoista suostumusta, on kuitenkin herättänyt paljon niin yleistä keskustelua, kuin akuutin tarpeen dataan liittyvien ilmiöiden ja prosessien tutkimukselle. Yksinkertaistettuna, jos sinun tai minun huonon olon selvittämiseen tarvitaan verikokeita, luovutamme varmasti mielellämme verta (dataa), joka analysoidaan ja analyysia hyödynnetään terveydentilamme parantamiseen. Verikokeen tulokset jäävät terveydenhuollon järjestelmiin, joista niistä hyötyy pääasiassa potilas itse. Mahdollisesti voisimme antaa myös luvan tietojen käyttöön lääketieteellisissä tutkimuksessa. Mutta jos meiltä kysyttäisiin, voisiko kaiken meistä irrotettavan veren luovuttaa yritykselle, joka myisi sitä eteenpäin missä tahansa muodossa tuottaakseen voittoa omistajilleen, olisimme ehkä eri mieltä.

Nyky-yhteiskuntamme toiminta perustuu verkkovälitteiseen dataan ja erityisesti kansalaisten luottamukseen siitä, että dataa käsitellään asianmukaisesti. Yhteiskunta romahtaisi, jos emme voisi luottaa siihen, että tarkastaessani mobiilisovelluksesta pankkitilini saldon, summa pysyy samana ellen itse käytä omia rahojani. Datan keräämisen siirtyessä enenevissä määrin julkishallinnolta yrityksille, on samalla muuttunut myös ihmisten suhde dataan ja suhtautuminen sen keräämiseen, tallentamiseen ja käyttöön. Toisaalta huoli datan

keräämisestä ja käytöstä ei ole yleistettävissä koko yhteiskuntaan, sillä ilmiöstä eniten tietoiset henkilöt ovat siitä myös eniten huolissaan (Kennedy et al. 2021, 2).

2.2 Dataistuminen

Yksinkertaisimmillaan dataistumisella (engl. datafication, joissain yhteyksissä suomennettu myös datafikaatioksi) viitataan prosessiin, jossa jostain ilmiöstä tai informaatiosta tehdään luokiteltavaksi, mitattavaksi ja uudelleen esitettäväksi kelpaavaa dataa (Mayer-Schönberger & Cukier 2013, 78). Tämä prosessi jakautuu kahteen vaiheeseen: ensin tiedonmuruset muunnetaan dataksi, jonka jälkeen kerättyä dataa hyödynnetään, eli sen avulla luodaan lisäarvoa tuotteille tai palveluille (Mejias & Couldry 2019, 3). Historiallisessa kontekstissa dataistumisen voidaan katsoa olevan viimeisin vaihe ihmisten välisen kommunikaation kehityksessä: koneellistumisen, sähköistymisen ja digitalisoitumisen jälkeen olemme parhaillaan siirtymässä kohti seuraavaa vaihetta eli dataistumista (Couldry & Hepp 2016, 34). Ihmisten välisen kommunikaation kehityksen lisäksi dataistuminen voidaan nähdä myös yhteiskunnallisen jatkumon vaiheena, jossa maatalouden, teollistumisen ja tietoyhteiskunnan jälkeen siirrytään dataistuneeseen yhteiskuntaan.

Ilmiönä dataistumisen juuret ovat sosiaalisen kanssakäymisen sijaan liiketaloudessa, jossa on jo pitkään hallittu esimerkiksi tuotanto- ja logistiikkaketjuja kerätyn datan ja sen hallinnan avulla (Mejias & Couldry 2019, 4). Viime vuosina mediatutkimuksessa dataistumisen käsite on alkanut vakiintua tarkoittamaan ilmiötä, jossa nimenomaisesti ihmisiä koskevaa verkkovälitteistä informaatiota koostetaan dataksi, usein taloudellisen hyödyn saavuttamiseksi (Mejias & Couldry 2019, 1) ja mahdollisesti niin, etteivät Internetin käyttäjät ole siitä tietoisia (van Dijck et al. 2018, 33). Dataistuminen voidaankin määritellä sisältämään myös laajempi ihmiselämän muutos, jossa elämä muuntautuu loppumattomaksi datanlähteeksi (Mejias & Couldry 2019, 2) ja datan mitattavuudella perusteltuna yhteiskunnan toimintoihin luodaan illuusio luottamuksesta ja objektiivisuudesta (Couldry & Hepp 2016, 138).

Termin tarkoitusta voidaan laajentaa entisestään käsittämään myös ne metodit, joilla (ihmisiä koskevaan) dataan päästään käsiksi, ja joilla voidaan tarkkailla ihmisten käyttäytymistä (van Dijck 2014, 1478). Tällaisella määrittelyllä dataistumiseen liittyvien ilmiöiden tutkimus lähestyy sosiologiaa, kun määritelmä avautuu keinoihin mitata ja sitä kautta ymmärtää ihmisten sosiaalista käyttäytymistä. Lisäksi dataistumiseen voidaan sisällyttää myös kapitalismikritiikki toteamalla sen olevan prosessi, jossa ihmiskokemusta käytetään raakamateriaalia, jota jalostamalla vaikutetaan, ennustetaan ja ohjataan ihmisten toimintaa

(Zuboff 2019, 186–187) ja jonka avulla saavutettavat monetaariset hyödyt jäävät kansainvälisille suuryrityksille (Mejias & Couldry 2019, 6). Tässä mielessä dataistumisen voidaan katsoa olevan myös postkolonialismin ilmentymä (Mejias & Couldry 2019, 6), jossa kolonialistisia käytäntöjä toteutetaan maa-alueiden, luonnonvarojen ja orjatyön riistämisen, alistamisen ja valloittamisen sijaan ihmisten sosiaalisella pääomalla.

Historiallisesti verkon käyttäjistä kerättyyn dataan, joka sisälsi käyttäjien tietoisesti antaman informaation lisäksi esimerkiksi matkapuhelimista automaattisesti kerättyä metadataa (mm. aikaleimat, sijaintitiedot, verkkosivuilla ja sosiaalisen median alustoilla tehdyt toimet), suhtauduttiin verkkoalustojen sivutuotteena (van Dijck et al. 2018, 33). Nykyään teknologiayhtiöistä esimerkiksi Google on kuitenkin nimenomaisesti datayhtiö, ja keskeisin osa sen toimintaa on käyttäjädatan kerääminen, hyödyntäminen ja kierrättäminen mainostajien käyttöön. Datayhtiöiden tuottaessa pääosan ihmisiä koskevasta datasta, muita dataistumisen osapuolia ovat esimerkiksi valtiot ja kansainväliset yhteisöt, mutta myös vaikkapa terroristit ja hakkerit[4], jotka voivat kerätä ja hyödyntää dataa (Mejias & Couldry 2019, 2), sekä teknologiat, jotka mahdollistavat valtavien datamäärien keräämisen, hyödyntämisen ja säilyttämisen. Dataistuminen on punoutunut niin syvälle yhteiskuntaamme, että sen tutkimus on välttämätöntä.

Verkon käyttäjät, tai datan ”lähteet” eli ihmiset, kuitenkin myös hyötyvät dataistumisesta (van Dijck et al. 2018, 33). Palveluiden käyttäjät vastaanottavat heitä kiinnostavaa markkinointisisältöä (Aguirre et al. 2016, 98), löytävät tarvittavaa tietoa helpommin hakukoneoptimoinnin avulla (Bol et al. 2020, 1997), seuraavat ”kavereiden” liikkeitä sosiaalisen median alustoilla ja saavat tietoa lähialueen tapahtumista. Lisäksi älypuhelin muistaa salasanat ja muistuttaa tärkeistä tapahtumista. Keskeiset kysymykset dataistumisesta liittyvätkin ihmisten ja datan kerääjien suhteeseen: etiikkaan, käyttäjien tietoisuuteen ja suostumukseen, sekä kontrolloinnin mahdollisuuksiin heitä koskevan datan käyttämisessä. Lisäksi huomio tulee kiinnittää myös siihen, ketkä jäävät dataistumisen ulkopuolelle ja mitä se tarkoittaa. Dataistuminen voikin johtaa, tietoiseen tai tiedostamattomaan, syrjintään (Mejias & Couldry 2019, 3), kun etenkin haavoittuvassa asemassa olevat yksilöt voivat jäädä

[4] Hakkeri-sanalla on alun perin viitattu tietotekniikasta syvällisesti kiinnostuneisiin henkilöihin. Hakkerit voidaan jakaa valko- musta- ja harmaahattuhakkereihin riippuen hakkeroinnin tarkoitusperistä. Kaikki hakkerit eivät pyri vahingoittamaan kohteitaan, vaan esimerkiksi valkohattuhakkerit voivat harrastustoimintana pyrkiä löytämään tietoturva-aukkoja organisaatioiden verkkojärjestelmistä. (Haasio 2013, 99–100.) Tässä tutkielmassa käytän hakkeri-sanaa viittamaan henkilöön, joka lain vastaisesti hankkii pääsyn ihmisten dataan, riippumatta hakkerin tarkoitusperistä.

näkymättömäksi tai virheellisesti representoiduksi (König et al. 2020, 1952). Datayhtiöt ovat myös käyttäneet vähemmän sääntelyn (eli EU:n ulkopuolisia) maita ja alueita koeryhminä erilaisille datankeruuprojekteille, joissa yksityisyydensuojaa on loukattu (Arora 2019), jonka lisäksi kehittyvistä maista kerättävän datan analysointia tehdään usein länsimaalaisten tutkijoiden toimesta, joilla ei ole ymmärrystä alueellisesta kulttuurista (Taylor 2016).

Syrjintä voi kohdistua datan keräämisen ulkopuolelle jäävien ihmisryhmien lisäksi myös vaikkapa työn- tai lainanhakutilanteisiin, joissa tarkastellaan yksilöstä kerättyä ja yrityksille saatavilla olevaa dataa. Toisaalta yritykset voivat hyödyntää keräämäänsä dataa kohdentamalla markkinointia ja muuta viestintää erityisen haavoittuvassa asemassa oleviin henkilöihin (Bol et al., 2020, 1999) tai muuttamalla hinnoitteluaan verkon käyttäjästä kerätyn datan perusteella. Ihmisen suhde dataistumiseen voikin vaihdella (tuhoon tuomitusta) länsimaisen ihmisen pyrkimyksestä päästä sen ulottumattomiin, yritykseen tulla nähdyksi ja datan keräämisen kohteeksi.

Koska dataistuminen läpäisee kansallisesti ja kansainvälisesti yhteiskuntaa niin monien eri osa-alueiden ja sitä kautta tieteenalojen läpi, on sitä tutkimuksellisestikin välttämätöntä lähestyä monitieteisestä näkökulmasta. Kriittinen datatutkimus keskittyykin dataistumisen seurauksiin: valvontaan, yksilöä koskevien tietojen kaupallistamiseen, yksityisyydensuojaan ja yksilöiden kyvyttömyyteen hallita ja tiedostaa omaa seuraamistaan verkossa (Helles et al. 2020, 1960; 1971). Tutkimuksellisesti tulee huomio kiinnittää myös (piilotettuihin) valtarakenteisiin, eli siihen, kuka päättää miten ja mitä dataa voidaan kerätä ja miten sitä hyödynnetään

José van Dijck jatkaa ajatusta dataistumisesta erottamalla siitä vielä ideologian, dataismin, joka pohjautuu dataistumisen mahdollistamiin keinoihin objektiivisesti mitata ja seurata ihmisten käyttäytymistä mediateknologioiden avulla (2014, 198). Yuval Noah Harari vie dataismin käsitteen vielä pidemmälle, toteamalla Big Datan olevan ”uusi jumala” ja dataismin nostavan tieteen jumalalliseen asemaan sen voidessa luoda uuden lajin, eli tekoälyn (2017, 36). Jumalvertausta käyttää myös esimerkiksi Pentland kuvaillessaan henkilökohtaisen datan tuottamaa tutkimusaineistoa jumalan näkökulmaksi ihmistieteisiin (2012, 37).

Dataismin ydin on ihmisten luottamus ja usko siihen, että meistä kerättävää dataa käyttävät vain luotettavat tahot oikeudenmukaiseen toimintaan (van Dijck 2014, 198). Suurta kansainvälistä (media)huomiota saaneet datavuodot, kuten esimerkiksi Cambridge Analytica -tietovuoto (2018) ovat kuitenkin horjuttaneet tätä uskoa. Toisaalta yksilötasolla suuret

tietovuodot eivät ole saaneet verkon käyttäjiä huolestumaan, koska datavuotojen ei koeta koskettavan itseä. (Hinds et al. 2020.) Viime vuosina ainakin Suomessa datauskoa on horjuttanut suuret tietovuodot, esimerkiksi psykoterapiakeskus Vastaamon potilastietojen vuotaminen pimeään verkoon vuonna 2020, joten tulevina vuosina tulemme näkemään miten tiivistyvä keskustelu tietovuodoista vaikuttaa ihmisten käsitykseen omasta datastaan.

2.3 Datavuoto

Suomalaisessa tutkimuskirjallisuudessa datavuoto -termi on vielä vakiintumaton. Esimerkiksi Google Scholar -hakukoneen avulla hakusanoilla ”datavuoto” ja ”datavuodot” löytyi helmikuussa 2022 yhteensä vain 10 hakutulosta. Datavuodon (tai sen englanninkielisten vastineiden *data leak* tai *data leakage*) määritelmät vaihtelevat suuresti tieteenalan mukaan. Julkisessa keskustelussa datavuotoon viitataan usein tilanteissa, joissa verkkopalvelimille tallennetuista tiedoista tulee kenen tahansa saavutettavissa olevia. Tällaisista tapauksia ovat esimerkiksi WikiLeaks -sivusto ja useat ns. pimeän verkon[5] sivustot. Näitä ”vuotoja” ei kuitenkaan tapahtuisi, ellei tietoja olisi siirretty verkkosivustoilta ja -palvelimilta toisille.

Tässä tutkielmassa lähestyn datavuotoa yksilölähtöisten tietojen siirtymisenä kolmannelle tai kolmansille osapuolille ilman yksilön tietoista suostumusta. Tietoisella suostumuksella viitataan tilanteeseen, jossa yksilö on ymmärtänyt tietojensa käyttötarkoituksen ja sen, kelle kaikille ja millä ehdoin tietoja voidaan käyttää. Tietoiseen suostumukseen ei voida katsoa riittävän geneerisen käyttöehtosopimuksen, jonka luetuksi ilmoittaminen on tunnetusti internetin suurin valhe[6]. Määritelmä kattaa datayhtiöiden toiminnan, mutta ei sulje pois valtiollisen valvonnan harjoittamaa datankeruuta tai muita epäkaupallisia toimijoita. Tähän määritelmään ja sen erottamiseen tietoturva- ja tietosuojaloukkauksista sekä tietomurroista (*data breach*) palataan tämän tutkielman alaluvussa 4.1 jossa tarkastelen myös datavuotoilmiön historiallista muutosta.

[5] Pimeä verkko (engl. dark web) sisältää sivustoja, joita ei näy tavallisilla hakukoneilla ja niiden selaamiseen tarvitaan erityinen verkkoselain. Pimeän verkon käyttäminen ei itsessään ole laitonta, mutta se sisältää useita sivustoja, joissa esimerkiksi käydään kauppaa laittomilla palveluilla tai tavaroilla.

[6] Obar & Oeldorf-Hirsch (2020) käyttöehtosopimusten ja tietosuojakäytäntöjen lukemista mitanneessa tutkimuksessa vain kaksi prosenttia koeryhmästä huomasi sopimusten sisältävän mm. ehdon siitä, että kuvitteellisen sosiaalisen verkostoitumisen palvelun käytöstä maksuna perittäisiin ehtojen hyväksyjän ensimmäinen lapsi.

2.4 Dataa ympäröivä infrastruktuuri ja sääntely

Teknologian ja samalla datan hallinnan suuryrityksiin viitataan englanninkielisissä lähdeteksteissä ilmaisulla Big Tech tai Big 5. ”Isoon viisikkoon” kuuluvat Amazon, Apple, Microsoft, Alphabet (eli Google) ja Meta (jonka sivustoja ovat esimerkiksi Facebook, Instagram ja WhatsApp). Nämä yritykset vastaavat infrastruktuurista, joka pyörittää sosiaalista, poliittista ja taloudellista elämäämme (Birch & Bronson 2022, 1). Viime vuosina lainsäätäjien huomio ympäri maailmaa on kiinnittynyt näiden yritysten monopolia muistuttavaan asemaan ja sen mahdolliseen hyväksikäyttöön, sisältäen tutkimuksia esimerkiksi markkinoiden sääntelystä sekä ihmisten ja yritysten datan kohtuuttomasta keräämisestä ja käytöstä (Birch & Bronson 2022, 2) sekä psykografisesta markkinoinnista (Lindgren et al. 2019, 75).

Osa ison viisikon yrityksistä (Apple, Microsoft) on tietokoneiden tai muiden teknologialaitteiden valmistajia, jotka ovat siirtäneet liiketoimintamalliansa myös datan keräämiseen ja hyödyntämiseen. Nykyään erityisesti Metan ja Alphabetin eli Googlen liiketoimintamallit perustuvat niiden tarjoamien verkkoalustojen käyttäjien datan keräämiseen ja jälleenmyymiseen mainostajille (Couldry & Mejias 2016, 49). Markkinataloudessa nämä yritykset ovat arvoketjussa korkeimpana ja kontrolloivat muiden yritysten pääsyä ihmiseen (Lindgren et al. 2019, 75) valtavalla datamäärällään ja sen mahdollistamalla potentiaalilla kohdennettuun markkinointiin.

Sosiaalisen median alustojen lisäksi yrityksillä on hallinnassaan valtava määrä muita sivustoja ja teknologioita, joiden avulla dataa kerätään. Yritykset tarjoavat esimerkiksi maksupalveluita (mm. Google Pay) sekä sivustoille luotettavaan kirjautumiseen ja tunnistautumiseen liittyviä ratkaisuja. Esimerkiksi ihmisen bottiroboteista erottava CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) -teknologia on Alphabetin omistuksessa. Käytännössä monet erityisesti verkossa toimivat pienet ja keskisuuret yritykset luottavat ison viisikon palveluihin niin suuresti, että esimerkiksi Alphabetin palvelimen kaatuessa yritysten maksuliikenne ja verkkokauppa lakkaavat toimimasta kokonaan.

Dataa siirteleviä ja sitä kautta datavuotoja mahdollistavia yrityksiä ovat myös esineiden internet -yritykset (Internet of Things, IoT), jotka yhdistävät arkipäivän esineitä internetiin niiden liikkeiden ja toiminnan seuraamisen mahdollistamiseksi. Lisäksi henkilökohtaisen datan parissa toimii sen analysoimiseen keskittyviä yrityksiä eri toimialoilla ja tiedonvälittämiseen, eli datan keräämiseen, paketointiin ja jälleenmyyntiin keskittyviä (data

broker) yrityksiä (Couldry & Mejias 2016, 51–53). Eri tavoin ihmisten dataa käsittelevät ja välittävät yritykset tallentavat tiedot palvelimille, ja oman liiketoiminta-alansa muodostaa myös datakeskukset, joissa palvelimet fyysisesti sijaitsevat. Huomionarvoista on myös, että näiden datakeskusten sähkönkulutus vastaa jo nyt prosentista koko globaalista energiantarpeesta (Obringer et al. 2021, 1) ja merkittävä osa kerätystä datasta on yksilöiden henkilökohtaista dataa, jonka ainoa käyttötarkoitus on sen jälleenmyyminen tai käyttäminen markkinointitarkoituksiin.

Datayritysten toiminta perustuu algoritmeihin, joilla on keskeinen rooli henkilökohtaisen datan käsittelyssä. Algoritmit ovat useita vaiheita sisältäviä laskutapoja, joiden avulla voidaan ratkaista ihmisen määrittelemä ongelma ja muuntaa sisään tuleva raakadata haluttuun muotoon (Couldry & Hepp 2016, 133; Louridas & Pietiläinen 2021, 15). Algoritmit itsessään eivät tee päätöksiä tai ajattele, vaan ovat ohjelmoijien toteuttamia monivaiheisia laskutapoja. Kaikki algoritmit voisi toteuttaa myös ihmisen toimesta, vaikka kynällä ja paperilla, mutta tietokoneet nopeuttavat algoritmien läpi laskemista niin, että niiden tehokas käyttö on mahdollista. (Louridas & Pietiläinen 2021, 15–18; 23; 38.)

Arkipuheessa algoritmit mystifioidaan helposti eräänlaiseksi ihmisiä ohjaavaksi itsenäiseksi tekijäksi. Ihmisten suhtautumista algoritmeihin ohjaavat esimerkiksi datayhtiöiden käyttämien algoritmien toimintalogiikoiden salassapito ja huhut siitä, että Googlen omatkaan työntekijät eivät pysty hallitsemaan tai edes käsittämään algoritmien toimintaperiaatteita (D'Onfro 2018). Käytännössä algoritmi on aina ihmisen antamien ohjeiden tulos, mutta niistä voi toki muodostua äärimmäisen monimutkaisia kokonaisuuksia.

Datayritysten toimintaa pyritään säätelemään kansainvälisesti ja paikallisesti, mutta ongelmana on teknologian nopea kehitys ja yritysten toiminnan globaali skaalautuminen. Euroopassa yksilön datanhallinnan, ja sitä myötä datavuotojen akateemisen tutkimuksen, osalta keskeinen muutos tapahtui vuonna 2018, jolloin Euroopan Unionin yleinen tietosuoja-asetus eli GDPR (General Data Protection Regulation) tuli voimaan. Tietosuoja-asetus nostaa EU:n kansalaisten henkilökohtaisen datan hallinnan ihmisen perusoikeudeksi toteamalla, että "Luonnollisten henkilöiden suojelu henkilötietojen käsittelyn yhteydessä on perusoikeus […] jokaisella on oikeus henkilötietojensa suojaan" (Euroopan parlamentin ja neuvoston asetus 2016/679). Sama asetus sisältää myös EU:n kansalaisten oikeuden tulla unohdetuksi internetissä. Käytännössä GDPR esimerkiksi velvoittaa yritykset ja organisaatiot pyydettäessä

esittämään ihmisestä kerätyn datan luettavassa muodossa sekä henkilön niin pyytäessä poistamaan henkilöä koskevat julkiset tiedot internetistä.

Tietosuoja-asetus on lisännyt ihmisten tietoisuutta datan keräämiseen ja tallentamiseen liittyvissä kysymyksissä, mutta kerätyn datan saaminen, puhumattakaan sen hallinnasta, on edelleen haasteellista (Tsekoura & Panagopoulou 2020, 48–49). GDPR:n myötä ihmisten tietoisuus datan keräämisestä, sen arkipäiväistymisestä ja siihen liittyvistä riskeistä on kasvanut, mutta samalla on kasvanut myös ahdistus, pelko ja luottamuksen puute datan käyttämistä kohtaan (Kennedy et al. 2020, 2–3). Lisäksi asetuksen käytännön vaikutukset ovat jääneet vähäisiksi, jättäen datan käyttöön liittyvät markkinat varsin sääntelemättömiksi (König et al. 2020, 1946).

Markkinoiden jakautuminen valtaville kansainvälisille suuryrityksille rajoittaa kilpailua estäessään pienten yritysten pääsemistä markkinoille. Yhteiskunnallisilla rajoituksilla pyritään estämään ison viisikon monopoliaseman hyväksikäyttöä, mahdollistamaan vapaata kilpailua ja takaamaan markkinoiden toimivuus (Lindgren et al. 2019, 75). Toisaalta Euroopan Unionin poikkeuksellisen kireää suhtautumista dataan voidaan pohtia myös sisämarkkinoiden tukemisen kannalta. Ison viisikon yritykset ovat amerikkalaisia. Vastaavasti esimerkiksi Kiinassa ja Venäjällä, joissa yksilönvapauksia on muutenkin rajoitettu, ei ihmislähtöisen datan keräämistä ja käyttöä säädellä samalla tavoin. Toisaalta EU:n tiukka lainsäädäntö on saanut myös kritiikkiä, koska esimerkiksi julkisia tietoja yhdistämällä voi ikään kuin vahingossa muodostaa GDPR:n suojaaman henkilörekisterin (Doss 2020, 286) ja kireä sääntely aiheuttaa toimintaa hankaloittavaa byrokratiaa esimerkiksi voittoa tavoittelemattomille organisaatioille ja pienyrityksille.

Yhdysvalloissa datan keräämistä ja käyttämistä koskeva lainsäädäntö vaihtelee osavaltioittain, mutta sen pohjana on kuluttajasuojalainsäädäntö, jossa ongelmiksi tunnustetaan lähinnä luottokorttitietoja ja identiteettivarkauksia koskevat tilanteet ja yksityishenkilön ei katsota kärsineen vahinkoa datan käytöstä, mikäli hänelle ei ole koitunut siitä taloudellista vahinkoa (Doss 2020, 278). Yhdysvalloissa huomio on säännöllisesti kiinnittynyt valtion suorittamalle datan keräämiselle ja sen analysoinnille ilman kansalaisten suostumusta (Doss 2020, luku 14). Väljempi, yksityisen datan kauppatavaraan rinnastava ja sen myymisen mahdollistava lainsäädäntö (Lindgren et al. 2019, 85–87) mahdollistaa yritysten tehokkaan ja voitollisen toiminnan. Toisaalta merkittävä ero mantereiden lainsäädännöissä koskee yritysten tietomurtojen julkistamista, joka on Yhdysvalloissa pakollista, mutta Euroopassa ei (Lindgren

et al. 2019, 83). Koska eurooppalaiset uutiset eivät ole täynnä tietomurroista ja niiden aiheuttamista datavuodoista kertovaa materiaalia, voimme tuudittautua virheelliseen turvallisuudentunteeseen omaa dataamme koskien.

Euroopan unionin tietosuoja-asetus sisältää myös käytännön vaatimuksia verkkosivustojen ylläpitäjille. Sivustojen käyttäjien on esimerkiksi hyväksyttävä sivustolla mahdollisesti käytettävät evästeet (cookies) ja tietosuojaseloste (privacy policy). Evästeiden avulla verkkosivulla käymistä voidaan seurata ja verkkosivun palveluntarjoaja saa tietoa esimerkiksi sivustolla käytetystä ajasta, linkkien klikkaamisesta ja muista toimista. Evästeet tallentuvat palvelimelle, jolloin sivustolla uudelleen vieraillessa sivusto ikään kuin muistaa kävijän tiedot ja voi niiden avulla esimerkiksi muokata sivuston näkymää. (Voutilainen 2020, 180.) Käytännössä evästeiden avulla näemme esimerkiksi verkkokaupassa tuotteita, joita olemme aiemmin selanneet tai joita meille suositellaan aikaisemmin selaamiemme sivujen perusteella. Tietosuoja-asetuksen mukaisesti evästeiden käyttöön tulee saada käyttäjän sivustolle ja käyttötarkoitukseen yksilöity suostumus (Voutilainen 2020, 180), jonka myötä verkkosivuille mennessä käyttäjän on ensin klikattava ”hyväksyn” ilmoitukseen, jossa evästeistä kerrotaan käyttäjälle. Lisäksi verkkosivuston tietosuojaseloste, eli sivuston selaajan henkilökohtaisen datan käyttöön liittyvä informaatio, on oltava läpinäkyvästi ja ymmärrettävästi esillä sivustolla (Voutilainen 2020, 94). Käytännössä niin evästeiden kuin tietosuojaselostuksen käyttöön liittyy monenlaisia haasteita, alkaen siitä, että niitä ei todellisuudessa lue juuri kukaan eivätkä käyttäjät edes tiedä mihin ne liittyvät tai mitä ne tarkoittavat.

3 Kirjallisuuskatsaus

Tässä luvussa kuvaan aineistonkeruuprosessini, eli kirjallisuuskatsauksen etenemisen, ja perustelen käyttämäni menetelmät käsittelemieni tutkimusartikkelien keräämiseksi. Lopuksi pohdin myös Google Scholar -hakukoneen käyttöön liittyviä näkökulmia erityisesti datavuototutkimuksen kontekstissa, jossa isona osallisena on Googlen tyyppisten yritysten keräämä ja hyödyntämä data.

Aikaisempaan tutkimukseen perehtyminen on sisäsyntyinen osa akateemista tutkimusta ja se on välttämätöntä merkityksellisen tutkimuksen tuottamiselle (Silverman toim. 2020, 387; Rumrill et al. 2010, 399). Sen avulla tarkastellaan tutkimusaihetta yleisellä tasolla, paikannetaan aukkoja aiheen tietotasossa, ja sitä myötä kartoitetaan uuden tutkimuksen tarpeellisuutta (Silverman toim. 2020, 412). Aiempaan tutkimukseen perehtymisen ollessa osa kaikkea tutkimusta, kirjallisuuskatsaus on tutkimusmetodi, jonka avulla käydään tutkimusaihetta koskevaa aikaisempaa tutkimusta läpi järjestelmällisemmin. Riippumatta kirjallisuuskatsauksen laajuudesta ja tavoitteista, vähintään systemaattisuuteen pyrkivä lähestymistapa auttaa tutkijaa pysymään puolueettomampana ja saavuttamaan riittävän määrän aineistoa tutkittavasta ilmiöstä, jonka lisäksi järjestelmällinen lähestymistapa lisää tutkijan tehokkuutta (Booth et al. 2016, 2).

Kirjallisuuskatsaus ei kuitenkaan tyydy vain esittelemään aikaisempaa tutkimusta, vaan siinä analysoidaan, arvioidaan ja yhdistellään aikaisemman kirjallisuuden esiin tuomia teorioita ja näkemyksiä (Booth et al. 2016, 9). Tällöin kirjallisuuskatsauksen tuloksina saavutetaan kriittinen ymmärrys tutkimusaiheen tietotasosta, sekä paikannetaan tulevan tutkimuksen tarvetta ja mahdollisuuksia (Efron & Ravid 2019, 2). Kirjallisuuskatsauksen koostaminen vaatii tekijältään kattavaa tietoa tutkimusaihetta koskevasta olemassa olevasta tutkimuksesta, ja taitoja ja tekniikoita aihetta koskevan tiedon löytämiseen, kriittiseen analyysiin ja yhdistelemiseen (Efron & Ravid 2019, 16).

Kirjallisuuskatsauksen laajuus, materiaalin valintakriteerit ja käytetyt metodit valitaan tutkimuskysymysten perusteella (Booth et al. 2016, 13). Sopiva kirjallisuuskatsauksen tyyppi riippuu tutkimuskysymyksen lisäksi tieteenalasta ja tutkimuksen tarkoituksesta, joten sopivaa kirjallisuuskatsauksen tyyppiä valitessa tulee huomioida katsauksen fokus, tavoitteet, kirjoittajan näkökulma, kattavuus, sekä tutkimuksen organisaatio ja yleisö (Efron & Ravid 2019, 15; 24). Erityyppisiä kirjallisuuskatsauksien jaotteluja on lukemattomia, mutta

karkeasti ne voidaan jakaa kvantitatiivisiin (systemaattisiin) ja kvalitatiivisiin katsauksiin. Erilaisia kvalitatiivisia katsauksia ovat esimerkiksi nopea kirjallisuuskatsaus, perinteinen kirjallisuuskatsaus, kartoittava kirjallisuuskatsaus, narratiivinen katsaus, strukturoitu katsaus ja tutkimussynteesi (Arksey & O'Malley 2005, 20). Lisäksi kirjallisuuskatsauksien tyypeistä voidaan erottaa myös laadullista ja määrällistä tutkimusta yhdistävä monimenetelmällinen mixed methods -menetelmä (Efron & Ravid 2019, 16). Eri tutkijat käyttävät erilaisia jaotteluja ja yhdistelevät eri menetelmiä kirjallisuuskatsauksia tehdessään, joten muutamaa poikkeusta lukuun ottamatta katsauksille ei ole yleisesti käytössä olevia määritelmiä tai kriteerejä (Arksey & O'Malley 2005, 20).

Puhtaan systemaattisen katsauksen toteuttaminen tässä tutkielmassa olisi, tutkimuskysymykset, tutkielman laajuus ja käytettävät resurssit huomioon ottaen, ollut mahdotonta. Lisäksi tutkimuskysymykseni suuntautuvat spesifisti aiemman tutkimuksen menetelmiin ja aineistojen keruutapoihin varsinaisten tutkimustulosten sijaan. Näin ollen päädyin monimenetelmälliseen (engl. mixed methods) katsaukseen, jossa yhdistän laadullisen tutkimuksen kartoittavaan kirjallisuuskatsaukseen (engl. mapping review tai scoping review) systemaattisen kirjallisuuskatsauksen elementtejä soveltuvin osin. Pyrin esimerkiksi toteuttamaan toistettavuuden ja systemaattisuuden periaatteita (kts. esim. Silverman toim. 2020, 413) kartoitusprosessin läpinäkyvyyden ja toistettavuuden osalta kirjoittamalla kartoitusprosessin auki ja käyttämällä systemaattisen kirjallisuuskatsauksen työkalua artikkelien sisäänotto- ja poissulkukriteereistä. Lisäksi tietokantahaut tehtiin systemaattiselle katsaukselle tyypillisesti tarkasti määritellyillä hakusanoilla, toisin kuin kartoittavassa katsauksessa, jossa hakuprosessi voi olla polveilevampi (Arksey & O'Malley 2005, 22). Painopiste kirjallisuuskatsauksessani on kuitenkin kvantitatiivisen analyysin sijaan laadullisessa tutkimuksessa.

Kartoittavaan kirjallisuuskatsaukseen päädyin, koska sitä käyttämällä voidaan tarkastella laajoja aihealueita ja vastaavasti sen avulla ei pyritä arvostelemaan katsaukseen sisällytettyjen tutkimusten laatua. Kartoittava katsaus sopii esimerkiksi tutkimuskentän laajuuden ja löydösten kuvaamiseen, tutkittavan aiheen uusien tutkimusmahdollisuuksien etsimiseen, sekä täysin systemaattisen katsauksen tarpeen määrittelemiseen. (Arksey & O'Malley 2005, 20–22.) Siinä missä kartoittavan katsauksen avulla saavutetaan laajahko yleiskuva tutkittavasta ilmiöstä, systemaattinen kirjallisuuskatsaus pyrkii aiemman tutkimuksen kuvaamiseen syvällisesti ja puolueettomasti, luoden toistettavan ja laskettavissa olevan käsityksen tieteenalaa koskevasta tutkimuksesta kokonaisuudessaan (Silverman toim. 2020, 413; Efron

& Ravid 2019, 16). Tällöin toistettavasti toteutettu systemaattinen kirjallisuuskatsaus tarjoaa yleisesti kattavimman kuvan rajatun tieteenalan tutkimuksesta (Gusenbauer & Haddaway 2019, 182).

Tutkimusaiheeni lähestymistä kvalitatiivisen tutkimuksen kautta puoltaa myös tutkimusten sirpaloituneisuus ja nuori tutkimusala. Vakiintumattoman terminologian ja historiallisen tutkimusperinteen johdonmukaisuuden puuttumisen vuoksi puhtaasti kvantitatiivisen kirjallisuuskatsauksen jalkoihin voisi jäädä esimerkiksi tutkimusmetodeja, joita ei ole vielä laajasti käytetty, mutta joiden avulla pystyisin vastaamaan tutkimuskysymyksiini. Kvalitatiivisen tutkimuksen lähtökohtiin kuuluu tiedon subjektiivisuuden hyväksyminen ja se, että tutkimuksella ei pyritä selittämään sosiaalista maailmaa, vaan ymmärtämään sitä ”osallistujan” näkökulmasta (Efron & Ravid 2019, 17). Näin ollen laadullista lähestymistapaa tukee myös tutkimusaiheeni, joka on nimenomaisesti yksilöiden henkilökohtaisen datan tutkimus. Lisäksi iso osa kirjallisuuskatsaukseni tutkimusartikkeleissa käytetyistä metodeista oli niin ikään laadullisia tai monimenetelmällisiä (*mixed method*), kuten tämän tutkielman luvussa 4 käy ilmi.

Datavuototutkimukselle, kuten verkkoalustoja koskevalla tutkimukselle ylipäätään, on tyypillistä, että akateemisella tutkimuksella saavutettava tieto vanhenee nopeasti. Toisaalta julkaisuja tehdään määrällisesti paljon, ja monitieteisen tutkimusalueen informaatio pirstaloituu laajasti eri alojen julkaisuihin ja sitä kautta eri verkkohakemistoihin. Näin ollen pelkästään nopean kirjallisuuskatsauksen, tai edes muutamiin tietokantoihin tai julkaisuihin keskittyneen puhtaan kartoittavan kirjallisuuskatsauksen uhkana olisi ollut keskeisten artikkelien jääminen tutkimusaineiston ulkopuolelle. Lisäksi monitieteisiin ilmiöihin keskittyvän tutkimusalan, kuten esimerkiksi datavuototutkimuksen, ollessa kyseessä, ennakkoluulottomalla ja tieteenalojen rajoista välittämättömällä kirjallisuuskatsauksella saadaan esiin ilmiötä koskevan tutkimuksen poikkeavuudet ja yhteneväisyydet (Bearfield & Warren 2008, 64). Erityyppisten kirjallisuuskatsausten avulla voidaan myös arvioida, kehittää ja jopa rakentaa uutta teoriaa sekä paikantaa lisätutkimuksen tarpeita (Booth et al. 2016, 11).

Nuorten tutkimusalojen, joka datavuototutkimus on, tutkimuksen uhkana on myös tiedon kumuloitumisen ja tieteellisen diskurssin muodostumisen kärsiminen, jos tutkimusta ei pystytä kytkemään aikaisempiin tutkimuksiin (Gusenbauer & Haddaway 2019, 182). Lisäksi voidaan päätyä tekemään toistavaa tai päällekkäistä tutkimusta muiden hankkeiden tai tutkijoiden kanssa tai toistaa jo aikaisemmin tehtyjä metodologisia virheitä (Efron & Ravid

2019, 2). Tämän tutkielman luvussa 5 käynkin läpi aiemmassa datavuototutkimuksessa toistuneita haasteita ja ongelmia sekä sitä, millaisista lähtökohdista tutkimusta voitaisiin tulevaisuudessa lähteä tekemään.

3.1 Kirjallisuuskatsauksen prosessi

Kirjallisuuskatsauksen laajuuden ja syvyyden määrittelemisessä tulee ottaa huomioon tutkimukseen käytettävissä olevat ajalliset ja työryhmän osaamiseen liittyvät resurssit, sekä tutkimuksen tarkoitus ja yleisö (Booth et al. 2016, 36–40). Tässä alaluvussa kuvaan tutkielmani prosessin hakusanojen määrittelemisestä kirjallisuuskatsauksen keskeisimpien artikkelien rajaamiseen. Rajaamiseen vaikuttaneita tekijöitä edellä mainittujen lisäksi olivat pro gradu -tutkielman luonne ja ohjepituus, eli työryhmän rajoittuminen yhteen henkilöön ja tutkimuskysymyksen rajaaminen niin, että tutkielma ei ylitä annettua ohjepituutta. Systemaattisen kirjallisuuskatsauksen tekeminen olisi vaatinut tätä enemmän resursseja.

Artikkelien valintaprosessissa pyrin systemaattisuuteen ja toistettavuuteen siltä osin kuin se verkkohakuihin pohjautuvassa katsauksessa on mahdollista. Verkkohakemistojen toiminta perustuu algoritmeihin, joiden toimintalogiikat eivät ole julkisia, joten on täysin mahdollista, että samojen hakujen toistaminen eri verkon käyttäjän toimesta saa aikaan täysin eri hakutuloksia. Systemaattisuuteen pyrin erityisesti kuvaamalla hakuprosessin riittävällä tarkkuudella seuraavissa alaluvuissa ja käyttämällä auki kirjoitettua kriteeristöä katsaukseen valituille artikkeleille. Yleisesti hakuprosessini seuraa Arksey & O'Malley (2005, 22) viittä kartoittavan kirjallisuuskatsauksen vaihetta; tutkimuskysymysten määrittelemisen jälkeen identifioin olennaiset tutkimusartikkelit, tutustuin valittuihin artikkeleihin, lajittelin niistä tarpeellisen datan ja lopuksi vertailin ja raportoin katsaukseni tulokset.

3.2 Hakusanojen määritteleminen

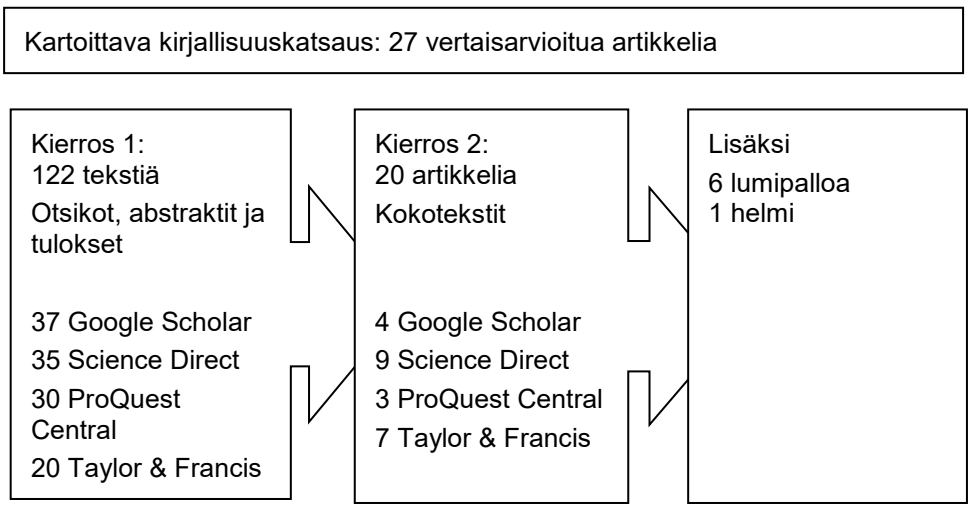
Verkkohauilla toteutetun kirjallisuuskatsauksen onnistumisen kannalta on keskeistä, että tutkimuskysymys, ja siitä johdetut verkkohaut, on määritelty riittävällä tarkkuudella (Gusenbauer & Haddaway 2019, 182). Useissa lähteissä (kts. esim. Efron & Ravid 2019; Gusenbauer & Haddaway 2019) kehotetaan konsultoimaan hakuehtojen laatimisessa organisaation asiantuntijoita, eli tässä tapauksessa yliopiston kirjaston henkilökuntaa, joka on erikoistunut hakuohjelmistojen käyttöön. Keskustelin tämän tutkielman hakumenetelmistä Turun yliopiston kirjaston informaatikon kanssa, jolta pyysin kommentteja ja kehitysehdotuksia suunnittelemiini hakusanoihin, käytettäviin tietokantoihin ja artikkelien

käsittelyyn. Informaatikon neuvojen avulla pystyin kohdentamaan fraasit tarkoituksenmukaisiksi, esimerkiksi suuntaamaan haut kaikissa hakemistoissa kokoteksteihin pelkkien otsikoiden sijaan, ja lisäksi kohdentamaan haut relevantteihin hakemistoihin (Hintikka 2022).

Yliopistoni informaatikon (Hintikka 2022) kanssa olimme samaa mieltä myös siitä, että hakusanat ”data” ja ”leak” ovat haastavia niiden yleisyyden takia. Esimerkiksi vuoden 2018 jälkeen julkaistuissa artikkeleissa, tutkimusalasta ja -aiheesta täysin riippumatta, data sana voi sisältyä pelkkänä viittauksena GDPR:n noudattamiseen. Lisäksi ”leak”, eli vuoto, on yleinen ilmaisu todella monilla tieteenaloilla niin metaforisessa kuin konkreettisessakin merkityksessä. Tästä syystä päädyin lisäämään hakufraaseihin sanan ”personal”, joka ohjasi haut jättämään pois edellä kuvailtujen kaltaiset artikkelit.

Monialaisille tutkimuskentille on tyypillistä, että terminologiassa esiintyy vaihtelevuutta ja saman ilmiön kuvaamiseen voidaan eri tieteenaloilla viitata eri termein (Kugley et al. 2016, 25). Tutkimukseni tekstihakujen haasteeksi osoittautuikin datavuoto-termin englanninkielisten vastineiden käyttö eri tarkoituksissa ja muodoissa: sanoja *data leak* ja *data leakage* käytettiin toistensa synonyymeina etenkin tietojärjestelmätieteiden tutkimuksissa, kun taas humanistisyhteiskunnallisissa artikkeleissa kyseisen termin lisäksi käytettiin esimerkiksi ilmaisua *data breach*. Kuten alaluvussa 2.3 toin ilmi, suomenkielisiä tutkimuksia ei datavuoto-sanalla ja sen eri muodoilla noussut esiin katsauksen kannalta merkittävää määrää, joten kaikki haut toteutettiin englanninkielisillä hakusanoilla.

3.3 Vaihe 1: Kartoittavat haut



Kuva 1 Kartoittavan kirjallisuuskatsauksen hakuprosessin eteneminen

Monitieteisen tutkimusaiheen kanssa ensimmäiset artikkelihaut kannattaa pitää laajempina, jotta tutkija saa yleiskäsityksen aihetta koskevasta tutkimuksesta (Booth et al. 2016, 114). Tämän takia tein ensimmäiset artikkelihaut Google Scholar -hakukoneella, pitäen hakusanat ja -kriteerit avoimempina, saavuttaakseni yleisnäkemyksen tutkimukseni aiheeseen liittyvästä aiemmasta tutkimuksesta. Google Scholariin kohdistuvasta kritiikistä (kts. alaluku 3.6 tässä tutkielmassa) huolimatta koin hakukoneen käytön perustelluksi, koska hakukone yhdistää eri julkaisijoiden artikkeleja ja on käyttöliittymältään joustava. Lisäksi yliopiston kirjaston omaan hakukoneeseen verrattuna Google Scholar -hakukoneen tulokset eivät riippuneet yliopiston ja julkaisijoiden välisistä käyttöoikeussopimuksista.

Ensimmäisissä rajaushauissa (haut 1–3) tarkoitukseni oli kartoittaa eri tietokannoista yhteen keräävän hakukoneen avulla aihettani koskevaa tutkimusta yleisellä tasolla. Samalla aloin hahmottaa keskeisiä julkaisijoita, joiden omiin tietokantoihin tein myöhemmin erillisiä hakuja. Vaikka tutkimuskysymykseni koskee nimenomaisesti humanistisyhteiskunnallisia tieteitä, en ensimmäisessä haussa edes pyrkinyt rajaamaan muitakaan tieteenaloja pois. Sen sijaan alusta saakka hyväksyin kartoitukseen vain vertaisarvioituja artikkeleja, rajaten niin sanotut white paper -julkaisut ja opinnäytetyöt pois käsittelystä. Lisäksi kartoittavat haut sisälsivät artikkeleja, jotka eivät auttaneet vastaamaan tutkimuskysymyksiini, mutta sisälsivät tutkielmani taustoitukseen tärkeää lähdemateriaalia. Toisin sanoen tällaiset artikkelit eivät päätyneet varsinaisen kirjallisuuskatsauksen viimeiseen vaiheeseen, vaan lähdeteksteiksi tutkielmani teoriaosaan.

Tutkimukseni toistettavuuden ja läpinäkyvyyden edesauttamiseksi luetteloin kaikki hakusanat ja -ajankohdat, sekä niiden kautta löydetyt artikkelit, jonka lisäksi tallensin artikkelit suoraan julkaisijan tietokannasta (kts. esim. Kugley et al. 2016, 9). Lisäksi annoin jokaiselle artikkelille #0.00-muotoisen tunnisteen, jossa ensimmäinen numero kuvasi mistä hakukierroksesta artikkeli oli peräisin ja jälkimmäinen numero monesko artikkeli kyseisestä hausta oli kyseessä. Lisäksi lumipallo- ja helmitekniikoilla kerätyt artikkelit saivat numeron alkuun tunnisteen ”SB” tai ”H” (esim. #SB1.2; #H.1). Näin jokaisella artikkelilla oli alusta saakka uniikki tunniste, joka prosessin edetessä helpotti omaa työskentelyäni huomattavasti. Käytin tunnisteita myös tallentaessani artikkelit tietokoneelleni ja artikkelien toisistaan erottamisen lisäksi tunnisteista oli heti nähtävissä missä vaiheessa hakuprosessia artikkeli oli löytynyt, millä keinoin (tietokantahaku vai lumipallo tai helmi) ja lisäksi monesko artikkeli kyseiseltä hakukerralta oli kyseessä. Haasteita aiheuttivat vain artikkelit, jotka tulivat esille

useammassa lähteessä, eli niin sanotut tuplat, mutta niitä oli koko hakuprosessin aikana kolme kappaletta.

Alusta saakka artikkelien luetteloinnin työkaluna käytin Microsoft Excel -taulukkoa (kuva 2), johon koostin tiedot esimerkiksi artikkelin kirjoittajista, julkaisijasta, julkaisuvuodesta, artikkelissa ilmoitetuista avainsanoista, tutkimuskysymyksestä ja tärkeimmistä löydöksistä. Hakuprosessin ensimmäisessä vaiheessa luin kaikista artikkeleista otsikon lisäksi vähintään abstraktin ja lopputulokset, jonka jälkeen seuraavilla kierroksilla perehdyin myös kokoteksteihin. Kahden haun jälkeen lisäsin taulukkoon myös erikseen sarakkeen sille, onko etiikka mainittu tutkimusartikkelissa vai ei. Lisäksi värikoodasin artikkelit, eli käytin niiden relevanttiuden mittaamiseen liikennevaloasteikkoa, jossa tumman vihreällä merkityt olivat tutkimusaiheeni kannalta relevantteja ja vastaavasti spektrin toisessa päässä tumman keltaisella merkityt artikkelit eivät vaikuttaneet keskeisiltä.

	Tutkija	Jul	Kieli	Otsikko	Julkaisija	Sivumää	Taustaorganisaatio	Tieteenala	Kvali	Metod	Aineisto	Keräystap	Tutkimuskysymy	Keskeiset tulokse	Avainsanat
#1.13	Swart, I.P.;	2013	Englanti	Visualization	Institute of El	8	University of Rhodes, Uni	Tietojärjestelm	Mixed	Teemoitte	Henkilökoht	Vapaasti verko	Mitä dataa ja missä m	Yksilön tunnetasolla a	data leakage; Ir
#2.6	Maris, Elen	2020	Englanti	Tracking sex:	new media &	21	Microsoft Research; Carn	Mediatutkimus	Mixed	Data scrap	22 484 porno	Ei varsinaisest	Mitä ja miten paljon k	Pornosivustoilla yhdist	Consent, data, p
#2.8	Valentinus	2018	Englanti	Impact of use	Institute of El	6	University of Indonesia (	Tietojärjestelm	Mixed	Kysely (e-m	340 sosiaali	Verkkokysely 2	Miten tietoisuus sosia	Tietoisuus yksityisyyte	Facebook; Twitt
#5.5	Gan, Diane	2015	Englanti	Social Networ	Future Interne	28	University of Greenwich (	Tietojärjestelm	mixed	Empiiriner	Large scale s	Vapaasti saata	Mitä kaikkea tietoa lä	Pystyttiin seuraamaan	Tweets;
#5.7	Schufrin, M	2021	Englanti	Visualizatio	IEEE Transacti	10	Fraunhofer IGD (Saksa)	Tietojärjestelm	mixed	Empiiriner	GDPR:n peru	Prototypin koe	Miten GDPR:n avulla s	Datan visualisointi aut	Data visualizati
#5.8	AlSabah, M	2018	Englanti	Your culture i	Computers & S	15	Qatar Computing Researc	Tietojärjestelm	Kvali	Teemoitte	400 000 pank	Data vuodettu	Miten kulttuuri ja kieli	Salasanan valintaan va	Usable security;
#6.3	Hayley Spa	2016	Englanti	Reflecting on	Geoforum	7	The University of Auckla	Maantiede	Kvali	Tapaustutk	PhD tutkimu	n/a	Millaisia eettisiä riski	Datavuodosta huolima	Digital technolo

Kuva 2 Kuvankaappaus viimeisen vaiheen artikkelien luokittelusta

Haku #1 13.1.2022, Google Scholar https://scholar.google.com/

> personal data leak
> anywhere in the article
> Return articles dated between 2010–2022
> Sort by relevance

Käytyäni hakutulosten ensimmäiset 10 artikkelia läpi toistin täsmälleen saman haun (haku #2) 17.1.2022. Hakutuloksien kymmenen ensimmäisen artikkelin joukkoon nousi viisi uutta artikkelia, jotka lisäsin artikkelilistaani. Suurin osa näistä tutkimuksista käsitteli datavuotoja tietojenkäsittelytieteiden näkökulmasta, joten iso osa tutkimuksista oli irrelevanttia omalle tutkimuskysymykselleni. Tutkimuksen toistettavuuden ja Google Scholarin algoritmien toimintalogiikan näkökulmasta on kuitenkin huomionarvoista, että ”merkittävimpien” artikkelien sisältö vaihtui puolella kahden eri haun välillä, vaikka hakujen välissä oli vain neljä päivää ja haut tehtiin samalla tietokoneella, selaimella (Google Chrome) ja IP-osoitteella, sekä tyhjentämättä selaimen evästeitä.

Seuraavaksi tarkensin hakua kohdistumaan vain tutkimuskysymykseni kannalta relevantteihin tieteenaloihin:

Haku #3 20.1.2022, Google Scholar https://scholar.google.com/

> personal data leak ("media studies" OR humanities OR "social sciences")
> anywhere in the article
> Return articles dated between 2010–2022
> Sort by relevance

Näistä hakutuloksista valitsin 20 ensimmäistä artikkelia tarkasteltavaksi. Google Scholar -hakukonetta käyttämällä keräsin yhteensä 35 artikkelia.

Google Scholar -hakujen tulokset auttoivat hahmottamaan, millä tavoin hakukriteerejä tulisi vielä tarkentaa päästäkseni mahdollisimman relevantteihin tuloksiin omien tutkimuskysymysteni kannalta.

Haku #4 30.1.2022, ProQuest Central

> ("data leak" OR "data leaks" OR "data leakage" OR "data leakages" OR "data breach" OR "data breaches") AND personal
> Scholarly journals
> Publication date 2010–2022

Haulla löytyi 280 tutkimusartikkelia. Käytyäni ensimmäiset kymmenen tulosta läpi otsikko- ja abstraktitasolla, ja seuraavat kymmenen otsikoiden tasolla, jouduin toteamaan, että ”breach”-sanan käyttö laajentaa hakua liiaksi käsittelemään tietomurtoja. Näin ollen hakufraasi ei johtanut artikkeleihin, jotka käsittelisivät riittävän tarkasti tutkimaani ilmiötä, eli datavuotoja.

Haku #5 9.2.2022, ProQuest Central

> ("data leak" OR "data leaks" OR "data leakage" OR "data leakages") AND personal
> Scholarly journals, peer reviewed
> Publication date 2010–2022
> Sorted by relevance

Haulla löytyi 54 tutkimusartikkelia. Alkupään hakutuloksissa oli merkittävä määrä tietojärjestelmätieteiden ja tietojenkäsittelytieteiden artikkeleja. Yritin tämän takia rajata hakua vielä tarkemmaksi hakuehtojen ”humanities” ja ”social” (jolloin hakuun olisi rajautuneet niin ”social sciences” kuin ”social media”), mutta nämä fraasit rajasivat hakutulokset vain muutaman artikkelin mittaiseksi. Ensimmäiset 20 hakutulosta kävin läpi

otsikon, abstraktin ja lopputulosten asteella. Näistä kolme osoittautui omien tutkimuskysymysteni kannalta keskeisiksi. Hakutulokset 21–54 kävin läpi otsikon ja abstraktin tasolla, mutta artikkelit eivät olleet relevantteja. Näistä artikkeleista yksi päätyi vaiheeseen 2. Hakutulosten perusteella totesin hakufraasin sinänsä olevan sopiva tarkoitukseensa.

Haku #6 15.2.2022, Science Direct

> "data leak" OR "data leaks" OR "data leakage" OR "data leakages" AND personal
> Research articles
> Publication date 2010–2022
> Social sciences
> Sorted by relevance

Haulla löytyi 213 tulosta. ScienceDirectin hakemiston hyvä puoli oli se, että hausta sai rajattua heti pois ”Computer Science” tieteenalan (joka käsittää sekä tietojärjestelmätieteiden, että tietojenkäsittelytieteiden artikkelit), joten tulosten rajaaminen käsittelemään tutkimuskysymykseni mukaisesti humanistisyhteiskunnallisia tieteitä oli mahdollista.

Haku #7 16.2.2022, Taylor & Francis

> ("data leak" OR "data leaks" OR "data leakage" OR "data leakages") AND personal
> Article
> Publication date 2010–2022
> “Humanities” ja “Legal, Ethical & Social Aspects of IT”
> Order by relevance

Humanities -kategorian haulla löytyi 30 tulosta. Näistä ensimmäisellä kierroksella kävin läpi ensimmäiset 20 tulosta, joista valitsin 7 seuraavalle kierrokselle.

3.4 Vaihe 2: Lumipallo ja helmet

Lumipallotekniikalla viitataan kirjallisuuskatsauksen kontekstissa menetelmään, jossa tutkittavan aiheen kannalta olennaisia artikkeleja etsitään etenemällä yhden artikkelin lähteistä eteenpäin, tai haetaan samaa lähdettä käyttäneitä muita artikkeleja (Booth et al. 2016, 121; 315). Lumipallotekniikkaa voidaan käyttää myös jäljittämällä artikkeleja, joissa on käytetty lähteenä keskeiseksi koettua artikkelia. Tässä tutkielmassa keskeisimmät artikkelit olivat aivan viime vuosilta, joten niitä ei ole vielä ehditty käyttää lähteenä uudemmissa julkaisuissa. Vastaavasti helmi-hakutekniikassa (joissain yhteyksissä myös helmenkasvatustekniikka) kartoitettavasta aiheesta paikannetaan yksi tai useampi avainartikkeli, jonka avulla pyritään

lähestymään tutkittavaa aihetta ja määrittelemään tarkemmin esimerkiksi terminologiaa (Booth et al. 2016, 114).

Tässä tutkielmassa hyödynsin varsinaisten tietokanta- ja hakukonehakujen jälkeen lumipallotekniikkaa. Tietokantahauista (haku #2) nousi esiin yksi artikkeli, joka oli julkaistu New Media & Society -lehden erikoisnumerossa ”The Tracked Society: Interdisciplinary Approaches on Online Tracking”. Kyseisen erikoisnumeron seitsemästä artikkelista viisi valikoitui mukaan kirjallisuuskatsaukseni viimeiseen vaiheeseen. Lisäksi lumipallotekniikkaa käyttämällä löytyi myös ”International Journal of Human-Computer Studies” -julkaisu. Julkaisun artikkelien sisällöstä tehdyllä ”data leak” haulla löytyi 24 artikkelia, joista kaksi artikkelia valikoitui lopulliseen artikkelikatsaukseen.

Lumipallotekniikalla löydettävissä olevia ja kirjallisuuskatsaukseni kannalta relevantteja artikkeleja jäi varmasti katsauksen ulkopuolelle, mutta hakuprosessia ei voinut jatkaa loputtomiin. Artikkelien jääminen alkuperäisten hakukriteerien ulkopuolelle esimerkiksi terminologian osalta kertoo käsitteiden moninaisuudesta ja mahdollisesti myös hakutietokantojen algoritmien toiminnasta. Tietomurtoja (data breach) käsittelevien artikkelien lisäksi datan keräämistä koskevat artikkelit jäivät tarkoituksella katsauksen ulkopuolelle, koska tutkimuskysymykseni käsittelevät datavuotoja koskevaa tutkimusta.

Helmiartikkeleiksi valikoidaan useimmiten niin sanottuja klassikkotekstejä. Tutkimusaiheeni uutuuden, nopean kehittymisen, monialaisuuden ja tutkimuskysymysteni asettelun huomioiden en paikantanut klassikkoasemaan nousseita julkaisuja. Lumipallotekniikan lisäksi hyödynsin helmitekniikkaa paikantamalla toiseen vaiheeseen päässeistä artikkeleista tekstit, joiden metodiksi olin määritellyt kirjallisuuskatsauksen (yhteensä kaksi artikkelia: #SB1.1 König et al. 2020; #SB1.6 Breuer et al. 2020). Kävin näiden artikkelien tutkimusaineistot (eli kirjallisuuden, joka oli otettu mukaan kirjallisuuskatsaukseen) läpi otsikko- ja abstraktitasolla. Lisäksi aineistossa oli yksi vuonna 2015 julkaistu kirjallisuuskatsausartikkeli, joka jäi oman tutkimukseni ulkopuolelle, koska siinä läpikäydyt artikkelit olivat lähinnä historiallisia huomioon ottaen ilmiön nopean muutoksen. Kaksi muuta julkaisua tarjosivat hyvää taustamateriaalia esimerkiksi datatutkimuksen etiikkaan ja tutkimustulosten säilytykseen liittyen. Tässä vaiheessa kartoitusta oli kuitenkin jo selvää, että kartoitus koski terminologisesti ”data leak” ja ”data leakage” tutkimusta. Sellaisten artikkelien, jotka eivät sisältäneet kyseisiä termejä, lisääminen artikkeleihin tässä vaiheessa olisi voinut vaikuttaa

kirjallisuuskatsauksen tuloksiin merkittävästi. Yksi artikkeli läpäisi ”vuotoseulan” ja se lisättiin kartoituksen vaiheeseen kaksi.

3.5 Vaihe 3: Sisäänotto- ja poissulkukriteerit ja kokotekstit

Sisäänottokriteerit	Poissulkukriteerit
Tutkimus käsittelee datatutkimusta ihmistieteille merkityksellisellä tavalla	Tutkimus on luonteeltaan tietotekninen
Julkaisun kieli on englanti	Julkaisun kieli on joku muu kuin englanti
Kyseessä on artikkeli, joka on julkaistu vertaisarvioidussa julkaisussa	Kyseessä on kirja-arvostelu, kirjan luku, white paper julkaisu tai muu kuin artikkeli
Artikkeli on saatavilla Turun yliopiston tietokantojen kautta tai julkisesti verkosta	Artikkeli ei ole saatavilla yliopiston tietokantojen kautta tai julkisesti verkosta

Vaiheista 1 ja 2 valikoituneet 28 artikkelia luin läpi kokonaisuudessaan. Koska edelliset hakuvaiheet olivat kestäneet noin kuukauden, epäilin artikkelien hyväksymiskriteerien hieman tarkentuneen kartoituksen edetessä, joten tarkastin artikkelit lukemisen yhteydessä vielä suhteessa tutkimuskysymykseeni. Tässä yhteydessä päädyin rajaamaan artikkeleista ulos vielä sellaiset artikkelit, joissa keskityttiin sellaisten teknologisten ratkaisujen kehittämiseen, jotka eivät olleet suoraan tekemisissä käyttöliittymän kanssa, tai jotka keskittyivät pohdintaan tutkimuksen tärkeydestä. Lisäksi tähän vaiheeseen saakka oli jostain syystä päätynyt yksi konferenssipapereiden esittelyartikkeli, joka niin ikään poistettiin. Tässä vaiheessa kokotekstejä lukiessani lisäsin artikkeleihin teemoittelun varsinaisen analyysini pohjaksi. Nämä artikkelit vastasivat tutkimuskysymyksiini 1 ja 3, jonka lisäksi vaiheen 1 artikkelit ja koko kirjallisuuskatsauksen prosessi auttoivat vastaamaan erityisesti tutkimuskysymykseeni numero 3, joka koskee datavirtojen tutkimiseen liittyviä haasteita.

3.6 Google Scholar akateemisen tutkimuksen työkaluna

Julkaisujen määrän noustessa ja painottuessa yhä enemmän verkkojulkaisuihin myös akateeminen tutkimus on enenevissä määrin alisteista dataistumiselle (kts. tämän tutkielman alaluku 2.2). Verkkohakemistot käyttävät toiminnassaan algoritmeja, joiden ohjelmoinnin painotusten takia samat hakunimikkeet ja -ehdot eri tietokoneilla voivat tuottaa eri otoksen esimerkiksi verkkoselaimeen tallennettujen evästeiden takia. Sama haku voi tuottaa eri tuloksia riippuen myös haun ajankohdasta, kuten aiemmin tässä luvussa totesin käyneen tämänkin tutkielman kirjallisuuskatsauksessa Google Scholar -hakukonetta käytettäessä.

Ylipäätään akateemisten artikkelien verkkohakemistojen ja hakukoneiden systemaattista tutkimusta on tehty verrattain vähän (Gusenbauer & Haddaway 2019, 181–182), mutta algoritmien toimintaan ja evästeiden vaikutukseen perehtyminen akateemisen tiedon hakuprosessissa ylittää tämän tutkielman resurssit ja tutkimuskysymyksen.

Koska tutkielmani käsittelee datavuotoja koskevaa tutkimusta, koin ongelmalliseksi keskittää tutkimukseni Big Tech -datayhtiö Alphabetin hallinnassa olevaan Google Scholar -hakemistoon sen suosiosta huolimatta. Google Scholar -hakemisto on noussut julkaisijoiden hakemistojen rinnalle sen kattavuuden takia, mutta sen käyttöön liittyy monia ongelmia. Google ei esimerkiksi luovuta tietoja siitä, miten artikkelit sen tietokantoihin kerätään, mitä julkaisijoita jää datakaavinnan ulkopuolelle, miten usein tietokantaa päivitetään tai mitkä tekijät vaikuttavat tulosten luokitteluun tärkeiksi (sort by relevance) (Baneyx 2008, 369). Lisäksi hakukoneoptimointi, eli verkkosivujen sisällön muotoilu niin, että sivusto näkyisi mahdollisimman korkealla Googlen hakutuloksissa on vuotanut myös akateemiseen kirjoittamiseen ja Google Scholar -hakemiston merkityksellisyys (relevance) algoritmi vaikuttaisi nostavan englanninkieliset artikkelit muiden kielien yläpuolelle hakutuloksissa (Rovira et al. 2021, 2; 13–14).

Tässä tutkielmassa päädyin kuitenkin käyttämään artikkelien keräämisessä ensin Google Scholar -hakukonetta, jonka jälkeen siirryin tekemään rajatumpia hakuja ScienceDirect, ProQuest Central ja Taylor & Francisin tietokantoihin. Tutkielmassani Google Scholar toimi ikään kuin kristallipallona, jonka avulla pystyin karkeasti arvioimaan tutkimani ilmiön eri tieteenalojen välisiä painopisteitä ja toisaalta hakusanojeni relevanttiutta.

Poissulkukriteerinä kirjallisuuskatsaukseni hauissa oli muu kuin englanninkielisyys. Kirjallisuuskatsauksen hakuprosessissa mukana olleista 130 artikkelista vain kaksi olivat muita kuin englanninkielisiä, eli jäivät katsauksen ulkopuolelle kielen perusteella. Näistä molemmissa oli artikkelin otsikko, abstrakti ja avainsanat artikkelissa myös englanniksi. Couldry & Hepp (2017, 36) varoittavat medioitumiseen liittyvien ilmiöiden liian eurosentrisestä tarkastelusta. Etenkin keskittyminen pelkästään globaalien datayhtiöiden toimintaan liittämättä sitä uskonnollisiin ja poliittisiin organisaatioihin tai huomioimatta sitä, miten eri tavoilla samat yritykset toimivat eri maantieteellisillä alueilla, tuottaa yksipuolisia ja vääristyneitä tutkimustuloksia, joita ei voida ainakaan yleistää pätemään Euroopan ulkopuolella. Vertailevaa tutkimusta on jo jonkin verran tehty, erityisesti lainsäädäntöön liittyen, mutta lainsäädännön alati muuttuessa ja kehittyessä kerran vuosikymmenessä tehtävät

vertailevat tutkimukset eivät riitä kuvaamaan valtavan nopeaa historiallista muutosta, jota elämme parhaillaan. Tässä yhteydessä tunnistan myös oman tutkielmani eurosentrisyyden, eli sen, että toistettuna esimerkiksi kiinankielisillä hakusanoilla tutkimustulokset luultavasti muuttuisivat eikä tutkielman sisältöä voi yleistää pätemään koko maailman datavuototutkimukseen.

Erityisesti Google Scholar -hakujen osalta kirjallisuuskatsauksen toistettavuuteen vaikuttivat hakukoneen algoritmit, eikä yhtä suurta muutosta hakutuloksissa näkynyt muilla suorittamillani hauilla eri verkkohakemistoissa. Toisaalta kaikki haut niiden lähteestä riippumatta suoritettiin järjestelmien omilla ”järjestä tärkeyden mukaan”-suodattimilla, eivätkä minkään hakujärjestelmän kriteerit algoritmin toimintaan, eli tärkeyden määrittelemiseen, olleet julkisia. Erityisesti Googlen osalta voidaan kuitenkin pohtia yrityksen mahdollisuuksia suodattaa tuloksia siirtämällä keskeisiäkin artikkeleja kauemmaksi hakutuloksissa, etenkin kun Googlen varsinaisen hakukoneen tiedetään jo manipuloivan hakutuloksia (Gilbert 2019). Datavuototutkimuksen osalta Google Scholar voisi teoreettisesti painottaa tuloksia, joissa datavuotoa käsitellään ihmislähtöisenä ongelmana, jossa verkon käyttäjät itse vuotavat itsestään liikaa tietoja verkkoon. Palaan tähän näkökulmaan ja datavuotodiskurssin muutokseen tarkemmin tämän tutkielman alaluvussa 4.1, jossa käsittelen datavuotojen määritelmien muutosta viimeisen kymmenen vuoden aikana.

4 Datavuototutkimus 2010–2022

Datavuotojen tutkimuksessa kuvataan henkilökohtaisen datan keräämistä ja käyttöä lähestyen sitä lähtökohdasta, jossa datan alkuperänä olevalla ihmisellä ei ole tietoa datansa liikkumisesta. Erilaisten älysovellusten mitattavana olemisen kokemista ja mittaamisen toimintalogiikoita on tutkittu laajemmin. Tällaisista lähtökohdista on tutkittu esimerkiksi ihmisten halukkuutta esittää omaa hyvinvointi- ja liikuntadataa puettavilla näytöillä[7] muille ulkoilijoille (Colley et al. 2020), liikuntasuoritusten sovelluksiin kirjaamisen kokemista (Lomborg et al. 2018) ja ihmisruumiin ja hyvinvointidatan yhdistämistä ”oman itsen laboratorioksi” (Kristensen & Ruckenstein 2018). Tällaisissa tutkimuksissa tutkimuskohteet ovat antaneet tietoisen suostumuksensa datan käyttöön niin palveluntarjoajille kuin tutkijoillekin. Henkilökohtaisen datan siirtymistä toisille ja kolmansille osapuolille ei ole tutkittu yhtä laajasti. Toisaalta edellä mainittuja tutkimuksiakin olisi voitu jatkaa tutkimalla sitä, mihin kerätty data jatkaa palveluidentarjoajilta matkaansa, eli vuotaa. Tässä luvussa tarkastelen edellisessä luvussa kuvailemani kirjallisuuskatsauksen tuloksia aloittamalla datavuotoihin liittyvien määritelmien ja sitä kautta yleisen diskurssin muutoksesta, minkä jälkeen käyn läpi tutkimusartikkeleissa käytettyjen aineistojen alkuperää, määrää ja saatavuutta sekä lopulta esille tulleita eettisiä kysymyksiä liittyen datavuototutkimukseen.

Kirjallisuuskatsaukseni artikkelit painottuvat vuosiin 2018–2022, huomioiden erityisesti, että vuoden 2022 julkaistuista artikkeleista mukana ovat vain tammi-helmikuussa julkaistut tekstit. Tämä ei kuitenkaan välttämättä kerro tutkimuksen määrän kasvamisesta viimeisen neljän vuoden aikana, vaikka se on erittäin mahdollista. Muita vaikuttavia tekijöitä voivat olla esimerkiksi hakujärjestelmien algoritmit, jotka mahdollisesti painottavat uudempia julkaisuja hakutuloksissa. Algoritmeja voi ohjata myös esimerkiksi tutkimusartikkeleihin osoitettu kiinnostuksen, eli klikkausten, määrä tietietyltä ajanjaksolta, jolloin hiljattain julkaistut artikkelit jälleen painottuisivat hakutuloksissa.

Tutkimusartikkeleja lainattiin hakemistojen ilmoittamien statistiikkojen mukaan muissa artikkeleissa varsin vähän. Ainut yli sata kertaa muissa julkaistuissa artikkeleissa lainattu teksti oli artikkeli #7.20[8] ”The biggest lie on the Internet: ignoring the privacy policies and

[7] Tutkimuksessa koehenkilöt käyttivät puettavaa teknologiaa (engl. wearable technology), jossa päähineisiin oli integroitu henkilön aktiivisuustietoja ympäristöön visualisoiva näyttö (engl. wearable display).
[8] Luvuissa 4 ja 5 viittaan kirjallisuuskatsaukseni artikkeleihin niille antamillani uniikeilla tunnisteilla (kts. tämän tutkielman alaluku 3.3) tekstin luettavuuden ja lähteiden tulkinnan helpottamiseksi. Tarkemmat tiedot

terms of service policies of social networking services”, joka käsitteli yksityisyydensuojalausekkeiden ja käyttöehtosopimusten lukemista. Useimpia artikkeleja oli julkaisijoiden sivustojen mukaan lainattu alle kymmenessä muiden tekijöiden artikkelissa. Huomioiden humanististen tieteiden tieteellisten artikkelien julkaisuviiveen olevan yli vuoden (Björk et al. 2013, 919) ja kirjoittamisprosessiin kuluvan ajan, vuosina 2021 ja 2022 julkaistuja artikkeleja ei yksinkertaisesti ole ehditty kovin laajasti käyttää muissa julkaistuissa vertaisarvioiduissa artikkeleissa.

Ajallisesti datavuototutkimuksessa pinnalla olleita teemoja, tai trendejä, voidaan ainakin osin selittää aiheisiin liittyvien maailmanlaajuisten tapahtumien kautta. Teemalliset piikit tutkimuksessa kulkevat sykleittäin merkittävien datavuotojen ja tietoturvaloukkausten kanssa. Vuoden 2016 Panaman paperit -tietovuotoa käsiteltiin kattavasti vuonna 2018 ja GDPR sekä Cambridge Analytican -tietovuoto aiheisia artikkeleja julkaistiin ahkerasti vuodesta 2020 eteenpäin. Nähtäväksi jää, toistuuko vastaava syklisyys jatkossa, ja miten viime vuosien suuret datavuodot, kasvava kritiikki datayhtiöitä kohtaan ja esimerkiksi Brexit näkyvät lähivuosina datavuototutkimuksessa.

Kirjallisuuskatsaukseni artikkeleissa ja muussa lähdekirjallisuudessa toistui toive ja tarve innovatiiviseen ajatteluun tutkimusmetodeja ja -aineistoja ideoidessa ja käytettäessä. Useassa lähteessä toistui metafora datan tai algoritmien mustan laatikon (”black box”) ulkopuolelta ajattelusta, jotta päästäisiin tutkimaan algoritmien aikaansaamaa ja datayhtiöiden keräämää sisältöä (Bol et al. 2020, 1997; Doss 2020, 278), koska niin sanotusti perinteisillä tutkimusmenetelmillä ja käytetyillä aineistoilla ei ole päästy riittävästi käsiksi tutkimuksen kannalta välttämättömään dataan. Harvoissa artikkeleissa kuitenkaan esiteltiin käytännön kehitysideoita menetelmien käytöstä ja aineistojen keräämisestä.

Artikkelissa #SB1.1 tarkastellaan monitieteisiä lähestymistapoja verkossa tapahtuvan seurannan, dataistumisen ja alustatalouden tutkimukseen. Artikkelissa todetaan olemassa olevan tutkimuksen keskittyvän teknologioiden toimintaan, jolloin yhteiskunnalliset ja yksilötason vaikutukset jäävät vähemmälle huomiolle. Suhteessa datavuotoihin tämä tarkoittaisi siirtymää datavuotojen kulun seuraamisesta kohti niiden aikaansaamia merkityksiä yhteiskunnassa ja yksilöissä. Tällainen muutos on toki jo havaittavissa, ja tässä luvussa

---

artikkeleista löytyvät tämän tutkielman liitteistä (liite 1 Artikkelikatsauksen viimeisessä vaiheessa mukana olleet artikkelit, liite 2 Kirjallisuuskatsauksen artikkelien koonti).

käsittelen tarkemmin sitä, miten datavuototutkimus on muuttunut viimeisen kymmenen vuoden aikana.

4.1 Datavuotojen määritelmät ja evoluutio

Taulukko 1 Datavuodon määritelmiä tutkimusartikkeleissa

Lähde	Määritelmä
#1.13 Swart et al. (2013), #5.8 AlSabah et al. (2018), #6.4 Olivero et al. (2020)	Datavuoto on hakkerien aikaansaannos, jossa henkilökohtainen data vuotaa julkisesti verkkoon saatavaksi palvelimeen kohdistuneen hakkerihyökkäyksen jälkeen.
#2.8 Paramarta et al. (2018), #7.19 Leering et al. (2020), #SB1.3 Rosso et al. (2020)	Datavuodon yleisin syy on ihminen, jonka tehtävä on suojata dataa. Vuoto voi olla esimerkiksi työntekijän sosiaaliseen mediaan kirjoittama salatuksi tarkoitettuja tieto yrityksestä tai virhe yrityksen tietojen suojaamisesta vastaavan henkilön työnteossa.
#2.6 Maris et al. (2020, 2018)	Datavuodossa verkkosivustot vuotavat tietoja käyttäjistään kolmansille osapuolille, kuten Googlelle ja Metalle, ilman käyttäjän tietoista hyväksyntää.
#7.2 Yu & Shen (2021)	Datavuoto on poliittisesti aktiivisten henkilöiden tietojen vuotaminen valtion toimielinten saataville.
#5.5 Gan & Jenkins (2015, 87)	Käyttäjät vuotavat henkilökohtaisia tietoja vapaaehtoisesti ja ymmärtämättä, että kuka tahansa voi lukea niitä, ei vain ystävät ja seuraajat.
#6.3 Sparks et al. (2016, 40; 43)	Datavuoto voidaan käsittää Deleuzen pakolinjana, pakenemisen lentoratana ja mutaationa, joka kehittyy eteenpäin asettaen uusia yhteyksiä entiteettien välille. […] Vuoto muodosti pakolinjan, paon näennäisesti suljetusta tilasta tutkijan ja osallistujan välisessä käyttöliittymässä, joka johti ensin muihin tiloihin verkkosivulla ja sitten osallistujien muodostaman putken läpi laajemmille alueille käyttöliittymän ulkopuolelle.

Harvoissa kirjallisuuskatsaukseni artikkeleista eksplisiittisesti määriteltiin datavuoto. Lisäksi artikkelit, joissa tarkempia määritelmiä käytettiin, olivat tietojärjestelmä- tai tietojenkäsittelytieteiden julkaisuja, jotka eivät muuten olleet tutkielmani kannalta oleellisia. Määritelmissä oli myös suuria eroja, kuten yllä olevista esimerkeistä (taulukko 1) käy ilmi. Tarkkojen määritelmien sijaan tulokulma datavuotoon oli löydettävä artikkelin aineiston kautta. Esimerkiksi verkkoon vuodettujen salasanalistojen tutkijat lähestyivät datavuotoja verkkoon julkisesti hakkerien toimesta vuodetun materiaalin kautta. Toisaalta spesifien määritelmien puuttuessa ei voida myöskään sulkea pois sitä, että tutkijat kokisivat datavuodon käsitteen sisältävän muutakin kuin vain tutkimusartikkelinsa lähtökohtien mukaisen tilanteen. Nämä lähtökohdat huomioon ottaen kirjallisuuskatsaukseni ulkopuolelle jäi epäilemättä

artikkeleja, joissa jostain näkökulmasta käsitellään datavuotoon verrattavaa tilannetta, mutta siitä ei käytetty ilmaisua ”data leak” tai ”data leakage”.

Ajallisesti etenkin 2010-luvun alun artikkeleissa datavuoto määriteltiin henkilön, esimerkiksi yrityksen työntekijän, toiminnan vuoksi vuotaneiksi arkaluontoisiksi tiedoiksi tai hakkerin rikollisilla tavoilla käsiinsä saamien tietojen vuotamiseksi verkkoon. Tällainen diskurssi, jossa datavuoto määritellään ihmisen syyksi joko niin, että rikollinen jakaa henkilökohtaista dataa verkkoon, tai datavuoto on käyttäjän itsensä huolimattomuudellaan tai tietämättömyydellään aiheuttama, tukee datayhtiöiden toimintaa. Niin kauan kuin keskustelu pysyy riittävän vahvojen salasanojen keksimisen ja ihmisten oman verkkokäyttäytymisen ympärillä, huomio ei kiinnity datayhtiöiden toimintaan. Vielä vuonna 2009 Googlen silloinen toimitusjohtaja Eric Schmidt totesi verkkoyksityisyyteen liittyen, että ”If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place” (Esguerra 2009) ja vuotta myöhemmin Facebookin perustaja Mark Zuckerberg totesi yksityisyyden lakanneen olemasta sosiaalinen normi (Johnson 2010). 2020-luvulle siirryttäessä datavuototutkimuksen lähtökohdat ovat siirtyneet aiemmasta yksityisyyden ja salailun rinnastavasta, verkon käyttäjän vastuuta omista tiedoistaan kertomisessa, painottavasta stalkkeritutkimuksesta (#5.5, 2015) sen selvittämiseen, millaisia stalkkereita esimerkiksi Google ja Meta ovat (#2.6, 2020).

Uhrin syyttämisen retoriikka ei ole kuitenkaan kokonaan poistunut datavuotojen ympärillä käytävästä keskustelusta. Huomion kiinnityttyä datayhtiöiden toimintaan yritykset ovat perustelleet datan keräämistä ja käyttöä käyttäjien ajan säästämisellä, verkossa toimimisen helpottamisella ja sillä, että käyttäjät itse haluavat dataistuneen käyttökokemuksen eri yritysten palveluille (Doss 2019, 93). Vastaavasti dataa keräävät yritykset vetoavat käyttöoikeussopimuksiin ja tietosuojakäytäntöihin, jotka käyttäjä on raksinut hyväksyvänsä. Käsittelen näihin sopimuksiin liittyvien ongelmien tutkimusartikkeleita tarkemmin tämän tutkielman alaluvussa 5.3.

Määritelmällisesti lähellä datavuotoa on tietoturvaloukkauksen ja tietomurron käsite (engl. data breach), joka näkyi myös kirjallisuuskatsaukseni artikkeleissa ja kirjallisuuskatsaukseni hakuprosessissa. Doss (2020, 13) määrittelee tietoturvaloukkauksen tapahtumaksi, jossa yksilöstä tallennettuja, identiteettivarkauden tai taloudellisen hyväksikäytön mahdollistavia (esim. luottokorttitiedot, sosiaaliturvatunnus), tietoja vuotaa kolmannelle osapuolelle. Määritelmä perustuu usean Amerikan osavaltion kuluttajasuojalainsäädäntöön ja sen tarkoitus on asettaa tietoja vuotaneet yritykset ilmoitusvastuuseen yksilölle. Dossin määrittelemä

tietoturvaloukkaus eroaa datavuodosta sen tarkoituksessa ja tietojen laadussa; tietoturvaloukkauksen (tai tietomurron) alkuperä on aina laiton, datavuoto sen sijaan ei sitä ole.

Tosin molemmissa prosesseissa yksilöä koskeva data liikkuu yksilön ja yrityksen väliltä kolmannelle osapuolelle, eikä eroa ole myöskään sillä, tapahtuuko se yksilön tiedostamana vai tiedostamatta. Niin ikään molempiin prosesseihin liittyy riskejä identiteettivarkauteen ja taloudelliseen hyväksikäyttöön. Jos ajatellaan esimerkiksi Googlen välittämiä tietoja kolmansille osapuolille, verkonkäyttäjän hyväksikäyttö on vain huomattavasti hienovaraisempaa, kun luottokortti pyritään tyhjentämään ja poliittinen äänestysvalinta muuttamaan psykografisesti kohdennettujen mainoskampanjojen avulla sen sijaan, että rikolliset tahot suoraan tyhjentävät luottokortin kopioidulla numerolla. Lisäksi datavuodon tapauksessa datan lähteelle eli verkon käyttäjälle ei kukaan ole ilmoitusvastuussa siitä, mitä tietoja kolmansille osapuolille on vuotanut ja miten niitä käytetään[9].

Datavuodon (data leak, data leakage) ja tietoturvaloukkauksen (data breach) käsitteitä käytetään osin päällekkäisesti akateemisessa kontekstissa. Tässä tutkielmassa olen sisällyttänyt kirjallisuuskatsaukseen myös joitain artikkeleja, joissa terminologisesti käsitellään tietoturvaloukkauksia (#6.19, 2022; #7.19, 2020; #SB2.2, 2020). Tämä ensinnäkin siksi, että tutkielmani alaluvussa 2.3 määrittelen käsitteleväni datavuotoa yksilölähtöisten tietojen siirtymisenä kolmannelle tai kolmansille osapuolille ilman yksilön tietoista suostumusta, joten tietomurrosta johtuvat vuodot mahtuvat määritelmääni datavuodosta. Toiseksi kyseiset artikkelit ovat löytyneet edellisessä luvussa määrittelemieni hakukriteerien perusteella. Kirjallisuuskatsauksen hakuprosessin edetessä pohdin myös hakutermien laajentamista, koska osa tietoturvaloukkauksia käsittelevistä artikkeleista käsitteli selkeästi datavuotoja. Resurssit ja tutkimuskysymykset huomioon ottaen se ei olisi ollut mahdollista.

Uniikeista datavuodon määritelmistä esille nousi artikkeli #6.3 (2016), jossa datavuoto nähdään Deleuzen ja Guattarin (1988) termein pakolinjana. Artikkelin tapaustutkimuksessa datavuoto tapahtui eräänlaisessa mikroympäristössä, jossa yliopiston kurssille osallistuvien oppilaiden päiväkirjanomaisia tekstejä vuosi verkko-opiskeluympäristössä kaikkien

[9] Lisäksi viime vuosina on keskustelu myös yksityishenkilöihin ja sosiaalisen median vaikuttajiin kohdistuvasta doksauksesta (engl. doxing, doxxing), jossa julkisesti saatavilla olevia tietoja, kuten kotiosoitteita tai puhelinnumeroita, jaetaan internetissä vastoin henkilön tahtoa ja usein satuttamistarkoituksessa. Doksauksessa harmia aiheutetaan vuotamalla periaatteessa julkisia tietoja uusille alustoille. Ilmiö voitaneen laskea yhdeksi näkökulmaksi datavuotoihin, vaikka sitä ei tutkimukseni artikkeleissa suoranaisesti käsiteltykään.

muidenkin oppilaiden luettavaksi. Tällainen käsitys datavuodoista on kirjallisuuskatsaukseni artikkeleista ehkä laajin. Se ottaa huomioon myös vuotojen kerrannaisvaikutukset vuodon kohteena olevan ihmisten lisäksi näiden ympärillä oleviin ihmisiin. Näin datavuoto ei oikeastaan pääty koskaan, vaan jatkaa kiertoaan ihmisten keskellä, jolloin datavuodoista muodostuu yhteiskunnan rihmasto. Samalla vuoto myös osittain siirtyy verkon ulkopuolelle, ihmisten välisiin suhteisiin ja heidän kokemiinsa tunnereaktioihin. Ajatus kestää tarkastelua aivan yksinkertaisella konkretian tasollakin; datayhtiöille liikkuneiden datavuotojen aikaansaannoksena muodostuu älylaitteidemme näytöille jokaiselle uniikki todellisuus, joka ohjaa elämäämme ja samalla omien läheistemme elämää.

Datavuotojen tutkimus ja datavuodon määritelmä ovat muuttuneet 2010-luvulta eteenpäin käsi kädessä teknologian kehittymisen ja datayhtiöiden nousun kanssa. Vielä 2010-luvun alussa keskityttiin tutkimuksissa esimerkiksi siihen, miten käyttäjän itse itsestään julkaisemia tietoja yhdistelemällä niin sanotut stalkkerit voivat selvittää esimerkiksi käyttäjän kulkureittejä ja työpaikan sijaintia (#5.5, 67). Vuosikymmenen lopussa monet tutkijat kiinnittivät huomionsa isojen kansainvälisten yritysten datavuotoihin, joissa internetiin päätyi muun muassa asiakkaiden osoite- ja pankkitietoja (#1.13). Lisäksi erilaiset lainsäädäntöjen vertailut (#6.17) olivat tavallisia aineistoja. 2020-luvulla fokus on siirtynyt datayrityksiin ja erityisesti käyttäjän tiedostamattomiin ja tahattomiin datavuotoihin, joissa datayhtiöt keräävät ja tallentavat käyttäjistä tietoja. Monet tutkijat ovat kiinnostuneet datayritysten keräämästä informaatiosta ja sen jakamisesta kolmansille osapuolille (esim. #2.6) ja yrittävät löytää keinoja päästä käsiksi tällaisia datavuotoja kuvaaviin aineistoihin (esim. #5.7).

Terminologisesti datavuoto on muuttunut naiivin internetin käyttäjän itse verkkoon luovuttamista tiedoista kohti datayritysten laajamittaisen datankeräämisen käsittelemistä datavuotona. Samalla henkilökohtaista dataa tavoitteleva osapuoli on vaihtunut muista ihmisistä dataa keräävіin ja hyödyntäviin yrityksiin. Verkon käyttäjästä on näin ollen tullut aktiivisen tekijän sijaan passiivinen objekti, jonka mahdollisuudet vaikuttaa datavuotoihin ovat rajalliset. Samalla vuotojen hyödyntäjät ovat muuttuneet muista, usein rikollisista, verkonkäyttäjistä suuryrityksiksi. Toisaalta tutkimusartikkeleissa todettiin myös, että vaikka mahdollisuuksia vaikuttamiseen olisi, harvat kuitenkaan tarttuvat niihin. Esimerkiksi artikkelin #6.2 tutkimuksessa tietotekniikan maisteriopiskelijoista koostuneen koeryhmän jäsenet kertoivat kyselytutkimuksessa datan tallentamisen ja käyttöön vaadittavien suostumusten vaikuttavan ostopäätöksiin sovelluksia ladatessa, mutta käytännön

etnografisessa asetelmassa samat henkilöt latasivat sovelluksia välittämättä vaadituista datan käyttöön ja tallentamiseen liittyvistä suostumuksista.

Samalla kun akateemisen tutkimuksen lähestymistapa ja suhtautuminen datavuotoihin on muuttunut vuosikymmenen aikana ja edelleen muuttumassa, datayhtiöt pyrkivät pitämään kiinni diskurssista, jossa datavuotojen ”syyllinen” on joko ihminen itse, tai verkossa pahoilla aikeilla liikkuva hakkeri tai stalkkeri. Sen sijaan, että yritykset toisivat selkeästi ilmi oman datan tallentamisensa laajuuden ja laadun, ne keskittyvät muistuttamaan verkon käyttäjiä kirjautumaan ulos palveluistaan julkisia tietokoneita käyttäessä ja käyttämään erikoismerkeistä ja satunnaisista kirjaimista koostuvia salasanoja, joita kukaan ei voi muistaa, mutta jotka voi kätevästi tallentaa vaikkapa Googlen palvelimille, jolloin niitä ei tarvitse itse muistaa.

4.2 Menetelmälliset lähestymistavat

Tässä alaluvussa käsittelen keinoja, joiden avulla datavuototutkimukseen on kerätty ja saatu aineistoja. Kirjallisuuskatsaukseni artikkelit lähestyivät datavuotojen tutkimusta useimmin kvalitatiivisista lähtökohdista, tai ainakin hyödynsivät laadullisen tutkimuksen menetelmiä tutkimuksessaan ollen monimenetelmällisiä tutkimuksia. Ainoa poikkeus oli ison otannan (6691 vastausta) kyselytutkimusta hyödyntänyt artikkeli #7.2, jossa ei ollut ollenkaan avoimia kysymyksiä ja aineistoa käsiteltiin tilastollisesti.

Tutkimusmetodilla, eli -menetelmällä, viitataan niihin tapoihin, joilla konkreettisesti kerätään ja analysoidaan tutkimusaineistoa (Puusa et al. 2020, 9). Kirjallisuuskatsaukseni artikkeleissa käytetyimmät tutkimusmenetelmät olivat kysely, erilaisilla empiirisillä analyyseilla ja kokeilla toteutetut tapaustutkimukset, sekä verkkolähteistä tehdyt datakaavinnat. Tämä on osittain ristiriidassa aikaisemman tutkimuksen kanssa, jonka mukaan viime vuosina datavuotoja on tutkittu yhteiskuntatieteissä lähinnä omaehtoisen, yksilön itse suorittaman digitaalisen mittaamisen näkökulmasta ja empiiristä tutkimusta on tehty vähän, koska datavuotojen kolmansien osapuolten toimintamallit koetaan liian monimutkaisiksi (Helles et al. 2020, 1959). Useissa artikkeleissa käytettiin myös eri menetelmien yhdistelmiä.

Taulukko 2 Kirjallisuuskatsauksen artikkelien tutkimusmenetelmät

Erityyppisiä kyselyitä käytettiin ainakin osittaisena tutkimusmenetelmänä kahdeksassa artikkelissa. Kyselyillä mitattiin sosiaalisen median alustojen vuotojen tiedostamisen vaikutusta käyttäjiin (#2.8), verkkopalveluihin väärien henkilötietojen antamisen syitä (#6.14), omien matkapuhelimien työpaikoilla käyttämisen datavuotoriskejä (#6.23), teknologisen osaamisen, tietoturvariskien tiedostamisen ja varallisuuden vaikutuksia mobiilisovellusten ostopäätöksiin (#6.32), poliittisen aktiivisuuden vaikutusta yksityisyyden suojaamiseen verkossa (#7.2), työpaikan tietoturvaohjeiden noudattamiseen liittyviä tekijöitä (#7.19) ja erityisesti haavoittuville ihmisryhmille suunnatun markkinointisisällön näyttämistä sosiaalisen median käyttäjille (#SB1.4). Näistä tutkimuksista mobiilisovellusten ostopäätöksiä tarkasteltiin kyselyn ja empiirisen kokeen yhdistelmällä, ja markkinointisisällön kohdentamisen tutkimuksessa kyselyyn osallistujat latasivat lisäksi verkkoselaimeensa lisäosan, joka mittasi verkkoliikennettä.

Taulukko 3 Kirjallisuuskatsauksen kyselyt

Tutkimus	Tutkimuskysymys	Aineiston keruumenetelmä
#2.8 Paramarta & al. (2018), Indonesia	Miten tietoisuus sosiaalisen median alustojen datavuodoista vaikuttaa käyttäjiin?	Kysely verkossa
#6.14 Zhou et al. (2022), Kiina	Mitkä tekijät saavat verkkopalveluiden käyttäjän antamaan vääriä henkilökohtaisia tietoja?	Skenaariokysely verkossa
#6.19 Wang et al. (2022), Kiina, Australia, Englanti	Milloin ja miten hotellialalla voidaan ja kannattaa kompensoida datavuotojen uhreja?	Skenaariokysely verkossa
#6.23 Ameen & al. (2021), Englanti, Oman, Yhdysvallat	Miten omien matkapuhelinten käytön aiheuttamia riskejä voitaisiin vähentää työpaikoilla erityisesti Gen-Mobile (18-35v.) työntekijöiden keskuudessa?	Kysely kasvotusten

#6.32 Barth & al. (2019), Alankomaat	Vaikuttaako yksityisyysparadoksiin käyttäjän teknologinen osaaminen, tietoisuus tietoturvasta ja/tai varallisuus?	Taustatietojen kerääminen kyselyllä, lisäksi tapaustutkimus
#7.2 Yu & Shen (2021), Kiina (Hong Kong)	Miten politiikkaan osallistuminen verkkoympäristössä vaikuttaa yksityisyyden suojaamiseen liittyvään käytökseen verkossa?	Kysely, kysely-yrityksen tuottama
#7.19 Leering & al. (2020), Alankomaat, Suomi, Ruotsi	Mitkä tekijät vaikuttavat työntekijöiden piittaamattomuuteen yrityksen tietoturvaohjeista?	Kysely, sähköpostitse
#SB1.4 Bol et al. (2020), Alankomaat	Miten sosiodemografiset tekijät vaikuttavat sosiaalisen median (Facebook) käyttäjille näytettävään markkinointisisältöön huomioiden erityisesti haavoittuvissa asemissa olevat ihmisryhmät?	Taustatietojen kerääminen kyselyllä, lisäksi mitattiin osallistujien verkkoliikennettä

Kyselyihin vastanneet olivat tyypillisesti joko yliopisto-opiskelijoita, vähintään kandidaatin tutkinnon suorittaneita tai työntekijöitä teknologia-alan yrityksissä, joiden tietosuojaohjeiden noudattamista tutkittiin. Keskeisiä haasteita kyselyiden käyttämisessä oli riittävän vastausmäärän saaminen. Erityisesti tämä vaikutti tutkimukseen, jonka tarkoituksena oli perehtyä haavoittuvien ihmisryhmien näkemään kohdistettuun markkinointiin ja mahdollisiin eroihin nähdyssä sisällössä eri sosiodemografisten ihmisryhmien välillä (#SB1.4). Tutkimus toteutettiin seurantatyökalun ja kyselyn yhdistelmänä ja yhteistyössä hollantilaisen tutkimuspaneelin kanssa, jolloin alkuperäinen otanta koostui väestön keskiarvokotitalouksista. Yli seitsemänsataa kotitaloutta suostui osallistumaan tutkimukseen lataamalla verkkoliikennettä mittaavan lisäosan selaimeensa ja 567 osallistujaa vastasi kyselyyn. Tutkimuseettisistä syistä verkon käyttö oli mahdollista myös yksityisessä tilassa, jolloin liikennettä ei mitattu. Verkon käyttöä mitannutta aineistoa saatiin lopulta vain 97 osallistujalta, joista 80 oli vastannut myös kyselyyn. Tutkimuksen loppuun suorittaneet taas olivat väestöllistä keskimäärää nuorempia, naisia ja tottuneempia toimimaan verkossa, jolloin tutkimuksen tilastollinen merkittävyys kärsi ja nimenomaisesti haavoittuvia ihmisryhmiä koskevaan tutkimuskysymykseen vastaaminen vaikeutui.

Toisaalta erityisesti laajoihin verkkokyselyihin liittyy myös tutkimuseettisiä haasteita. Kyselytutkimuksia tekevien yritysten ympärille on muodostunut halpatyöympäristö, jossa heikosti toimeentulevat ihmiset täyttävät pitkiä verkkokyselyitä saaden palkkioksi rahan sijasta yrityksen omia ”pisteitä”, joita voi tarpeeksi kerättyään muuttaa lahjakortiksi tai

tuotepalkinnoiksi. Vastineeksi saatavien pisteiden rahallinen arvo on erittäin alhainen verrattuna kyselyihin käytettyyn aikaan, puhumattakaan erittäin henkilökohtaisten tietojen jakamisesta yritykselle ja siitä mahdollisesti eteenpäin kolmansille osapuolille. (Doss 2020, 156–158). Toki voidaan pohtia myös sitä, onko tietoinen henkilökohtaisten tietojen luovuttaminen edes nimellistä maksua vastaan reilumpi vaihdanta kuin tiedostamaton tietojen vuotaminen yritykselle vastineeksi tarjotun verkkosivun tai -palvelun käyttämisestä ”ilmaiseksi”. Joka tapauksessa tällaisia massakyselyitä tuottavia yrityksiä käytettiin ainakin kahden tutkimuksen kyselyosuuden toteuttamisessa (#7.2 ja #6.14).

Kyselyihin ja haastatteluihin tutkimusmenetelminä liittyy myös haaste siitä, että tutkimuksen kohteina olevat ihmiset voivat antaa virheellisiä tai vääriä tietoja, tietoisesti tai tiedostamatta (Hyvärinen et al. 2017, 113). Tämä kävi erityisen hyvin ilmi yhdistelmämenetelmätutkimuksessa, jossa tarkasteltiin mobiilisovelluksen ostajien teknologisen osaamisen, tietoturvatietoisuuden ja varallisuuden vaikutusta ostopäätökseen (#6.32). Tutkimuksessa 66 tietojärjestelmätieteiden maisteriopiskelijaa tekivät taustatietokyselyn, jonka jälkeen heidän piti valita ja ladata mobiilisovellus omiin puhelimiinsa. Kyselyn perusteella sovelluksen vaatimat käyttöoikeudet olivat merkittävä tekijä sovellusta valitessa, mutta empiirisen seurannan perusteella eniten vaikuttavat tekijät olivat hinta, sovelluksen saamat arvostelut ja sen design. Kirjallisuuskatsaukseni artikkelien perusteella kysely tutkimusmenetelmänä olisi hyvä yhdistää johonkin toiseen menetelmään, koska yhdistelmämenetelmillä toteutetut tutkimukset poikkeuksetta osoittivat epäjohdonmukaisuuksia kyselyvastausten ja tutkimuskohteiden toiminnan välillä.

Väärien tietojen antamiseen liittyen artikkelissa #6.14 lähestyttiin datavuotoja tutkimalla eri verkkoalustoilla mahdollisesti datavääristymiin johtavaa henkilökohtaisten tietojen vääristelemistä. Neutralisaatioteorian ja skenaariokyselyn yhdistelmällä pyrittiin selvittämään syitä henkilökohtaisten tietojen vääristelemiselle. Kysely teetettiin edellä mainitun kaltaisella kiinalaisella kyselyitä tuottavalla yhtiöllä. Lisäksi tutkimukseen yhdistettiin viiden suuren persoonallisuuspiirteen teoria, jonka avulla pyrittiin selittämään taipumusta vääristellä verkkoon annettuja henkilökohtaisia tietoja. Asetelma väärien tietojen antamista koskevasta kyselytutkimuksesta on mielenkiintoinen, koska mikään ei varsinaisesti estä kyselyyn kymmenen sentin palkan (#6.14, 8) toivossa vastaavia henkilöitä vääristelemästä tietojaan tai naputtamalla mitä tahansa vastausta päästäkseen älypuhelimella kyselyn läpi siihen ”vastanneena”.

Toinen tutkimuseettinen näkökulma datavuototutkimuksen suhteesta niin kyselyiden kuin myös haastattelujen käyttöön liittyy siihen, että ihmisiä pyydetään kertomaan omaan verkkovälitteiseen yksityisyyteensä liittyvistä asioista. Ne lähtökohtaisesti kuuluvat, etenkin EU:n alueella, tietosuojan piiriin. Haastattelujen tai kyselyiden aineistosta voi muodostua henkilörekisteri, jossa toistetaan samoja tietoja, joiden keräämiseen, varastointiin ja uudelleenkäyttöön suhtaudutaan kriittisesti ja joiden vuotamista kolmansille osapuolille tutkitaan. Lisäksi mahdollisuutena on tutkimusdatasta lähtevän datavuodon päätyminen tutkimukseen osallistuneiden henkilöiden kannalta haitallisille tahoille. Etenkin kyselyitä tuottavien yritysten kanssa toimiessa mahdollisuutena on myös, että kerättyä dataa käytetään lopulta kaupallisiin tarkoituksiin yritysten myydessä datapakettejaan myös muille kuin akateemista tutkimusta suorittaville tahoille.

Artikkelissa #7.2 tutkittiin politiikkaan aktiivisesti osallistuvien henkilöiden oman yksityisyyden suojaamiseen liittyviä käyttäytymismalleja ja huolta datavuodoista. Kymmenessä Aasian maassa toteutetun kyselyn perusteella poliittinen aktiivisuus korreloi positiivisesti aktiivisen omien tietojen suojaamiseen pyrkimisen kanssa ja vastaavasti matalan kyberturvallisuuden maissa pelko omien tietojen vuotamisesta on suurempi. Kiinaa ei kuitenkaan voitu ottaa tutkimukseen mukaan, koska kyselydataa poliittisesta osallistumisesta ei voitu saada. Valtiollista valvontaa ja datavuotoja erilaisiin kansallisiin rekistereihin tutkittiin tämän lisäksi vain artikkelissa #7.4, jossa perehdyttiin Intian avustusjärjestelmärekisteriin, johon valtion tukea haluavien ihmisten on pakko rekisteröityä. Valtion valvontaa sivuavien artikkelien aineistoja kerättiin näin ollen vain Aasian maissa ja Intiassa.

Taulukko 4 Kirjallisuuskatsauksen haastattelut

Tutkimus	Tutkimuskysymys	Aineiston keruumenetelmä
#6.12 Haynes & al. (2016), Englanti	Miten Englannin datasuojalaki (Data protection act 1998) toimii ja millainen on Englannin datasuojan toimintaympäristö erityisesti sosiaalisen median alustat huomioiden?	Asiantuntijahaastattelu
#SB2.2 Hinds et al. (2020), Englanti	Miten Cambridge Analytica -skandaali vaikutti Facebook-profiileihin ja asenteisiin datan käytöstä?	Teemahaastattelu

Vaikka erilaiset haastattelut ovat laadullisen tutkimuksen eniten käytetyin aineiston keruumenetelmä yhteiskuntatieteellisessä tutkimuksessa (Puusa et al. 2020, 99), kirjallisuuskatsaukseni artikkeleista haastattelututkimuksia oli vain kaksi (6.67 %). Haastatteluiden avulla tutkittiin henkilökohtaisen datasuojan toimintaympäristöä erityisesti sosiaalisen median alustat huomioiden Isossa-Britanniassa (#6.12) ja Cambridge Analytica -skandaalin vaikutuksia Facebook-profiileihin sekä asenteisiin henkilökohtaisen datan käytöstä (#SB2.2). Datasuojan toimintaympäristöön ja lainsäädäntöön keskittyvässä tutkimuksessa haastateltiin kymmentä eri alojen asiantuntijaa kasvotusten tai puhelimitse ja Cambridge Analytica -tapauksen vaikutuksia tutkittiin teemahaastatteluilla, joihin osallistui 30 yliopisto-opiskelijaa tai yliopistolla työskentelevää henkilöä.

Niin asiantuntijoiden haastattelemiseen kuin teemahaastatteluihin liittyy erityispiirteitä. Haastattelujen tekemisen ja käytön haasteina ovat muun muassa arvojen ja merkitysten tutkimisen haasteellisuus, sekä haastattelijan ja haastateltavan välisen suhteen ja kommunikaation vaikutus saatavaan aineistoon (Puusa et al. 2020, 102–104). Erityisesti haastattelutilanteen hierarkiaan ja haastattelutilanteeseen tutkimusartikkeleissa vaikutti se, että teemahaastattelut järjestettiin yliopisto-organisaation sisällä, kun taas asiantuntijahaastattelut toteutettiin tutkimusorganisaation ulkopuolella vaikuttavien henkilöiden kanssa. Asiantuntijuudelle ei ole olemassa spesifiä määritelmää, joten viime kädessä asiantuntijuuden määrittelee oman tutkimusaiheensa mukaisesti tutkija itse (Hyvärinen et al. 2017, 182). Aineistoni artikkelissa asiantuntijarooliin oli valittu useita korkeamman yksityisyydensuojan puolesta puhuvien, voittoa tavoittelemattomien, organisaatioiden johtohenkilöitä, oikeustieteellisen tiedekunnan apulaisprofessori, sekä verkkomainonta- ja suoramarkkinointiyhdistyksen puheenjohtajat. Toisaalta mukana oli myös henkilöitä, joiden ammatillista roolia ei erikseen artikkelissa määritelty. Tutkimuksen tekijät olivat myös tutkimansa aiheen asiantuntijoita, joten haastattelutilanteissa ei vaikuttanut olleen hierarkkista asetelmaa verrattuna esimerkiksi tilanteeseen, jossa akateemista tutkimusta tekevä henkilö tutkii jotain aivan oman alansa ulkopuolista ilmiötä tai aihetta.

Sen sijaan yliopiston opiskelijoilla ja työntekijöillä toteutetussa haastattelututkimuksessa (#SB2.2) tavoitteena oli saada mahdollisimman monimuotoinen ihmisryhmä huolimatta siitä, että tutkimukseen rekrytoitiin osallistujia vain yliopiston sisäisten viestintäkanavien välityksellä. Haastatteluihin valittiin esimerkiksi eri-ikäisiä, eri vaiheessa opintojaan olevia opiskelijoita ja yliopiston hallinto-, myynti- ja kirjanpitotehtävissä työskenteleviä henkilöitä, joilla ei ollut suoraa yhteyttä akateemiseen tutkimukseen tai välttämättä edes akateemista

koulutusta. Tutkimusmenetelmän haasteina oli kuitenkin edustavan otoksen saaminen, koska toteutunut otos ei edustanut kansallista keskiarvoa. Otoksen ulkopuolelle jäivät kokonaan esimerkiksi lapset, eläkeläiset, työttömät ja nuoret, jotka eivät ole yliopisto-opiskelijoita. Lisäksi haastattelut koskivat osittain haastateltavien Facebookin käyttöä, jota ei mitattu millään tavalla, eli tiedot käyttäytymisen mahdollisesta muutoksesta tulivat tutkimuskohteilta itseltään. Verrattuna esimerkiksi aiemmin tässä alaluvussa käsiteltyyn mobiilisovellusten ostopäätöksiin vaikuttaneita tekijöitä mitanneeseen tutkimukseen (#6.32), ihmisillä on taipumus kaunistella yksityisyyteensä liittyvää käyttäytymistä verkossa, vaikka todellisuudessa tietoisuus datavuodoista ei vaikuttaisi siihen millään tavalla.

Taulukko 5 Kirjallisuuskatsauksen tapaustutkimukset

Tutkimus	Tutkimuskysymys	Aineiston keruumenetelmä
#5.7 Schufrin & al. (2021), Saksa	Miten GDPR:n avulla saatavilla olevista datavuodoista voitaisiin muotoilla käyttäjälle selkeitä?	Tapaustutkimus, prototyyppiohjelmiston kehittäminen 37 vapaaehtoisen henkilökohtaisella datalla
#6.3 Sparks et al. (2016), Uusi-Seelanti	Millaisia eettisiä riskejä ja dilemmoja digitaaliseen tutkimustyöhön sisältyy? Mitä tapahtuu, kun arkaluontoista tutkimusdataa vuotaa?	Tapaustutkimus, (auto)etnografinen kuvaus datavuodosta yliopistokurssin oppimisympäristössä
#6.32 Barth & al. (2019), Alankomaat	Vaikuttaako yksityisyysparadoksiin käyttäjän teknologinen osaaminen, tietoisuus tietoturvasta ja/tai varallisuus?	Tapaustutkimus, lisäksi taustatietojen kerääminen kyselyllä
#7.20 Obar & Oeldorf-Hirsch (2018), Kanada, Yhdysvallat	Miten ihmiset lukevat yksityisyydensuojalausekkeita ja verkko-ohjelmistojen käyttöehtosopimuksia?	Tapaustutkimus, mitattiin yksityisyydensuojalausekkeiden ja käyttöehtosopimusten lukemiseen käytettyä aikaa. Lisäksi itsearviointi.
#SB1.6 Breuer et al. (2020), Saksa	Miten digitaalisten jälkien dataa voidaan hankkia tutkimuskäyttöön, ja millaisia eettisiä ja käytännön haasteita siihen liittyy?	Tapaustutkimus, aineistona oma aiempi tutkimus. Lisäksi kirjallisuuskatsaus.
#H.1 Taylor (2016), Alankomaat	Millaisia eettisiä ja menetelmällisiä haasteita liittyy tutkimuksiin, joissa tarkastellaan ihmisten liikkumista matkapuhelimista kerättävän datan avulla matalan toimeentulon maissa?	Tapaustutkimus, aineistona teleoperaattorin tutkimuskäyttöön julkaisemasta datapaketista tehdyt tieteelliset julkaisut.

Kaikki laadullinen tutkimus ei ole tapaustutkimusta, vaikka tapaustutkimusta kutsutaankin usein spesifin menetelmän sijaan tutkimusstrategiaksi tai -lähestymistavaksi ja erityyppisiä

tapaustutkimuksia tehdään laajasti eri tieteenfilosofisissa kehyksissä (Eriksson & Koistinen 2015, 2; 4). Tapaustutkimukset voidaan jaotella yksittäis- ja monitapaustutkimuksiksi, joille yhteistä on tarkasti, esimerkiksi paikan, ajan, tai muun kriteerin perusteella määritellyn tapauksen analysointi ja ratkaiseminen (Eriksson & Koistinen 2015, 4). Molempiin tapoihin yhdistetään usein muita menetelmiä, kuten kyselyitä, avoimia haastatteluja, dokumenttien keräystä ja analysointia tai niiden havainnointia (Eriksson & Koistinen 2015, 3). Pelkkään kyselyyn ja sen analysointiin verrattuna tapaustutkimuksessa asetetaan tulokset määritetyn yksittäisen tapauksen kontekstiin ja tapaustutkimuksen tavoitteena on ymmärtää, selittää ja kuvata tutkimusaihetta mahdollisimman monipuolisesti, jolloin sen avulla voidaan tulkita monimutkaisia ja muuttuvia kokonaisuuksia (Puusa et al. 2020, 202; Eriksson & Koistinen 2015, 4).

Tapaustutkimuksen on todettu olevan hyvä lähestymistapa, jos tutkimuskysymyksen keskiössä ovat ”mitä”, ”miten” ja ”miksi” -kysymykset, tutkimuskohde on tämänhetkinen ilmiö, aiheesta on vähän aiempaa empiiristä tutkimusta tai tutkijan pystyy vaikuttamaan vähäisesti tapahtumiin (Eriksson & Koistinen 2015, 5). Näistä usea kriteeri täyttyy datavuototutkimuksessa oikeastaan automaattisesti, koska ilmiö itsessään on niin tuore, että aiempaa empiiristä tutkimusta on vähän ja kyseessä on myös äärimmäisen ajankohtainen tutkimusaihe. Tutkimuskysymysten osalta kirjallisuuskatsauksessani tapaustutkimuksen avulla vastattiin esimerkiksi kysymyksiin ”Miten GDPR:n avulla saatavilla olevista datavuodoista voitaisiin muotoilla käyttäjille selkeitä” (#5.7), ”Miten ihmiset lukevat yksityisyydensuojalausekkeita ja verkko-ohjelmistojen käyttöehtosopimuksia” (#7.20) ja ”Miten digitaalisten jälkien dataa voidaan hankkia tutkimuskäyttöön ja millaisia eettisiä ja käytännön haasteita siihen liittyy” (#SB1.6).

Kaikkiin edellä mainittuihin tutkimuksiin yhdistettiin eri menetelmiä. Datavuotojen visualisointiin kehitettiin prototyyppiohjelmisto, jonka lisäksi sen käyttäjille teetettiin kysely, jonka avulla selvitettiin koekäytön vaikutuksia koehenkilöihin. Yksityisyydensuojalausekkeiden ja verkko-ohjelmistojen lukemiseen käytetyn ajan mittaamiseen yhdistettiin itsearviointi ja kysely. Digitaalisten jälkien tutkimuskäyttöön hankkimisen tapauksena oli saman tutkijan aiempi oma tutkimus, jonka lisäksi artikkeliin sisällytettiin kuvaileva kirjallisuuskatsaus.

Erityisesti yksittäistapaustutkimuksiin kohdistetaan kuitenkin usein kritiikkiä tutkimuksen toistettavuudesta ja siitä, että tutkimuksen kohteena oleva tapahtumasarja voi olla

poikkeustapaus normin sijaan. Lisäksi etenkin yksittäistapaustutkimuksen ollessa kyseessä vertailun haasteellisuutta pidetään ongelmallisena. (Puusa et al. 2020, 197–200; 202.) Yksittäistapaustutkimuksen käyttöä on perusteltu esimerkiksi vakiintuneen teorian testaamisen tai täysin poikkeuksellisen tai harvinaisen tapauksen yhteydessä (Puusa et al. 2020, 202). Kirjallisuuskatsauksessani tällainen poikkeuksellinen yksittäistapaus oli tutkimusasetelma, jossa täysin muuta tutkimusta tehdessä opiskelijoiden tuottamia, luonteeltaan henkilökohtaisia, tekstejä vuosi verkko-oppimisympäristöön (#6.3). Palaan artikkeliin vielä tarkemmin tämän tutkielman alaluvussa 4.4., jossa käsittelen datavuototutkimuksen eettisiä haasteita, sekä alaluvussa 5.4., jossa pohdin (auto)etnografian laajemman hyödyntämisen mahdollisuuksia datavuototutkimuksen tulevaisuudessa.

Tapaustutkimuksen juuret ovat etnografiassa, ja etnografisia menetelmiä hyödynnetään edelleen tapaustutkimuksessa. Kaikessa tutkimuksessa tutkija jollain asteella tulkitsee aineistoaan, mutta etnografiassa tutkijan suora tulkinta ja kerronnallisen tiedon tuottaminen korostuvat (Eriksson & Koistinen 2015, 34). Tapaustutkimuksen osalta yksittäistapauksenkin ollessa kyseessä tutkimuskohteen vertailu tapahtuu sisäänrakennetusti; tutkijapositionsa vuoksi etnografi vertaa tutkimustapaustaan väistämättä omiin kokemuksiinsa ja mahdollisesti myös alkuperäisiin odotuksiinsa (Puusa et al. 2020, 204). Etnografisia piirteitä oli kirjallisuuskatsauksessani useammassakin tutkimuksessa, mutta erityisesti (auto)etnografinen tutkimusasetelma korostui jo edellä mainitussa #6.3 artikkelissa, jossa tutkija käsitteli myös omia tunteitaan datavuodon eräänlaisena aiheuttajana.

Taulukko 6 Kirjallisuuskatsauksen datakaavinnat ja vapaasti verkossa saatavilla olevat aineistot

Tutkimus	Tutkimuskysymys	Aineiston keruumenetelmä
#1.13 Swart et al. (2013), Etelä-Afrikka	Mitä dataa ja missä muodossa internetiin vuosi 2013 Etelä-Afrikan hakkeritapauksessa? Miten vuodetun datan visualisointi vaikuttaa sen kokemiseen?	Datakaavinta, aineisto tietomurrosta verkkoon vuodettu
#2.6 Maris, et al. (2020), Yhdysvallat	Mitä ja miten paljon käyttäjädataa pornosivustoilta vuotaa kolmansille osapuolille, mitä kolmannet osapuolet ovat, voiko seksuaaliset kiinnostuksenkohteet vuotaa kolmansille osapuolille?	Datakaavinta, verkkosivujen läpikäynti webXray-ohjelmistolla
#5.5 Gan & Jenkins (2015), Englanti	Mitä kaikkea tietoa Twitterin käyttäjästä voidaan saada pelkästään yhden (sijaintimerkityn) twiitin kautta?	Datakaavinta vapaasti verkossa saatavilla ohjelmistoilla julkisista Twitter-päivityksistä

Tutkimus	Tutkimuskysymys	Aineiston keruumenetelmä
#5.8 AlSabah et al. (2018), Qatar	Miten kulttuuri ja kieli vaikuttavat salasanojen valintaan?	Datakaavinta, aineisto tietomurrosta verkkoon vuodettu
#6.4 Olivero et al. (2020), Espanja	Mitä tietoja (dataa) kohdehenkilöistä löytyi julkisena internetistä, mitä informaatiota voidaan luoda yhdistelemällä näitä tietoja ja voiko henkilöstä itsestään lähtöisin oleva verkkopresenssi johtaa haavoittuvuuteen?	Datakaavinta, aineisto kyselytutkimuksen tietovuoto
#SB1.2 Helles et al. (2020), Tanska	Miten verkkosivujen kolmansien osapuolten (third-party services) käyttö eroaa eri EU-maiden, alueiden ja erityyppisten sivustoiden välillä?	Datakaavinta, verkkosivujen läpikäynti WebXray-ohjelmistolla
#6.9 Ong (2012), Kiina (Hong Kong)	Miten Malesian ja Hong Kongin datasuojalait eroavat muiden maiden vastaavista? Keskittyykö lainsäädäntö organisaatioiden toiminnan legitimoimiseen vai yksilön datan suojaamiseen?	Saatavilla vapaasti verkossa, lainsäädäntötekstit
#6.17 Greenleaf & Park (2014), Australia, Etelä-Korea	Miten Etelä-Korean Personal Information Privacy Act of 2011 eroaa muista vastaavista laeista?	Saatavilla vapaasti verkossa, lainsäädäntötekstit
#7.4 Pawan (2019), Australia	Millaisia yksityisyydesuojaan kohdistuvia ongelmia biometristä tunnistusta hyödyntävällä Aadhar avustusjärjestelmällä on Intiassa?	Saatavilla vapaasti verkossa, laki- ja asiakirjatekstejä sekä aktivistien nauhoittamia videohaastatteluja
#SB1.3 Rosso & al. (2020), Yhdysvallat, Kanada	Miten yksilöiden verkkohakukäyttäytyminen ja taloudelliset muuttujat (pörssikurssit) reagoivat Edward Snowdenin tietovuotoon?	Saatavilla vapaasti verkossa, verkkohakujen statistiikkoja, pörssikursseja

Datakaavintojen avulla saatiin käyttöön kahdenlaista tutkimusaineistoa. Julkisista lähteistä, kuten Twitteristä[10] ja verkkosivustoilta kaavittua dataa, sekä julkisesti verkossa saatavilla olevaa dataa, jonka alkuperä oli kuitenkin tietomurroissa. Jälkimmäiseen asetelmaan palaan tarkemmin alaluvussa 4.4, jossa käsittelen datavuototutkimuksen tutkimuseettisiä haasteita. Sosiaalisen median alustojen datakaavinnoissa ongelmana on, että alustojen käyttö vaatii

[10] Twitter siirtyi vuonna 2023 Elon Muskin yksityisomistukseen ja yrityskauppojen myötä palvelun nimi vaihtui X:ksi. Selkeyden vuoksi käytän tässä tutkielmassa alustasta sen aikaisempaa nimeä, koska aihetta koskevat tutkimusartikkelit on kirjoitettu ennen omistajanvaihdosta.

sisäänkirjautumista. Lisäksi julkisesti saatavilla on aina rajoitettu määrä dataa; jotkut päivitykset ovat julkisia, mutta useimmat ovat näkyvissä vain julkaisijan omalle sidosryhmälle (”ystäville” tai ”seuraajille”). Poikkeuksena tähän on Twitter, jossa käyttäjien julkaisut ovat julkisia. Siksi sitä pystytään hyödyntämään myös tutkimuskäytössä. Artikkelissa #5.5 tutkittiin datakaavinnan avulla sitä, kuinka paljon yksittäisestä käyttäjästä pystytään selvittämään tietoja pelkkien twiittien perusteella. Tässä yhteydessä sivuttiin myös muita sosiaalisen median alustoja, kun yhden alustan tietojen perusteella todettiin muiden käyttäjätunnusten löytämisen olevan varsin todennäköistä.

Vapaasti verkossa saatavilla oleva data oli useissa tutkimuksissa #6.9; #6.17; #7.4) eri maiden laki- ja asiakirjatekstejä. Julkisten asiakirjojen erityyppiset vertailut ovat toteutettavissa ilman monimutkaisia teknologisia innovaatioita, koska tekstit ovat pääasiassa vapaasti saatavilla mistä päin maailmaa tahansa. Toisaalta spesifisti datavuotoja koskeva lainsäädäntö muuttuu ja päivittyy usein, joten ajantasaisten vertailujen julkaiseminen on haastavaa.

Ainakin kahdessa tutkimuksessa (#2.6; #SB1.2) hyödynnettiin WebX-ray-ohjelmistoa, koska ohjelmisto mainitaan nimeltä tutkimusartikkeleissa. Kyseessä on veloituksetta verkossa saatavilla avoimen lähdekoodin ohjelmisto, jonka avulla voi kerätä tietoja, analysoida verkkosivujen liikennettä ja sisältöä, kerätä tietosuojaselosteita ja tietoja sivustoista, jotka keräävät käyttäjädataa. Lisäksi ohjelmiston avulla pystytään myös listaamaan kolmansien osapuolten sivustoja, joihin analysoitavista sivustoista on yhteys. (WebXray 2023.)

4.3 Aineiston määrä, saatavuus ja analysointi

Datavuotoja koskevan aineiston saatavuuteen ja käytettävyyteen vaikuttaa se, että iso osa henkilökohtaisesta datasta on suuryritysten keräämää. Yritykset käyttävät itse dataa palveluidensa parantamiseen ja niiden myymiseen muille yrityksille, jolloin datan saaminen tutkimuskäyttöön vaatii yhteistyötä yksityisten (datan omistavat yritykset) ja julkisten (tutkivat tahot, esim. yliopistot) sektorien välillä (#SB1.6, 2058). Viime vuosina useat datayhtiöt, esimerkiksi Meta (Facebook) ovat kuitenkin vaikeuttaneet tai aktiivisesti estäneet keräämänsä datan käyttöä akateemisessa tutkimuksessa (#SB1.6, 2071). Kun datan kerääminen ja sen myyminen muille yrityksille on yrityksen ydinliiketoimintaa, ei tietoja haluta vastikkeettomasti luovuttaa edes yleishyödylliseen käyttöön. Toisaalta asetelmaan linkittyy myös yksityisyydensuojaan ja tutkimusetiikkaan liittyviä kysymyksiä. Aineistossani teleoperaattorin tutkimuskäyttöön luovuttaman, puhelinliikennetietoja sisältäneen

anonymisoidun datapaketin, todettiin mahdollistavan puhelinliikenteen paikantamisen yksilötasolle saakka (#H.1, 331–332).

Vaikka dataa saataisiinkin tutkimuskäyttöön datayhtiöiltä, suuryritysten mahdollistamaan datan käyttöön linkittyy haasteita puolueettoman ja ohjailemattoman tutkimuksen tekemisen suhteen. Monipuolisimpien datapakettien käyttöön saaminen edellyttää usein sopimussuhdetta niitä tarjoavan yrityksen kanssa, jolloin tutkimus voi jäädä kokonaan datan keränneen yrityksen sisäiseen käyttöön, tai sen julkaisua voidaan ainakin valvoa ja rajoittaa (#SB1.6, 2061). Lisäksi yksityisyydensuojaa koskeva lainsäädäntö rajoittaa ja sääntelee myös tutkimustyöhön saatavia aineistoja, siis maissa, joissa lainsäädäntö on olemassa ja ajan tasalla. Näistä lähtökohdista tutkimusaineistojen saaminen puolueettomaan akateemiseen tutkimukseen dataa kerääviltä yhtiöiltä, eli taholta jonne dataa vuotaa, on lähes mahdotonta.

Kirjallisuuskatsaukseni tutkimusartikkelien aineistoista oli erotettavissa selkeästi julkisesti saatavilla olevat aineistot, kuten esimerkiksi laki- ja säädöstekstit, sekä sosiaalisen median alustoista Twitter. Lisäksi erotettavissa oli aineistokategoria, jossa tutkittiin ja vertailtiin salasanojen muodostamista, käyttämällä aineistoina tietosuojaloukkauksista peräisin olevia, vapaasti verkossa saatavilla olevia listoja salasanoista. Salasana- ja lainsäädäntövertailut olivat myös ainoita aineistoja, joiden avulla tehtiin vertailuja eri maiden ja kulttuurien välillä. Toisaalta salasanalistojen käyttäminen tutkimusaineistona herätti myös kysymyksen siitä, mitä julkisesti verkkoon vuodettua dataa voi tutkimuseettisesti käyttää tutkimusaineistona.

Lakitekstien vertailussa käytetty asetelma katsaukseni artikkeleissa (#6.9; #6.17) oli jonkun toisen alueen sääntelyn vertaaminen vastaavaan eurooppalaiseen sääntelyyn. Tutkimukset olivat ajalta ennen Euroopan unionin yleistä tietosuoja-asetusta, mutta vertailupohjana oli Euroopan unionin direktiivit, Iso-Britannian datasuoja-asetukset, sekä artikkelissa #6.9 OECD:n (Organisation for Economic Co-operation and Development, taloudellisen yhteistyön ja kehityksen järjestö) yksityisyydensuojaperiaatteet. Vertailujen ulkopuolelle jäi Yhdysvaltojen aihetta koskevat säädökset.

Monitieteisissä, tietojenkäsittelyä yhteiskuntatieteelliseen tutkimukseen yhdistelevissä, artikkeleissa yleinen tutkimusasetelma oli erilaiset empiiriset kokeet teknologian hyödyntämiseen liittyen. Tietojen vuotamista kolmansille osapuolille paikannettiin webXray:n avulla (#SB1.2) ja datavuotojen visualisointiin kehitettiin ohjelmistoprototyyppi (#5.7). Toisena luovempana tutkimusaineistojen käytön esimerkkinä esiin nousi erilaisten aineistojen yhdistäminen. Artikkeleissa yhdistettiin esimerkiksi etnografinen koe samalle

koeryhmälle tehtyyn kyselyyn (#6.32), ja anonyymien DuckDuckGo -verkkohakujen käyttöstatistiikkoja sekä tietoturvayhtiöiden pörssikursseja verrattiin ajallisesti Snowden-paljastusten kanssa (#SB1.3).

Nykyinen datan keräämistä ja käyttämistä kuvaava tutkimus jättää dataistumisen trendiä seuraten isoja ihmisryhmiä tutkimuksen ulkopuolelle. Aineistossani selkeästi yleisin tutkittava ihmisryhmä oli (korkeakoulu)opiskelijat (kts. esim. #SB1.4; #6.32; #SB2.1), joiden käyttäminen selittyy helpolla saatavuudella. Vaikka tutkimustasolla tiedostettaisiin opiskelijoiden olevan varsin homogeeninen ihmisryhmä, on erilaisten tutkimusharhojen muodostuminen erittäin mahdollista. Esimerkiksi tietoista suostumusta tietojen keräämiseen käsittelevät tutkimukset antaisivat hyvin erilaisia tuloksia vaikkapa vanhainkodin asukkailla tehtynä. Haavoittuvien ihmisryhmien, kuten lasten, vanhusten ja maahanmuuttajien, jääminen tutkimuksen ulkopuolelle on ongelmallista myös siinä mielessä, että juuri heihin kohdistuu esimerkiksi valvontaa, johon voidaan hyödyntää dataa keräävää teknologiaa. Valvontaa suoritetaan huolenpidon nimissä, johon sisältyy riski tietojen käyttämisestä muuhun kuin alkuperäiseen tarkoitukseen (ns. function creep) (Taylor 2016, 330).

Lisäksi esimerkiksi yksityisyydensuojalausekkeita tutkiessa on todettu niiden ymmärtämisen vaativan vähintään kahden vuoden korkeakouluopintoja (Maris et al. 2020, 2027), joten tietoon perustuvan suostumuksen tutkimuksessa tulisi käyttää myös muita kuin korkeakoulutettuja ihmisryhmiä. Toisaalta edes kyselytutkimukset, jotka toteutettiin käyttämällä tutkimusryhmän ulkopuolisia yrityksiä kyselydatan tuottamiseen, eivät tavoittaneet iällisesti heterogeenistä vastaajaryhmää (#7.2, 6). Koska monet haavoittuvista ihmisryhmistä ovat myös markkinointiyritysten otollisinta maaperää, olisi näiden ryhmien tavoittaminen myös tutkimuksen näkökulmasta ensiarvoisen tärkeää.

Monikansallisissa tutkimuksissa painottui aineistona lainsäädäntöä koskevat tutkimukset, joissa vertailtiin eri maiden datan keräämistä koskevaa sääntelyä. Tapaustutkimukset ja etnografiset seurannat suoritettiin länsimaiden maaperällä. Diskurssi kansainvälisistä ja globaaleista datankeräämiseen ja käyttöön liittyvistä haasteista vaikuttaakin todellisesti käsittävän pelkästään Euroopan ja Amerikan. Aineistoni artikkeleissa muutamissa sivuttiin tätä ilmiötä, esimerkiksi #H.1 artikkelissa huomioitiin kansainvälisyyden diskurssin kätkevän datamarkkinoilla vallitsevan epätasa-arvon.

4.4 Datavuototutkimuksen eettiset kysymykset

Kirjallisuuskatsaukseni viimeisen vaiheen artikkeleista (n=27) neljässätoista eksplisiittisesti mainittiin tutkimusetiikka jossain yhteydessä. Tarkastellessani tutkimusasetelmien toistettavuutta (EU-maassa), eettiset syyt olivat keskeisin este toistettavuudelle. Viidessä artikkelissa tutkimusasetelmat olivat eettisesti kyseenalaisia. Näistä kolmessa (#1.13; #5.8, #6.4) aineisto koostui tietomurtojen yhteydessä verkkoon vuodetusta datasta, ja sitä käytettiin ilman datavuotouhrien suostumusta. Artikkelissa #6.3 osana aineistoa oli tapahtumaketju, jossa yliopiston kurssilla muiden opiskelijoiden luettavaksi vuosi virheellisesti päiväkirjamaisia tekstejä, joten toistaminen vaatisi uuden vastaavan tapahtumaketjun tuottamisen. Tämän lisäksi kyseinen artikkeli käsitteli tutkimusartikkeleja, joiden aineistona oli kansainvälisen Orange-teleoperaattorin tutkimuskäyttöön luovuttama datapaketti. Paketti sisälsi tiedot kaikesta yrityksen asiakkaiden puhelinliikenteestä Norsunluurannikolla, eikä vastaavia tietoja enää tässä laajuudessa luovuteta kolmansille osapuolille. Toistettavuuden kannalta eettisten kysymysten lisäksi esteet ovat myös juridisia, kun dataa koskevaa lainsäädäntöä päivitetään jatkuvasti ympäri maailmaa. Kaksi jälkimmäistä tutkimusta nimenomaisesti käsittelivät akateemiseen tutkimukseen liittyviä haasteita, eli artikkelit eivät itsessään olleet eettisesti kestämättömiä.

Aineistojen keräämistavoista eettisesti haastavimpana erottui datakaavinnat verkossa julkisesti saatavilla olevista materiaaleista. Artikkeleista erottui selvästi artikkeli, jossa etiikkaa ei mainittu millään tavalla, ja aineisto oli tietomurrosta peräisin (#1.13). Kahdessa muussa tietomurtoaineistossa (#5.8, #6.4) tutkimusasetelma oli hyväksytty instituution omassa eettisessä neuvostossa ja tutkimusasetelman käyttö pyrittiin perustelemaan. Tietomurtoaineistojen käyttämisessä akateemisessa kontekstissa ongelmana on myös datan alkuperän, eli vuodon uhreiksi joutuneiden ihmisten, suostumuksen saaminen. Vaikka esimerkiksi salasanavuotoja analysoivissa artikkeleissa ei tietojen kohdentaminen tiettyyn henkilöön ole mahdollista, voi oman salasanan näkeminen tutkimusaineistossa, tai vaikkapa tutkimuksesta tehdyssä lehtiartikkelissa, aiheuttaa henkilökohtaista haittaa.

Useissa katsaukseni artikkeleissa (mm. #SB1.6; #SB1.4) peräänkuulutettiin innovatiivisia tutkimusmenetelmiä ja keinoja aineistojen keräämiseen. Paradoksaalisesti juuri tällaisten, poikkeavien, aineistojen ja menetelmien käyttöön vaikutti liittyvän tutkimuseettisiä haasteita. Useat ”innovatiiviset” tutkimukset myös toteutettiin alueilla, joiden datalainsäädäntö ei ole ajan tasalla. Esimerkiksi osana #H.1 artikkelia oli aineisto, jonka monikansallinen

teleoperaattoriyritys Orange luovutti tutkijoiden käyttöön Norsunluurannikolla. Monialaisen tutkimusprojektin aikana kävi ilmi, että harvaanasutuilla alueilla datan anonymisointi ei onnistunut. Samaan tulokseen päädyttiin myös artikkelissa #1.13 eteläafrikkalaisella aineistolla. Lisäksi ensimmäisessä tapauksessa länsimaalaisista tutkijoista koostetulta tutkijaryhmältä puuttui kulttuurillinen konteksti datan tulkitsemiseen, mikä johti vääriin tulkintoihin aineistosta. Artikkeli oli osa laajempaa tutkimusprojektia, jonka keskeisiä tuloksia oli, että vastaavanlaisen aineiston käyttö on eettisesti kyseenalaista, tarvitaan selkeitä ohjeistuksia henkilökohtaisten data-aineistojen luovuttamisesta tutkimuskäyttöön ja kansainvälisten suuryritysten tulisi suorittaa parempaa eettistä itsesääntelyä tapauksissa, joissa ne operoivat monien erilaisten lainsäädäntöjen alueilla. Jälkiviisaina voidaan pohtia, tarvittiinko tällaisiin tuloksiin teleoperaattorin käyttäjien puhelutietojen julkaisemista yhden vuoden ajalta tutkimuskäyttöön ilman käyttäjien tietoista suostumusta.

Käsittelemistäni tutkimusartikkeleista nousi esille myös tutkimuksia, joiden aineistona käytettiin verkossa vapaasti saatavilla olevia tiedostoja eri tietovuodoista. Käytännössä tiedot olivat rikollista alkuperää, koska ne olivat peräisin tietomurroista ja hakkereiden verkkoon vuotamia. Tällaisten aineistojen avulla visualisoitiin datavuotoja (#1.13, aineisto ja artikkelin kirjoittajien kotiyliopistot eteläafrikkalaisia) ja tarkasteltiin kielen ja kulttuurin vaikutusta salasanojen valintaan (#5.8, aineisto ja artikkelin kirjoittajien kotiyliopistot qatarilaisia). Molemmissa artikkeleissa aineistoja käytettiin kvantitatiiviseen analyysiin ja siitä tehtäviin johtopäätöksiin. Tutkimuksia yhdisti myös se, että datankeruuprosessia ei eksplisiittisesti määritelty. Aineistojen todettiin olleen saatavilla sosiaalisessa mediassa (#5.8) ja eri pastebin-sivustoilla (#1.13).

Niin datavuotojen visualisointia kuin salasanojen muodostamista tutkittiin kirjallisuuskatsaukseni artikkeleissa kuitenkin myös eettisesti kestävästi. Datavuotoja esimerkiksi visualisoitiin tutkimusryhmän itse kehittämän prototyyppiohjelmiston avulla ja käytetty data saatiin tutkimukseen osallistuneilta vapaaehtoisilta (#5.7). Lisäksi kyseinen työryhmä pystyi vastaamaan artikkelin #1.13 tutkimuskysymykseen siitä, miten vuodetun datan visualisointi vaikuttaa sen kokemiseen, johon ei artikkelissa #1.13 käytetyillä kvantitatiivisilla menetelmillä pystytty vastaamaan. Vastaavasti salasanojen keksimistä tarkasteltiin artikkelin #5.8 lisäksi artikkelissa #SB2.1 jossa salasana-aineisto koostui vapaaehtoisten opiskelijoiden keksimistä salasanoista. Aineistoni tutkimusartikkelien perusteella tietomurroista saatujen aineistojen käyttö ei siis ollut välttämätöntä, koska vastaavia tutkimuksia pystyttiin tekemään myös eettisesti kestävästi.

Koska eettisesti kyseenalaisimmat julkaisut ovat peräisin länsimaiden ulkopuolelta, aineistoni perusteella olisi helppo vetää johtopäätös siitä, että eettisesti haastavien tutkimusten tekeminen ei ole ongelma länsimaissa. Aineistonkeruukriteereinäni oli kuitenkin artikkelien julkaisu kansainvälisessä tiedejulkaisussa ja niin kauan, kun artikkeleja julkaistaan kansainvälisissä tiedejulkaisuissa, hyväksytään myös aineistojen ja menetelmien käyttö. Toki yhtenä valintakriteerinä oli myös koko artikkelin saatavuus englannin kielellä, joten ilman paneutumista muilla kielillä julkaistuihin artikkeleihin, ei eettisen kestävyyden maantieteellisistä eroista voida tehdä johtopäätöksiä tämän kirjallisuuskatsauksen perusteella. Datavuototutkimuksessa korostuu kuitenkin verkkoaineistoja koskeva eettinen kysymys siitä, voiko aineistoja käyttää vain sillä perusteella, että ne ovat verkossa vapaasti saatavilla.

Yhtä helppoa olisi kuvitella tutkimusetiikan kulkevan vain eettisemmän tutkimuksen suuntaan, mutta aineistosta erottui kaksi artikkelia, joissa kulku oli päinvastainen. Julkaisuissa #5.5 vuodelta 2015 ja #6.4 vuodelta 2020 tutkittiin millaisen kokonaisuuden yksittäisestä käyttäjää koskevasta tiedosta saa koostettua vapaasti verkossa olevasta materiaalista. Varhaisemmassa tutkimuksessa lähtödatana käytettiin muutaman vapaaehtoisen paikannustiedot sisältäneitä twiittejä, mutta myöhäisempi tutkimusasetelma oli rakennettu niin, että tutkittavia henkilöitä ei informoitu tehtävästä tutkimuksesta (kts. lainaus tämän kappaleen alapuolella). Tapauksessa nousee esille keskeinen kysymys tutkimusetiikasta, eli se, mitä kaikkea voi perustella "hyvällä aikomuksella".

> "Our aim was not to use these data for any commercial or harming purpose, but as a real scenario to validate the presented approach, without using synthetic data. Thus we performed the study without notifying the involved persons, because their informed consent or voluntary participation would have been a potential threat to validity." (#6.4, 2)

Artikkelin #6.4 tutkimus on läpäissyt niin kirjoittajien kotiyliopistojen eettisen neuvoston, kuin artikkelin julkaisijankin, vaatimukset julkaistavalle artikkelille. Artikkelissa viitataan lisäksi Euroopan unionin lainsäädäntöön ja artiklaan, jonka perusteella anonymisoitua dataa voidaan tutkijoiden tulkinnan mukaan käyttää ilman henkilön tietoista suostumusta. Tarkoitus pyhittää keinot -näkökulmassaan tutkimus ei ole kuitenkaan aineistossani ainutlaatuinen, vaan ainakin osin samalla mentaliteetilla datavuototutkimusta lähestytään myös esimerkiksi jo aikaisemmin tässä alaluvussa mainituissa artikkeleissa #5.8 ja #H.1. Jälkimmäisen tosin huomauttaessa myös, että datavuotoja koskevan tiedon anonymisoinnin tekniikoihin ja teknologioihin tulisi kiinnittää erityistä huomiota myös yliopistotutkijoiden koulutuksessa (#H.1, 41).

Datavuototutkimukseen liittyy monia eettisiä kysymyksiä, jotka linkittyvät aineistojen saatavuuden haasteisiin, teknologioiden käyttöön ja kehittymiseen, sekä alati kehittyvään lainsäädäntöön. Lisäksi tutkimusta hankaloittaa eräänlainen sisäänrakennettu vaatimus siitä, että julkisilla varoilla rahoitetun akateemisen tutkimuksen tulisi olla eettisesti ja moraalisesti korkeammalla tasolla datan kaupalliseen käyttöön verrattuna. Siinä missä datayhtiöt käyttävät datasuojalainsäädännöltään jäljessä olevia alueita testialueina, herättää vastaava toiminta akateemisessa kontekstissa kysymyksiä eettisyydestä, kuten esimerkiksi tässä tutkielmassa laajalti käsitellystä tutkimuksesta #H.1 käy ilmi. Mikäänhän ei sinänsä estä tutkijaa tai ketä tahansa muutakaan ihmistä menemästä pimeään verkkoon, lataamasta sieltä tietovuotomateriaaleja ja analysoimasta niitä, mutta viimeistään julkaisuvaiheessa ei voi olla törmäämättä eettisiin kysymyksiin.

Lisäksi datatutkimustyötä monimutkaistaa akateemisen tutkimuksen uskottavuuteen liittyvä vaatimus aineiston jaettavuudesta tutkimuksen toistettavuuden ja läpinäkyvyyden nimissä (Breuer et al. 2020, 2062). Yksilön oikeudesta omaan dataansa, henkilökohtaisen datan suojaamisesta ja aineiston jaettavuuden vaatimuksesta muodostuu ristiriita, jossa tutkija rinnastuu datayhtiöihin. Herää kysymys, millä tavoin tutkimukseen osallistunut henkilö voi antaa tietoisen suostumuksen tutkimukseen osallistumisesta, jos kerätyt tiedot jaetaan kolmansille osapuolille.

5 Datatutkimusta tulevaisuudessa

Tässä luvussa käsittelen kirjallisuuskatsauksessani esiinnousseita haasteita aikaisemmassa datavuototutkimuksessa ja pohdin ratkaisumahdollisuuksia niihin. Aineistoni pohjalta hahmottelen aikaisemman tutkimuksen tietoaukkoja paikantaen tulevan tutkimuksen tarvetta ja mahdollisuuksia kirjallisuuskatsauksestani esiin nousseiden teemojen avulla. Erityisesti käsittelen tutkimuskohteiden inklusiivisuuteen ja yleisemmin datanhallinnan valtarakenteisiin liittyviä haasteita, tutkimusryhmien monitieteistä ja kansainvälistä yhteistyötä ja lopuksi autoetnografian ja yhteistoiminnallisen autoetnografian mahdollisuuksia datavuototutkimuksessa.

5.1 Datavuototutkimuksen inklusiivisuus ja datanhallinnan valtarakenteet

Kuten aiemmin tässä tutkielmassa olen tuonut esille, ihmisten toimintaa ja asenteita kuvaamaan pyrkivä datavuototutkimus on painottunut aineiston saatavuuden vuoksi länsimaihin ja erityisesti jo valmiiksi diginatiiveihin opiskelijoihin. Tämän lisäksi tutkimukset on usein tehty pienellä otannalla; tyypillisesti korkeakoulun datatutkimukseen, yksityisyydensuojaan tai verkkoinfrastruktuuriin liittyvällä kurssilla. Tyypillisesti opiskelijat käsitellään yhtenä ryhmänä, eikä heitä profiloida pienempiin segmentteihin. Syyt opiskelijoiden tutkimiseen ovat ennen kaikkea käytännöllisiä ja taloudellisia, jonka lisäksi pienen otoksen profilointi ei ole yksityisyydensuojasyistä mahdollista. Tutkimus on logistisesti helppo ja nopea toteuttaa opiskelijoilla, koska heitä ei tarvitse erikseen rekrytoida tutkimukseen, jolloin myös tutkimusbudjetti on helpommin hallittavissa. Aineistoni perusteella datavuototutkimuksen ulkopuolelle jää enemmän ihmisiä, kuin sen sisäpuolelle. Lisäksi osa muita kuin länsimaalaisia nuoria aikuisia koskeneista tutkimuksista oli eettisesti vähintäänkin harmaalla alueella.

Tarve haavoittuvien ihmisryhmien tutkimukselle kuitenkin tiedostetaan. Tutkimuksissa on esimerkiksi pyritty perehtymään verkkosisällön kohdentamiseen spesifisti haavoittuviin ihmisryhmiin (#SB1.4). Tutkimuksen haasteena oli, että kyselyn jälkeen toteutetun seurannan jätti kesken yli 86 % tutkimukseen osallistuneista. Vaikka tutkimuksen tavoitteena oli saavuttaa keskimääräisotanta Hollannin kansalaisista, lopputuloksena oli seurantadataa nuorten, diginatiivien naisten datavuodoista. Valitettavasti keskeyttäneiltä ei kysytty syitä siihen, joten selvittämättä jäi, johtuiko keskeyttäminen esimerkiksi haluttomuudesta tiedostaa

omia verkkokäyttäytymismallejaan, tai vaikkapa tutkimukseen suunnitellun seurantatyökalun vaikeakäyttöisyydestä.

Yhdenkään aineistoni tutkimuksen kohderyhmänä ei eksplisiittisesti olleet esimerkiksi vanhukset, kehitysvammaiset, kuulo- tai näkövammaiset, alaikäiset lapset tai toimintarajoitteiset. Haavoittuvien kohderyhmien tutkimiseen liittyy läheisesti eettisiä kysymyksiä, joita ei voida ohittaa, mutta tutkimuksen painottuminen pelkästään työkykyisten, usein korkeasti koulutettujen ja jo valmiiksi datavuotoihin liittyviä ongelmia tiedostavien nuorten aikuisten, keskuuteen on ongelmallista. Lisäksi esimerkiksi pienituloiset, vanhukset ja lapset eivät ole datayhtiöiden näkökulmasta mielenkiintoisia johtuen rajoittuneesta ostovoimasta ja erityisesti lasten osalta markkinointiin liittyvistä rajoituksista (Doss 2020, 76–77), joten vertailevan otannan saaminen eri ihmisryhmistä kertoisi myös riippuuko datavuotojen aste yksilön rahankäytön potentiaalista.

Haavoittuvien ihmisryhmien lisäksi kirjallisuuskatsaukseni artikkeleista puuttui myös toinen yhteiskunnallinen ääripää: rikkaat. Varallisuuden merkitystä datavuotojen kokemiseen tai syntymiseen ei spesifisti tutkittu yhdessäkään artikkelissa. Yhdessä tutkimuksessa pyrittiin huomioimaan varallisuuden vaikutus mobiilisovelluksien hankintaan antamalla kokeeseen osallistuneille henkilöille riittävä varallisuus, eli kymmenen euroa, yhden sovelluksen hankintaan (#6.32). Mittaamatta jäi kuitenkin, olisiko päätöksentekoprosessi muuttunut, jos osallistujilla olisi ollut käytössään kymmenen euron sijaan vaikkapa sata, tai tuhat, euroa mobiilisovelluksen hankintaan. Mikään yhteiskuntaluokka ei ole immuuni datavuodoille, ja viime vuosina julkiseen keskusteluun ovat nousseet julkisuuden henkilöiden kohtaamat tietomurrot. Julkisuuteen nousseiden tietomurtojen kohteena ovat usein naiset, ja tarkemmin heidän esimerkiksi pilvipalveluihin tallentamat alastonkuvansa (Chavez et al. 2014; Flynn 2021; Pullen 2017). Julkisuuden ja korkean varallisuustason vaikutuksia datavuotoihin ei kuitenkaan aineistoni artikkeleissa suoraan käsitelty.

Verkossa tapahtuvia ilmiöitä koskevan tutkimuksen yhteydessä otettiin esille myös otantaharhan mahdollisuus silloin, kun tutkimukseen osallistetaan vain henkilöitä, joilla on esteetön pääsy internetiin (#6.3, 41). Vaikka äkkiseltään voisi ajatella datavuototutkimuksen edellytyksenä olevan tutkimuskohteiden esteetön ja rajoittamaton pääsy internetiin, ei näin kuitenkaan välttämättä ole. Vaikka Suomessa olemme tottuneet rajoittamattomiin datapaketteihin ja vähäisiin verkkoyhteyksien katvealueisiin, useissa maissa verkon käyttöä voivat rajoittaa siihen liittyvät maksut tai verkkoyhteyden saatavuus. Puhumattakaan

esimerkiksi lapsista ja vanhuksista, joiden verkon käyttöä voivat edellä mainittujen lisäksi rajoittaa myös vaikkapa lähisukulaiset, tekninen osaaminen tai laitteiden saatavuus.

Länsimaiden ulkopuolelta tulevissa aineistoissa, lainsäädäntötekstit poissulkien, esiintyi merkittäviä eettisiä haasteita koskien niin aineistojen hankintaa kuin aineistojen tulkintaa ilman riittävää ymmärrystä paikallisesta kulttuurista tai yhteiskunnasta. Kun tähän yhdistetään datayhtiöiden datan keräämiseen ja esimerkiksi verkon käyttöoikeuksiin liittyvä toiminta alemman sääntelyn maissa, ja toisaalta samojen yhtiöiden tarve varmistaa elektroniikkalaitteisiin tarvittavien raaka-aineiden, kuten koboltin, saatavuus, olisi kattavalle ja tutkimuseettisesti toteutetulle datavuototutkimukselle kysyntää myös länsimaiden ulkopuolisilla aineistoilla.

Ylipäätään mediahistoria kytkeytyy valtaan ja sen käyttöön (kts. esim. Couldry & Hepp 2017, 36), ja sama asetelma näkyy myös datavuotojen tutkimuksen maantieteellisessä ja historiallisessa jakaumassa, sekä tutkittavien aineistojen saatavuudessa. Datavuototutkimuksen osalta merkittäviä tekijöitä ovat esimerkiksi suurten datayhtiöiden käyttämä valta olla luovuttamatta aineistoja tutkimuskäyttöön, sekä paikallisten lainsäädäntöjen vaikutus käytettäviin aineistoihin. Vaikka tietyt aineistot ovat vapaasti tai ainakin lähes vapaasti saatavilla verkossa, paikallinen lainsäädäntö ja dataistumisen aste yhteiskunnassa vaikuttavat siihen, mitä ja millä aineistoilla tutkitaan. Alueet, joilla dataistumisen aste on alempi, myös datan käyttöä ja keräämistä rajoittava lainsäädäntö on usein löyhempi. Toisaalta vallankäyttö ja datavuotojen siirtyminen hallinnollisten elimien sijaan suuryrityksille näkyi ennen kaikkea siinä mitä ei ole kunnolla päästy tutkimaan, eli suuryrityksille jatkuvasti vuotavaa, kerättyä ja varastoitua dataa.

Aineistoni artikkelien julkaisuvuosina (2010–2022 kesäkuu) eurooppalaisille verkon käyttäjille näkyvin muutos datan keräämistä koskevaan lainsäädäntöön tapahtui vuonna 2011, kun EU:n direktiivi evästeiden käytön hyväksymisestä tuli voimaan. Nykylainsäädännön mukaan Euroopan unionin kansalaisille suunnatuilla verkkosivuilla tulee sivuston käyttäjältä saada lupa evästeisiin, joiden avulla seurataan käyttötutkimuksia mainontaa, analysointia ja markkinatutkimuksia varten (Your Europe 2023). Ainakin nuorille verkonkäyttäjille ”hyväksyn evästeet”-painikkeen klikkaaminen sen kummemmin tutustumatta hyväksyttäviin asioihin on jopa ”kulttuurillinen normi” (Obar & Oeldorf-Hirsch 2018, 142). Toisaalta jo tehdyistä tutkimuksista on käynyt ilmi, että etenkin nuoret tiedostavat yksityisyyteen ja datan keräämiseen liittyvät riskit ja sanovat olevansa niistä huolissaan, mutta huolehtiminen ei

kuitenkaan vaikuta toimintaan verkossa (#6.32; #SB2.2). Lisäksi edes nuorten aikuisten datasuojaan ja -infrastruktuuriin liittyvää tietotasoa ei ainakaan kirjallisuuskatsaukseni aineistossa tutkittu, vaikka osa kyselyistä ja haastatteluista sisälsivät alan sanastoa ja antoi osin viitteitä siihen, etteivät edes diginatiivit henkilöt ymmärtäneet mitä esimerkiksi datavuoto, evästeet tai psykografinen markkinointi tarkoittivat.

Verkon käyttäjien ja verkkopalveluiden tarjoajien välillä on valta-asetelma, joka ei anna käyttäjälle paljoa liikkumavaraa. Datayritysten tarjous on usein mallia ”ota tai jätä”, eli mikäli palveluita haluaa tarkoituksenmukaisesti käyttää, tulee niiden ehtoihin myös suostua. Vaikka EU:n lainsäädäntö edellyttää palveluntarjoajaa antamaan käyttäjille kattavat ja selkeät tiedot kerättävän datan käytöstä (Your Europe 2023) ja vastaavia säädöksiä on käytössä myös muilla alueilla, harva käyttäjä tiedostaa mihin kaikkeen antaa luvan hyväksyessään evästeet, yksityisyydensuojalausekkeen tai käyttöehdot (Maris et al. 2020, 2031).

Useissa aineistoni artikkeleissa tuotiin esille ajatus tietoisesta suostumuksesta verkkopalveluiden käyttöön koskien niistä kerättävää dataa. Erityisesti esiin nostettiin seksuaalisuuteen liittyvien sivustojen käyttö ja esimerkiksi pornosivustoilta vuotava data ja kyseisten sivustojen evästeiden avulla tehtävä profilointi (#2.6). Lisäksi tietoinen suostumus tuotiin esille myös suhteessa akateemisessa tutkimuksessa käytettävään dataan, etenkin silloin kun mahdolliset tutkimusaineistot ovat verkossa vapaasti saatavilla (#6.3). Verkon käyttäjä, oli kyseessä sitten verkkosivuston selaaminen tai akateemiseen tutkimukseen osallistuminen, on alimpana datavuototutkimuksen valtarakenteissa.

Haastatteluilla, etnografisella tarkkailulla, kyselyillä ja julkisten asiakirjojen vertailulla ei kuitenkaan päästä datavuotojen ytimeen kiinni. Voidaan tutkia, miten ihmiset kokevat oman datansa vuotamisen, mutta epäselväksi jää, mitä kaikkea julkisten ja yksityisten organisaatioiden palvelimille siirtyy. Kerätty data on datayritysten omaisuutta ja sen tutkimuskäyttöön luovuttamiseen liittyy myös yksityisyydensuojaan liittyviä, osin myös eettisiä, kysymyksiä. Euroopan Unionin alueella jokaisella on kuitenkin oikeus vaatia dataa keräämiä tahoja toimittamaan kaikki itsestä kerätty data luettavassa muodossa. Palaan tähän asetemaan autoetnografian näkökulmasta alaluvussa 5.3. Eri ihmisryhmien datavuototutkimuksen piiriin tavoittamisen osalta yksi mahdollisuus voisi kuitenkin olla kansalaisten osallistaminen tutkimukseen kansalaistieteen keinoin.

Kansalaistieteelle (engl. citizen science) ei ole tiukkaa määritelmää, mutta keskeistä siinä on niin sanotusti tavallisten kansalaisten tietoinen osallistuminen tieteellisen tutkimuksen

tekemiseen muutenkin kuin tutkimuksen kohteina. Osallistumisen aste voi vaihdella tutkimusdatan tietoisesta lahjoittamisesta tutkimussuunnitelmien tekemiseen, datan käsittelyyn ja tutkimustulosten käsittelyyn ja artikkelien kirjoittamiseen saakka. Datalahjoituksiin perustuvia tutkimuksia, joissa tutkimukseen osallistuvat henkilöt tietoisella ja autonomisella päätöksellä lahjoittavat esimerkiksi erilaisten älylaitteiden keräämää dataa itsestään, on tehty etenkin terveydenhoidon saralla. (Bietz et al. 2019, 1–2.)

5.2 Monitieteinen ja kansainvälinen yhteistyö

Aineistoni artikkeleista nousi esille jako tietojärjestelmätieteiden ja humanististen tieteiden välillä. Haasteita aiheutti se, että humanististen tieteiden piiristä ei löytynyt riittävää osaamista aineiston keräämiseen, kun taas vastaavasti tietojärjestelmätieteiden tutkimuksissa keskityttiin datan kaapimiseen ja datavuotojen estoon, eikä vuotanutta dataa pyritty analysoimaan kvalitatiivisesti. Hedelmällisimmät lähtökohdat tutkimuksille vaikuttivat olevan sellaisia, joissa tutkimusryhmän sisällä on riittävä tietotaito ratkaista ja tarvittaessa kehittää tietoteknisiä ratkaisuja aineistojen saatavuuteen ja sen jälkeen analysoida eri menetelmin saavutettua dataa. Aineiston lähteenä olevien ihmisryhmien diversiteetin lisäksi tarvitaan siis heterogeenisia tutkimusryhmiä.

Kuten tämän artikkelin luvussa 4 toin esille, kirjallisuuskatsaukseni kartoittavissa hauissa en vielä rajannut hakutuloksista mitään tieteenaloja pois, jolloin hakutuloksissa oli mukana merkittävä määrä tietotekniikkaan liittyviä julkaisuja. Toisaalta esiin tuli myös se, miten osa tutkimuksesta tapahtuu datayhtiöiden sisällä, jolloin tutkijat pääsevät käsiksi aineistoihin, joihin akateemisia väyliä käyttäen ei olisi pääsyä. Samalla tutkimustulokset voivat jäädä yritysten sisäiseen käyttöön tai tulla julkaistuksi ns. white paper -julkaisuina, joiden tarkoitus on edesauttaa taustaorganisaation tavoitteita, eikä tuottaa vertaisarvioitua, akateemista tietoa.

Kirjallisuuskatsaukseni aikana yllätyin kuitenkin monitieteisyyttä ajatellen siitä, että kauppatieteiden käyttöön tarkoitettua dataa käytettiin aineistoni artikkeleissa hyvin vähän. Datayhtiöistä useimmat kuitenkin pelaavat nykyisillä liiketoimintasuunnitelmillaan kaksilla korteilla. Toisilla järjestetään käyttäjien henkilökohtaista dataa myytäväksi muille yrityksille, ja toisilla laitetaan yritykset kilpailemaan keskenään näkyvyydestä eri alustoilla. Esimerkiksi Googlen mainostilaa kaupataan eräänlaisella huutokaupalla, jossa rahan lisäksi merkittäviä tekijöitä ovat esimerkiksi mainostilaa ostavan yrityksen mainoksen laatu ja laskeutumissivulla käytetty aika (Komulainen 2018, 160). Jokaisen datavuodoista kiinnostuneen henkilön kannattaisikin ehkä ensimmäisenä tehdä Googlemainos ja tarkastella, millä kaikilla keinoin

mainoksia voidaan kohdentaa, eli hyvin yksinkertaisella tavalla takaisin mallintaa datan kerääminen. Tällainen tutkimus olisi resurssinäkökulmasta varsin mahdollinen toteuttaa myös niin, että tutkimuskysymyksenä olisi esimerkiksi kartoittaa verkon käyttäjien asennemuutoksia kohdennettuihin sisältöihin ennen ja jälkeen mainoksen tekemisen. Yhdistämällä tällaiseen lähtökohtaan kansalaistieteen keinoja, olisi kohdehenkilöiltä mahdollista saada myös yksityiskohtaista dataa verkon selaamisesta tutkimusta edeltävältä ja seuranneelta ajalta tekemällä GDPR-keräyspyynnöt Googlelle.

Kärjistetysti datayhtiöiden toimintalogiikka on kerätä ihmisistä mahdollisimman paljon dataa, jotta mainostavilta yrityksiltä voidaan kerätä mahdollisimman paljon rahaa. Ihmiset eivät siis oikeastaan ole datayhtiöiden asiakkaita, vaan tuottajia. Datayhtiöiden hyödyntämisestä yritysten markkinoinnissa on näin ollen valtava määrä tietoa, jota ei kuitenkaan ainakaan kirjallisuuskatsaukseni artikkeleissa juurikaan hyödynnetty. Selkein intersektio humanistisen tutkimuksen ja kauppatieteiden välillä oli artikkelissa, jossa selvitettiin datavuotojen uhrien kompensoimiseen liittyviä tekijöitä (#6.19). Samalla esiin tuli kysymys yksilön datavuodon arvon kokemisesta, eli samaa tutkimusaineistoa hyödynnettiin kauppatieteissä sen selvittämiseen, miten yritysten kannattaisi toimia datavuodon tapahtuessa ja toisaalta kvalitatiivisiin pohdintoihin siitä, miten yksilöt kokevat datavuodon ja onko edes mahdollista määrittää vuodoille hintaa ihmisen (asiakkaan) näkökulmasta. [11]

Akateemisen tutkimuksen ja siihen johtavan koulutuksen kontekstissa voidaan pohtia myös sitä, miten ja missä tulevia datavuotojen hyödyntäjiä koulutetaan. Parhaillaan yliopistoissa koulutetaan tulevia vaali- ja markkinointikampanjoiden tekijöitä, joiden ammattitaito riippuu siitä, miten taitavasti dataa osataan kerätä käyttöön ja hyödyntää kaupallisissa tarkoituksissa. Samanlaisia taitoja tarvitsevat myös datavuototutkijat, mutta täysin eri syistä. Perehtymällä toimintatapoihin ja koulutukseen yli tiedekuntarajojen dataa voidaan sujuvasti käsitellä eri näkökulmista ja samalla tutkia sitä, miten eri tavoilla akateeminen yhteisö siihen suhtautuu. Henkilökohtainen data, ja sitä kautta datavuodot, ovat yhteiskunnassamme kapitalismin väline, eikä sitä näin ollen voida tutkia liiketaloudesta erillisenä kokonaisuutena.

[11] Monitieteisistä lähtökohdista henkilökohtaisen datan yrityskäyttöön luopumisen hintaa ovat kirjallisuuskatsaukseni ulkopuolella tutkineet myös mm. Acquisti et al. (2013). Tutkimuksessa yksilöiden henkilökohtaisen datan arvon kokemiseen vaikutti merkittävästi se, pitikö koehenkilön maksaa datansa suojaamisesta, vai saiko hän vastaavan summan vähemmän rahaa käyttöönsä. Koehenkilöt olivat valmiita ottamaan vastaan vähemmän hyödykkeitä (vähempiarvoisen lahjakortin) yksityisyytensä suojaamiseksi, mutta eivät olleet valmiita käyttämään lahjakortin varoja yksityisyydensuojaansa. Toinen merkittävä tekijä oli järjestys, jossa vaihtoehdot tarjottiin koehenkilöille.

Yksi aineistoni artikkeleissa hyödyntämätön ja julkinen aineisto oli myös kuluttajille suunnattu yritysten data- ja yksityisyysviestintä. Yksityisyydensuojalausekkeita tutkittiin kyllä, mutta kuluttajille suuntautuva markkinointiviestintä ”turvallisista” ja ”yksityisistä” henkilökohtaiseen dataan liittyvistä palveluista on varmasti muuttunut datavuotojen ympärillä vellovan julkisen keskustelun muuttuessa[12]. Koska julkinen viestintä vaikuttaa datavuotoja koskevaan diskurssiin, sen nykytilan ja muutoksen tutkiminen toisivat uusia näkökulmia datavuotojen yhteiskunnalliseen asemaan.

Yhteiskunnan kannalta voidaan ylipäätään pohtia sitä, missä määrin olemassa oleva tutkimus lopulta päätyy hyödyttämään datayhtiöitä itseään. Kun tutkimustulokset kertovat, että sosiaalisten medioiden ja erilaisten mobiilisovellusten käyttäjät tiedostavat datan keräämiseen liittyvät uhkat, mutta eivät välitä niistä, ja että kerätyn datan perusteella terveydestään eniten huolissaan olevat naiset ovat helpoiten ohjailtavissa, mikään ei estä datayhtiöitä käyttämästä akateemisesti tuotettua tutkimusta oikein kehystettynä omassa markkinoinnissaan.

Toisaalta eri maita ja kulttuureita verranneissa tutkimuksissa todettiin tulosten osalta isoja alueellisia eroja. Esimerkiksi verrattaessa Iso-Britanniaa, Yhdysvaltoja ja Arabiemiraatteja työntekijöiden asenteet omien matkapuhelinten käytön aiheuttamista datavuotoriskeistä vaihtelivat maittain kytköksissä alueellisten lainsäädäntöjen kanssa, mutta merkitystä oli myös työympäristöllä ja esihenkilötyöllä (#6.23). Vastaavia kansainvälisiä, länsimaita ja muita alueita vertaavia, vertailuja (lakitekstejä lukuun ottamatta) ei aineistossani ollut, joten tulevaisuudessa monikansalliselle tutkimukselle on varmasti vielä kysyntää esimerkiksi datavuotoihin liittyvien tunteiden (vrt. esim. #SB2.2, jossa tutkittiin datan käyttöön suhtautumista) kartoittamisessa. Tiivis kansainvälinen yhteistyö tutkimusryhmissä toisi myös tutkimuskohteet lähemmäksi tutkijoita ja auttaisi pienentämään kulttuurieroihin pohjautuvaa ja virheellisille tulkinnoille altistavaa kuilua tutkijoiden ja aineiston välillä (Taylor 2016, 320).

Datasuojaa ja datan keräämisen rajoittamista koskevan säädännön ja datavuotoihin suhtautumisen välistä korrelaatiota sivuttiin joissain artikkeleissa, mutta spesifisti valtiollisen rajoittamisen ja datavuotojen suhdetta ei tutkimuskysymysasteella tutkittu. Epäselväksi jäi

[12] Esimerkiksi Telian datakeskusmainoksessa kirjailija Tuomas Kyrön äänellä rinnastetaan henkilökohtainen data ja peruna toteamalla, että ”Tieto on raaka-aine niin kuin öljy, vilja tai vesi. Sitä louhitaan, jaetaan, tutkitaan ja jalostetaan, annetaan, otetaan, tuodaan ja viedään. Säilötään ja siirretään. […] Tarvitsemme omavaraisuutta ja varmuusvarastoja. Kellarissa täytyy olla perunoita ja datakeskuksen on hyvä sijaita omassa maassa”. Mainosvideo katsottavissa https://www.youtube.com/watch?v=4zLkEKhEgSI (linkki tarkistettu 29.5.2023).

esimerkiksi kansalaisiin kohdistuvan valvonnan ja datavuotojen suhde. Tässäkin tutkimusasetelmassa ongelmalliseksi tullee aineiston saatavuus, koska valtiollista valvontaa koskevat asiakirjat ovat pitkälti salattuja. Lähimmäksi tällaista näkökulmaa aineistossani pääsi artikkeli, jossa kartoitettiin kyselytutkimuksen avulla politiikkaan verkossa osallistumisen vaikutusta yksityisyyden suojaamiseen liittyvään käytökseen eri Aasian ja Lähi-idän maissa (#7.2). Toki tässäkin yhteydessä Kiina, joka korkean sääntelyn ja kansallisten rajoitusten maana olisi ollut erittäin kiinnostava tutkimusalue, jäi vertailun ulkopuolelle sen poliittisen tilanteen vuoksi.

5.3 Autoetnografia potentiaalisena lähestymistapana

Etymologisesti etnografialla tarkoitetaan kansojen kuvaamista tai niistä kirjoittamista, eli etnografian keinoin käsitellään ihmisiä kollektiivisesti, ei yksilöinä (Angrosino 2007, 2). Etnografian juuret ovat antropologian vieraiden kulttuurien tutkimuksessa, jossa vakiintui 1800- ja 1900-lukujen vaihteessa ajatus siitä, että pystyäkseen paljastamaan sosiaalisia ja kulttuurillisia rakenteita tutkijalle vieraasta kulttuurista tutkijan tulee itse päästä tarkkailemaan tutkimuskohteena olevia ihmisiä kulttuurin sisältäpäin (Seale 2012, 246; Angrosino 2007, 2). Euroopassa etnografian juuret ovat brittiläisen imperiumin kolonialistisen vallan alla olleiden alueiden tutkimuksessa, ja vastaavasti Amerikassa ensimmäiset tutkimuskohteet olivat jo suurilta osin tuhottujen alkuperäiskansojen kulttuurit (Angrosino 2007, 2). Nykynäkökulmasta etnografian ongelmana voidaan pitää asetelmaa, jossa tutkija penetroituu itselleen vieraaseen ”eksoottiseen” kulttuuriin ja tuottaa kokemuksistaan tutkimusta, jonka narratiivi on väistämättä kolonialistiseen perinteeseen nojaava (Lapadat 2017, 591).

Metodin syntyhistorian ja tutkijaposition sisäänrakennetun hierarkkisen ja ulkopuolisen, toiseuttavan, aseman myötä keskeiset haasteet etnografian käytössä liittyvät tutkimusetiikkaan ja toisaalta tutkimuksen uskottavuuteen ja toistettavuuteen (Berger 2016, 232; Hammersley 2018, 2). Lisäksi etnografiasta kiinnostuneen tutkijan kannalta haasteena on myös muuttunut, nopeatempoisempaan tutkimukseen ja julkaisemiseen kannustava, akateeminen ympäristö (Hammersley 2018, 2), jonka puitteissa pitkiä ajanjaksoja kenttätyötä vaativan tutkimustavan toteuttaminen on haasteellista.

Nykykäsityksen mukaan etnografian keinot soveltuvat parhaiten arkipäiväisten, usein toistuvien, toimintojen tutkimukseen. Siinä missä vielä 1960- ja 1970-luvuilla jokapäiväiset toiminnot kytkettiin fyysiseen toimintaan yksilöinä ja ihmisten välillä, nykypäivänä teknologia ja dataistuminen ovat lomittuneet ihmiselämään niin syvälle, että niitä ei voida

sulkea ihmisen toiminnan ulkopuolelle omaksi kokonaisuudekseen (Couldry & Hepp 2016, 19). Toisaalta kyse on edelleen toimintojen tarkastelusta. Sohvalla selällään makaaminen ei vielä juurikaan sisällä toimintaa (paitsi lääketieteellisestä näkökulmasta), mutta jos sohvalla maatessaan selaa etenkin verkkoon kytkettyä älypuhelinta, jokainen sekunti sisältää valtavan määrän toimintaa. Datatutkimuksen näkökulmasta ajateltuna tällainen toiminta on sekä tiedostettua (lähetetyt viestit, selatut sivut, sovelluksissa käytetty aika), että tiedostamatonta (iso osa datavuodoista).

Kuten aiemmin tässä tutkielmassa olen tuonut ilmi, etnografista tutkimusta datavuodoista on tehty vähän, vaikka ihmisten kokemusta oman itsensä mittaamisesta, mitattavana olemisen kokemisesta ja mittaamisen toimintalogiikoista palveluja ja teknologioita tarjoavien yritysten osalta on tutkittu varsin laajasti (self-tracking-tutkimukset). Etnografisia menetelmiä käyttäneissä tutkimuksissa nousi kuitenkin esille tarve lisätutkimukselle. Esimerkiksi kyselyn avulla todettiin koeryhmän jäsenten kertovan datan tallentamisen ja käyttöön vaadittavien suostumusten vaikuttavan ostopäätöksiin sovelluksia ladatessa, mutta käytännön etnografisessa asetelmassa samat henkilöt latasivat sovelluksia välittämättä vaadituista datan käyttöön ja tallentamiseen liittyvistä suostumuksista (#6.32). Etnografisia metodeja hyödynnettiin myös salasanatutkimuksissa, joista osassa koeryhmän henkilöitä pyydettiin itse keksimään salasanoja. Koeryhmissä itse keksittyjen salasanojen laatu oli huomattavasti parempi kuin vuodetusta datasta tehdyistä salasana-analyyseissä (#SB2.1). Etnografisella seurantatutkimuksella tulokset haastatteluiden, kyselyiden ja dokumenttien analysoinnin kanssa olivat lähes päinvastaisia.

Verkkovälitteisen kommunikaation yleistyttyä etnografia on saanut uusia muotoja, kun tutkijat ovat siirtyneet tarkkailemaan verkkoyhteisöjä ja yksilöiden toimintaa niissä. Verkkoetnografian keinoja datatutkimuksen osalta ei ole vielä täysin hyödynnetty. Toisaalta verkkovälitteisen etnografiaan (*digital ethnography, cyber-ethnography, virtual ethnography, net-nography*) (Silverman toim. 2020, 117) liittyy myös metodologisia ongelmia: jos tutkimus rajoittuu verkossa käydyn keskustelun analyysiin, onko kyseessä tekstianalyysi vai etnografinen tutkimus? Datavirrat ja -vuodot ovat osa jokapäiväistä elämää ainakin kaikilla puhelinta verkkoyhteyden välityksellä käyttävillä, tai muilla välineillä verkkoa selaavilla, ihmisillä, joten etnografian käyttö tutkimusmetodina datatutkimuksessa on perusteltavissa. Samalla kaventuu myös etnografisen tutkimuksen toiseuttava asetelma; kaikki Facebookin käyttäjät selaavat sitä samankaltaisessa verkkonäkymässä, ympäröivästä kulttuurista ja yhteiskunnasta riippumatta.

Vielä etnografista lähestymistapaa vähemmän datavuototutkimuksessa on käytetty autoetnografisia aineistoja ja tutkimusotetta. Autoetnografisessa tutkimusotteessa tutkija yhdistää oman henkilökohtaisen kokemuksensa kulttuurillisen kokemuksen tutkimiseen. Tutkija positioi itsensä sekä tutkimuksen kohteeksi (subjektiksi) että tutkijaksi, jolloin eettisestä näkökulmasta tarkasteltuna pystytään välttämään kritiikki etnografian toiseuttavasta lähestymisestä tutkimuskohteeseen. Autoetnografi ei puhu muiden puolesta, vaan keskittyy omien kokemustensa kautta tuomaan esiin piilossa olevia nyansseja kulttuurillisesta kokemisesta (Ellis & Adams 2020, 368). Toisaalta, juuri tutkija-aseman subjektiivisuuden aiheuttama etäisyyden puute tutkittavaan aiheeseen ja henkilökohtaisen kokemuksen linkittäminen osaksi sosiokulttuurillisia ilmiöitä syövät autoetnografisen tutkimuksen uskottavuutta (Lapadat 2017, 589–591).

Tutkijan oman kokemuksen yhdistämisen määrä tutkimuskohteensa tarkasteluun vaihtelee. Keskeistä on kuitenkin, että yhden ihmisen (tutkijan) elämä voi tarjota yleistettävää tietoa ihmisen kokemuksesta (Ellis & Adams 2020, 360). Osa tutkijoista huomauttaakin kaiken etnografisen tutkimuksen sisältävän myös autoetnografisia elementtejä (Ellis & Adams 2020, 360). Keskeistä on kuitenkin, että kulttuuri-ilmiöitä tarkastellaan itsekriittisestä ja sosiaalisesti tietoisesta näkökulmasta, jolloin tutkimus ei rajoitu pelkästään kirjoittajan oman kokemisen kuvailuun (Holman Jones et al. 2016, 23). Tämä erottaa autoetnografisen lähestymistavan autobiografisesta kirjoittamisesta; proosallisen kerronnan sijaan tutkijat pyrkivät selittämään tunnetason ilmiöitä tieteen keinoin, samalla pyrkien kohti yhteiskunnallista muutosta. Vaikka autoetnografiassa kuvaillaan yksilön eli tutkijan kokemusta, sen fokus on havaintojen tulkinnassa, vertailussa ja (itse)reflektiossa sekä niiden suhteessa kulttuuriin, politiikkaan ja yhteiskunnan valtasuhteisiin (Ellis & Adams 2020, 364).

Historiallisesti autoetnografian yleistyminen juuri 2000-luvulla voidaan kytkeä yleiseen kulttuurin muutokseen (Ellis & Adams 2020, 364) ja kommunikaatiovälineiden demokratisoitumiseen erityisesti internetin myötä (Lapadat 2017, 592). Muita 2000-luvulla esiin nousseita ilmiöitä ovat esimerkiksi tositelevisio ja yksilölähtöinen verkkosisältö (blogit ja hieman myöhemmin sosiaalinen media). Individualismi ja yksilökeskeisyys, sekä toisaalta oman itsen esiintuominen ja tiedon saatavuus eri medioiden välityksellä tapahtuivat samanaikaisesti autoetnografian yleistymisessä kvalitatiivisessa tutkimuksessa. Samanaikaisesti syntyi myös kohdennettu markkinointi, eli tietojen kerääminen yksityishenkilöistä heille mieleisen ja kohdennetun mainossisällön luomiseksi. Autoetnografit pyrkivät kuitenkin erottamaan nämä rinnakkaiset ilmiöt toisistaan (Ellis & Adams 2020, 365)

ja korostamaan autoetnografian kehittymistä osana kvalitatiivisen tutkimuksen historiaa ja vastauksena etnografian tutkimuseettisiin haasteisiin.

Datavirtojen tutkimukseen autoetnografia tarjoaa mielenkiintoisia mahdollisuuksia. Ensinnäkin sen käyttö tarjoaa ainakin näennäisesti yksinkertaisen ratkaisun yksilödatan aineiston saavutettavuuteen, koska tutkijalla on, ainakin EU:n alueella ja tietyin rajoituksin myös Englannissa, oikeus saada omaa itseään koskeva data käyttöönsä. Samalla ohitetaan tutkimuseettinen haaste muita koskevan yksilöidyn datan käytöstä tutkimuksessa. Lisäksi lainsäädännöllisestä näkökulmasta tarkasteltuna yksityisyydensuojalainpiiriin kuuluvan informaation käyttö on tutkimuskohteen, eli tutkijan, omissa käsissä. Autoetnografista dataa on siis mahdollista saada ja käyttää varsin yksinkertaisen prosessin avulla, varsinkin niin kauan, kun kyseessä on GDPR:n alainen data.

Tutkijan omaa henkilökohtaista dataa hyödyntävässä tutkimuksessa datavirtoja pystytään tarkastelemaan myös laadullisesta näkökulmasta, datavirtojen laajempien kokonaisuuksien analyysin sijaan. Edes niissä kirjallisuuskatsaukseni artikkeleissa, joissa pyrittiin laadullisesti tarkastelemaan kerättyä dataa, ei pystytty arvioimaan sitä, miltä osin kerätty data oli paikkansapitävää. Esimerkiksi pornosivuilta vuotanutta dataa koskevassa tutkimuksessa selvisi vuotojen suuri määrä ja niiden aiheuttama uhka erityisesti marginalisoiduille ihmisryhmille (#2.6). Tutkimuksessa erottuivat erityisesti sellaiset verkkosivujen osoitteet, jotka sisälsivät fetisseihin tai seksuaalivähemmistöihin viittaavaa sanastoa. Tutkimusasetelmalla ei kuitenkaan pystytty kartoittamaan sitä, missä määrin kolmansien osapuolten keräämä data pitää paikkaansa, eli olivatko sivustoja selanneet henkilöt aidosti kiinnostuneita verkko-osoitteen nimeämästä sisällöstä, kuten data ikään kuin olettaa, vai oltiinko sivustoilla jostain muusta syystä.

Autoetnografisen lähestymistavan käyttöä puoltaa myös sen soveltuvuus arkielämän ”näkymättömäksi muuttuneiden”, usein toistuvien, käytänteiden tutkimukseen ja sen luonteeseen yhdistää henkilökohtaisuus sosiaaliseen ja kulttuurilliseen ympäristöön (Uotinen 2010, 163). Ihmisistä kerättävää dataa ja sen hyödyntämistä kutsutaan myös uudeksi kolonialismin muodoksi, jossa riistettävä resurssi on maa-alueiden ja raaka-aineiden sijaan ihmisistä kerättävä data, ja riistävä taho suuryritykset, jotka tekevät datasta rahaa (Couldry & Mejias 2019, xi). Tästä näkökulmasta voisi ajatella tutkimushistoriallisesti kolonialismikritiikistä syntyneen tutkimuksellisen lähestymistavan sopivan myös uuden kolonialismin, eli datavuotojen tutkimiseen. Toisaalta tutkijan oman henkilökohtaisen datan

käyttäminen datavuotoja koskevassa tutkimuksessa ei poista kysymystä autoetnografisen aineiston soveltuvuudesta kuvaamaan kulttuurillista ilmiötä pelkän yksilökokemuksen sijaan. Lisäksi lähestymistapa ei lisää datavuototutkimuksen tasa-arvoa ja inklusiivisuutta aineiston tullessa akateemisesti koulutetulta datavuototutkijalta.

Autoetnografisen lähestymistavan sisällä on myös eroja henkilökohtaiseen aineistoon suhtautumisessa. Evokatiivisessa autoetnografiassa keskiössä ovat kirjoittajan henkilökohtaiset tarinat ja niiden tunnepitoinen kuvailu, kun taas analyyttinen autoetnografia kytkee henkilökohtaiset kokemukset osaksi valittua kulttuuri-ilmiötä (Denshire 2014, 833–835). Evokatiivinen autoetnografia on tarinallisempaa, yhteiskunta kuvataan yhden ihmisen kautta ja keskiössä ovat usein rankat emotionaaliset elämänmuutokset, kuten läheisen kuolema, seksuaalinen väkivalta tai rasismi, joiden avulla yhteiskunnallisia ilmiöitä kuvataan yksilön läpi. Vastaavasti analyyttinen autoetnografia pyrkii teoreettisempaan ja objektiivisempaan lähestymiseen (Denshire 2014, 835), jossa yksilön kokemusta käytetään esimerkkeinä laajemmasta ilmiöstä. Autoetnografien keskuudessa keskustelu evokatiivisen ja analyyttisen lähestymisen spektristä on osin polarisoitunut ja viime vuosina on kiistelty esimerkiksi siitä, hävittääkö analyyttisyyteen pyrkivä lähestyminen autoetnografian emotionaalisen ja tarinankerronnallisen ytimen (Lapadat 2017, 594–595).

Erityisesti evokatiivisen autoetnografian kannattajien keskuudessa nousisi varmasti esille keskustelu siitä, onko esimerkiksi datayritysten keräämän raakadatan käyttäminen edes autoetnografista aineistoa, mikäli siihen ei yhdistetä tutkijan kokemukseen sidottua narratiivia. Yritysten keräämä data on tutkijasta peräisin, mutta tutkijan asema datan keräämisessä on olla pelkästään lähde, josta tiedot kaavitaan. Henkilökohtainen data on henkilökohtaista, mutta sen keräämisen prosessiin tutkija ei juurikaan ole osallistunut. Toisaalta datayrityksien algoritmien toimintaan ei sisälly inhimillisiä virheitä eikä henkilön persoonasta johtuvaa vaihtelua, joten data itsessään on äärimmäisen objektiivista.

Autoetnografista lähestymistapaa hyödyntävä tutkija joutuisi myös väistämättä pohtimaan, missä määrin hän haluaa avata itseään akateemiseen tutkimukseen. Autoetnografiassa tutkija hyödyntää aina kahta identiteettiä itsestään; yksityistä minäänsä eli kokijaa aineiston tuottamiseen, ja akateemista minäänsä sen analysoimiseen (Ellis & Adams 2020, 368). Autoetnografinen tutkija on näin aina tutkimuksensa keskiössä, eikä tutkimuksen tekijää voida häivyttää aineiston tai tutkimustulosten taakse. Samalla häivytetään epätasa-arvoinen valtasuhde tutkijan ja tutkittavan henkilön välillä (Lapadat 2017, 593) ja kiistetään käsitys

siitä, että akateemisen tutkimuksen tulisi erotella tutkija ja tutkimus toisistaan (Campbell 2017, 13). Asetelmaan kytkeytyy uniikin tutkijaposition vuoksi väistämättä henkilökohtaisia ja eettisiä riskejä ja haasteita (Holman Jones et al. 2016, 19; Denshire 2014, 832). Tutkimuksessa pystytään hyödyntämään tietoa, johon ei välttämättä muilla keinoin päästäisi käsiksi, mutta samalla tutkija altistaa itsensä emotionaaliselle stressille ja epävarmuudelle (Ellis & Adams 2020, 359). Positioimalla oman elämänsä osaksi akateemista tutkimusta tutkija altistuu myös kritiikille siitä, millaisia valintoja omassa elämässään on tehnyt ja miksi niitä on käytettävä omassa tutkimuksessaan (Holman Jones et al. 2016, 24; Campbell 2017, 7). Anonymiteetin puuttuminen koskee myös esimerkiksi tutkijan läheisiä ihmisiä, mikäli tutkimuksen aineisto sisältää kuvausta esimerkiksi tutkijan perheenjäsenistä (Lapadat 2017, 593). Lisäksi tutkimuksen toistettavuuteen liittyvien kysymysten osalta aineiston luovuttaminen kokonaisuudessaan muille osapuolille ei ole yksinkertaista.

Historiallisesti autoetnografian kehittymistä on edesauttanut vastaliike lääketieteen ja psykologian klassikkotutkimuksille, joissa tutkittavia ihmisryhmiä kohdeltiin epäeettisesti (Ellis & Adams 2020, 366). Yksinkertaistettuna autoetnografisen tutkimuksen kohteilla ei ole mahdollisuutta tulla kohdelluksi kaltoin, koska tutkimuksen kohde on tutkija itse. Tosin voidaan pohtia sitä, missä määrin autoetnografit kokevat painostusta julkaista hyvinkin yksityisiä tietoja itsestään esimerkiksi tutkimuksen tärkeyttä painottaakseen. Viime kädessä päätös ja vastuu autoetnografiseen tutkimukseen suostumisesta ja aineiston käytöstä on tutkijalla itsellään, jonka voidaan olettaa olevan täysi-ikäinen ja päätösvaltainen henkilö.

Oman kysymyksensä muodostaa kuitenkin autoetnografisen aineiston tarkoituksellinen tuottaminen. Etenkin pitkään autoetnografista tutkimusta tehneet henkilöt puhuvat ”autoetnografisesta elämisestä ja tekemisestä” (mm. Jones et al. 2016, 21). Ajatus elämästä autoetnografisen aineiston tuotannon välineenä herättää kysymyksiä siitä, missä määrin omaa elämää muokataan tutkimukseen sopivaksi, ja miten aineistoon vaikuttaa se, että kaikki kokeminen on potentiaalista tutkimusmateriaalia, ja sellaisena se myös kirjataan ylös. Datavuotojen näkökulmasta: jos tutkija valmiiksi tiedostaa kaiken verkossa tapahtuvan liikenteensä potentiaalisesti päätyvän osaksi datavuototutkimusta, voiko tutkijan verkkokäyttäytyminen muuttua tietoisesti tai tiedostamatta?

Puhtaan autoetnografisen lähestymisen sijaan yksi uusi ja aikaisemmin hyödyntämätön näkökulma datavuototutkimuksen aineistoihin liittyviin haasteisiin voisi löytyä myös

yhteistoiminnallisesta autoetnografiasta[13]. Yhteistoiminnallisessa autoetnografiassa (engl. collaborative autoethnography, joskus myös duoethnography, collective autoethnography) ryhmä tutkijoita yhdistää autoetnografisen tutkimuksensa etsien eroja ja yhtäläisyyksiä, analysoiden ja tulkiten omia ja muiden tekstejä yhdessä ja erikseen (Lapadat 2017, 598). Yhteistoiminnallinen autoetnografia pyrkii säilyttämään autoetnografian henkilökohtaisen datan, hyödyntäen ryhmän suomia synergiaetuja aineiston analyysissa ja ymmärtämisessä (Chang et al. 2016, 17; 20–24). Sen syntyyn on vaikuttanut tarve evokatiivista autoetnografiaa analyyttisemmasta lähestymistavasta autoetnografiseen aineistoon. Yhden tutkijan ”ikkuna yhteiskuntaan” laajenee ja tutkittavaan aiheeseen saadaan useampi kuin yhden henkilön ääni.

Yhteistoiminnallisella autoetnografialla pystytään myös osittain vastaamaan autoetnografisen lähestymistavan tutkijan yksityisyyteen liittyviin eettisiin haasteisiin, koska tutkimustuloksia pystytään häivyttämään tutkimusryhmän sisäisiksi, profiloimatta niitä yhteen henkilöön. Autoetnografian alalajin haasteina ovat kuitenkin esimerkiksi tutkimusryhmän sisäinen toiminta ja luottamus, riippuvuus muiden tutkijoiden elämäntilanteista ja tutkimustulosten mukautuminen ryhmädynamiikan mukaisiksi (Chang et al. 2016, 30–36). Lisäksi akateemisen tutkimusryhmän datavuotojen yhdistäminen ei pysty luomaan kovin tasa-arvoista ja inklusiivista tutkimusdataa kaikkien osallistujien ollessa koulutukseltaan ja yhteiskuntaluokaltaan homogeenisiä. Toisaalta yhtenä sivupolkuna tätä tutkielmaa tehdessä tuli ilmi Google Scholar -tietokannan ongelmallisuus datavuototutkimuksessa. Tutkijoista vuotava data ja se, miten Google sitä hyödyntää esimerkiksi akateemisten tutkimusartikkelien hakemistossaan, vaikuttaisi olevan vielä tutkimatonta aluetta.

Autoetnografiaa on kritisoitu ennen kaikkea sen objektiivisuuden puutteesta suhteessa tutkittavaan aineistoon sen ollessa tutkijan itsensä tuottamaa. Vastauksena kritiikkiin tutkijan ja subjektin olemattomasta etäisyydestä (Lapadat 2017, 589) autoetnografit ovat tuoneet esille muiden tutkimuslähestymistapojen tutkimuskohteita toiseuttavia metodeja ja virheellisiä presentaatioita (Ellis & Adams 2020, 365), jotka johtuvat ennen kaikkea siitä, että tutkija ei pysty samastumaan tutkimuskohteeseensa. Nämä ongelmat olivat vahvasti läsnä myös niissä kirjallisuuskatsaukseni artikkeleissa, joissa länsimaiset tutkijat analysoivat itselleen vieraiden alueiden ja kulttuurien datavuotoaineistoja.

[13] Osassa suomenkielisiä tutkimuksia on käytetty myös termiä kollaboratiivinen autoetnografia (kts. esim. Hietamäki, 2021), itse suomentaisin termin yhteistoiminnalliseksi, joka sisältää ajatuksen autoetnografian eri muodoista ja samalla välttää työ-sanan mahdollisesti negatiiviset konnotaatiot.

Useissa aiemmissa datavuototutkimuksissa tuotiin esille käyttäjien valehtelu verkossa. Jos internetin yleisin valhe todella on ”olen lukenut käyttöehdot” (#7.20), niin meistä lähes jokainen on kamala valehtelija verkossa. Aineistoni artikkeleissa tätä ei kuitenkaan tarkasteltu lähemmin muuta kuin artikkelissa, jossa kartoitettiin väärien tietojen antajien persoonallisuustekijöitä (#6.14). Yhteistoiminnallisella autoetnografialla päästäisiin käsiksi myös niihin tietoihin, mitä muita valheita ihmiset verkossa kertovat ja miten se vaikuttaa heistä kerättyyn dataan. Näkevätkö datayhtiöt valheidemme lävitse ja todella tuntevat meidät paremmin kuin me itse?

5.4 Tiedon vastaanoton ja datavuotojen kokemisen tutkimus

Kyselyitä ja tutkimuskohteiden seurantaa yhdistäneissä tutkimusartikkeleissa toistui asetelma, jossa tutkimuskohteet tiedostavat datavuotoihin liittyvät ongelmat, mutta siitä huolimatta tieto ei muuta käyttäytymistä tai vaikuta verkossa toimimiseen. Haastattelututkimuksessa yliopisto-opiskelijat ja yliopiston henkilökunta olivat pääasiassa tietoisia muutama kuukausi ennen haastatteluja julkisuuteen nousseesta Cambridge Analytica -skandaalista, mutta se ei ollut vaikuttanut heidän käyttäytymiseensä verkossa (#SB2.2). Sen sijaan että he olisivat poistaneet sosiaalisen median tunnuksiaan, muuttaneet yksityisyysasetuksiaan tai vaikuttaneet edes huolestuneilta, haastateltavat kokivat olevansa itse immuuneja psykografisesti kohdennetulle mainonnalle. Toisaalta he eivät myöskään täysin ymmärtäneet miten algoritmit ja kohdennettu markkinointi toimivat.

Aineistoni artikkeleissa ei myöskään valitettavasti ollut yhtäkään länsimaiden ulkopuolella tehtyä tutkimusta, jossa olisi yhdistetty kysely ja etnografinen seuranta. Indonesiassa tehdyssä kyselyssä tietoisuus verkkopalveluiden yksityisyyteen liittyvistä ongelmista lisäsi yksityisyyttä suojaavien toimintojen käyttämistä (#2.8) ja kymmenessä eri Aasian maassa[14] toteutetun kyselyn perusteella verkossa aktiivisesti politiikkaan kantaa ottavat ihmiset kertoivat kiinnittävänsä enemmän huomiota omaan yksityisyydensuojaansa ja pelkäävänsä mahdollisia datavuotoja (#7.2).

Tulokset tutkimuskatsaukseni artikkeleiden perusteella kuitenkin näyttäisivät osoittavan, että ihmiset eivät juuri välitä datavuotojen mahdollisuudesta. Aineistosta puuttuivat tutkimukset sellaisten henkilöiden kanssa, jotka ovat kokeneet tosielämässä itselleen merkittävän,

[14] Kiina ei ollut mukana kyselyssä, koska kyselydatan tuottaminen poliittisen osallistumisen osalta osoittautui mahdottomaksi (Yu & Shen 2021, 12).

julkisesti verkkoon saatavilla olevaksi päätyneen, datavuodon. Tapaustutkimukset, joissa perehdyttäisiin niin sanotusti tavallisten ihmisten sosiaalisen median tilien kaappaamisen tai tietomurtojen yhteydessä verkkoon julkaistun sensitiivisen materiaalin vaikutuksiin olisivat tarpeellisia. Tällaisia aineistoja olisi tutkimuseettisesti kestävästi saatavilla ainakin kansalaistieteen tai yhteistoiminnallisen autoetnografian keinoin, koska omien julkisten datavuotojen kartoittaminen on mahdollista useiden verkkosivujen kautta[15]. Lisäksi samoja tutkimuksellisia lähestymistapoja hyväksikäyttäen voitaisiin tutkia tilanteita, joissa tietomurron seurauksena henkilön sosiaalisen median tilejä kaapataan ja pääsy niihin estetään käyttäjältä itseltään. Aiemmin tässä tutkielmassa käsitellyssä artikkelissa edes päiväkirjamaisia kirjauksia sisältänyt datavuoto oppimisympäristön sisällä ei aiheuttanut tutkimusryhmänä toimineiden opiskelijoiden keskuudessa niin paljoa negatiivisia tunteita, että yksikään opiskelija olisi halunnut keskeyttää tutkimukseen osallistumisen (#6.3). Opiskelijayhteisön sisällä omalle vertaisryhmälle vuotanut data ei kuitenkaan ole verrattavissa julkisesti saatavilla olevaan datavuotoon, varsinkaan kun kyseinen aineisto oli oppimispäiväkirjamaista tekstiä, joka oli tuotettu kurssin suorittamista varten.

Erityyppisiin datavuotoihin ja niiden kokemiseen liittyen henkilökohtaisesta datasta voidaan erottaa arkaluontoinen data (engl. sensitive data, Linnet 2016, 328; Breuer et al. 2020, 2072). Arkaluonteiseen dataan voidaan katsoa kuuluvan esimerkiksi data seksuaalisesta suuntautumisesta tai preferensseistä (Maris et al. 2020, 2032) ja terveydestä (Barth et al. 2019, 57). Etelä-Koreassa arkaluontoiseen dataan lasketaan edellisten lisäksi tiedot ideologiasta, uskonnosta ja ammattiliittoon tai poliittiseen puolueeseen kuulumisesta (Greenleaf & Park 2014, 498). Kirjallisuuskatsaukseni aineiston perusteella vaikuttaisi myös siltä, että verkon käyttäjät automaattisesti jaottelevat omat tietonsa arkaluonteisiin ja ”normaaleihin”, ainakin käyttämällä voimakkaampia salasanoja sivustoilla, joiden kokevat säilyttävän arkaluonteista dataa (#6.32, 61). Hypoteettisesti datavuotojen kokeminen riippuu siitä, minkälaista dataa vuotaa. Datan jaottelu erityyppisiin datoihin lähtökohtana sen merkittävyys alkuperälleen voisi toimia myös koko datan keräämiseen liittyvän infrastruktuurin ohjaavana tekijänä tulevaisuudessa.

---

[15] Tällaisia sivustoja on esimerkiksi Potsdamin yliopiston ylläpitämä https://sec.hpi.uni-potsdam.de/ilc/search?lang=en (linkki tarkistettu 29.5.2023), jonka kautta voi tarkistaa sähköpostiosoitteen esiintymisen julkisiksi päätyneissä tietovuodoissa. Palvelu listaa tietovuodoissa esiintyneen sähköpostin lisäksi siihen yhdistetyn salasanan, nimen, syntymäajan, osoitteen, puhelinnumeron, maksukorttitiedot, sosiaalisturvatunnuksen ja IP-osoitteen. Testatessani palvelua itse merkittävin tietovuotoni oli tapahtunut vuonna 2013 Neopets virtuaalilemmikkisivustolla, josta oli vuotanut sähköpostiosoitteeni lisäksi tiedot syntymäajastani, nimestäni ja IP-osoitteestani.

6 Lopuksi

Datavuototutkimukseen soveltuvien aineistojen saatavuuden osalta datavuototutkijoiden kädet ovat sidotut monilta eri suunnilta. Datayhtiöiltä aineistoja on saatavilla huonosti tai ei lainkaan, kansallinen lainsäädäntö ja kansainväliset sopimukset rajoittavat yksityisten tietojen käyttöä ja tutkimuseettiset kysymykset tekevät sitä entisestään. Tiettyjä aineistoja olisi saatavilla täysin julkisesti verkossa, mutta ne ovat usein rikollista alkuperää eli peräisin tietomurroista, joten niidenkään käyttö ei ole eettisesti kestävää. Aineistoja voisi myös periaatteessa kerätä löyhemmän datalainsäädännön alueilta, mutta niissä ongelmaksi muodostuu toiseuttava tutkimusasetelma ja tutkimuseettiset kysymykset.

Yksittäisen tutkijan pääsy suoraan verkon kautta eettisesti kestäviin ja akateemiseen tutkimukseen sopiviin aineistoihin on kirjallisuuskatsaukseni perusteella todella haastavaa. Menetelmällisesti ongelmia aiheuttaa kyselyihin ja haastatteluihin vastanneiden ihmisten taipumus kaunistella omaa verkkokäyttäytymistään ja huoltaan omista tiedoistaan. Samoin haavoittuvien ihmisryhmien tavoittaminen ja tutkimusten tekeminen heterogeenisella otannalla on harvinaista. Tutkimuksen ympärille tarvitaan monitieteinen joukko eri alojen osaajia, joiden avulla voidaan kehittää erilaisia datakeräysjärjestelmiä ja huomioida lainsäädäntöön ja etiikkaan liittyvät erityisseikat. Täysin toivotonta datavuotojen tutkimus ei kuitenkaan ole. Innovaatioita kuitenkin tarvitaan enenevissä määrin, ja erilaisten dataa keräävien teknologisten ratkaisujen lisäksi voidaan perehtyä erityyppisten julkisten aineistojen yhdistelemiseen ja toisaalta niissä tapahtuneisiin muutoksiin historiallisten trendien kuvaamiseksi.

Toisaalta monitieteisissä tutkimuksissa on toistaiseksi ohitettu kauppatieteiden yhdistäminen datavuototutkimukseen. Miten datayhtiöihin alisteisessa asemassa toimivat yritykset hyödyntävät yksilöiden dataa? Tai, kun datavuototutkimuksessakin on havaittu yliopisto-opiskelijoiden olevan tutkimukseen helpoiten saatavilla oleva ihmisryhmä, miten tulevia datankerääjiä ja -hyödyntäjiä koulutetaan verkkomarkkinoinnin kursseilla verrattuna vaikkapa mediatutkimuksen kriittiseen datatutkimukseen suuntautuviin kursseihin?

Kirjallisuuskatsaukseni ulkopuolelle jäivät myös tutkimuskysymykset datavuototutkimuksen ja siihen liittyvän julkisenkeskustelun muutoksesta. Esimerkiksi diskurssianalyysi datavuotoihin liittyvästä julkisesta keskustelusta tai datayhtiöiden julkilausumista olisi toteutettavissa julkisilla ja helposti saatavilla olevilla aineistoilla. Tällaisten aineistojen avulla

ei saada vastauksia siihen mitä ja millaista dataa yksittäisestä ihmisestä vuotaa verkkoon, mutta yhteiskunnallisen keskustelun ja samalla asenteiden muutoksen kuvaaminen datavuotojen osalta olisi mielestäni tärkeä tehdä. Akateemisessa tutkimuksessa datavuoto on kokenut evoluution, jossa käyttäjän virheistä johtuvista datavuodoista on siirrytty datayhtiöiden palvelemille vuotaviin, usein tiedostamattomiin, datavuotoihin. Mutta onko samalla muuttunut yhteiskunnallinen diskurssi, vai vieläkö datavuodot ovat yleisessä keskustelussa vain käyttäjien omaa syytä tai pahojen hakkerien aiheuttamia?

Autoetnografista tai yhteistoiminnallista autoetnografista tutkimusotetta ei ole ainakaan kattavasti hyödynnetty tähänastisessa tutkimuksessa. Autoetnografiaan liittyy kuitenkin monia erityispiirteitä tutkijaposition ja tutkimuskenttään asettumisen osalta. Keskeisenä hyötynä voisi kuitenkin olla datayhtiöiden keräämiin tietoihin kiinni pääseminen. Näitä tietoja on aikaisemmassa tutkimuksessa pystytty hyödyntämään erittäin rajallisesti, koska henkilökohtaisen datan tutkimista rajoittavat yksityisyyslainsäädäntö, tutkimuseettiset haasteet ja datayhtiöiden haluttomuus luovuttaa yritystoiminnalleen keskeisiä tietoja tutkimuskäyttöön. Toinen vaihtoehto tällaisten aineistojen keräämiseen voisi olla kansalaistiede.

Tutkimusprosessini aikana esille tuli myös Googlen lyöttäytyminen akateemiseen tutkimukseen. Jos pro gradu -tutkielman kirjoittajaa ohjeistetaan hakemaan tietoa Googlen tarjoamasta tietokannasta, koska se on parempi kuin yliopiston oma kankea tietokanta, on ensiaskeleet kohti metsää otettu myös akateemisesti. Oppilaitoksissa on kitketty avoimen lähdekoodin tietosanakirja Wikipedian käyttö akateemisessa kontekstissa pois, koska kuka tahansa voi muokata siellä olevia tekstejä. Toisaalta Googlen käyttöön jopa akateemisessa tiedonhaussa (Google Scholar) ei joko kiinnitetä huomiota tai jopa rohkaistaan, vaikka Googlen tarjoamat hakutulokset ovat loppujen lopuksi yhden yksityisen yrityksen käsissä.

Datan keräämistä ja sen hyödyntämistä ei saada loppumaan ja tekoälyn sekä algoritmien kehittyessä ja monimutkaistuessa datan merkitys yhteiskunnassa korostuu entisestään. Yhteiskunnallisesti tärkeää olisi alkaa kiinnittää humanistis-yhteiskunnallisissa tieteissä huomio datan määrästä ja keräystavoista myös siihen, miten yritykset sitä käyttävät. Kun tähänastinen tutkimus on yhtä mieltä ainakin siitä, että datavuodot huolestuttavat ihmisiä mutta eivät saa heitä toimimaan millään tavalla vuotoja estääkseen, tulisi tutkimuksen siirtyä etsimään keinoja siihen, miten datavuotojen kohteet, eli tavalliset ihmiset, saataisiin toimimaan. Tämänhetkisen tutkimuksen perusteella me verkon käyttäjät tunnemme itsemme

toivottomiksi tai valitsemme olla tiedostamatta datavuotoihin liittyvät ongelmat ainakin niin pitkään, kun niiden vaikutukset eivät suoraan henkilökohtaisesti koske meitä. Datavuotojen tyrehdyttämiseen ei yksilötasolla riitä mobiilipeleistä sudokulehtiin siirtyminen, ja datavuotojen näkyväksi tekeminen vaatii yhteistyötä ja innovaatioita yli tiedekuntarajojen ja tutkimusperinteiden.

Lähteet

Acquisti, A., John, L. K., & Loewenstein, G. (2013). What Is Privacy Worth? The Journal of Legal Studies, 42(2), 249–274. https://doi.org/10.1086/671754

Aguirre, E., Roggeveen, A. L., Grewal, D., & Wetzels, M. (2016). The personalization-privacy paradox: implications for new media. The Journal of Consumer Marketing, 33(2), 98–110. https://doi.org/10.1108/JCM-06-2015-1458

Angrosino, M. (2007). Doing Ethnographic and Observational Research. [Online]. London: SAGE Publications.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Arora, P. (2019). General data protection regulation—a global standard? Privacy futures, digital activism, and surveillance cultures in the global south. Surveillance & Society, 17(5), 717–725. https://doi.org/10.24908/ss.v17i5.13307

Baneyx, A. (2008). 'Publish or Perish' as citation metrics used to analyze scientific output in the humanities: International case studies in economics, geography, social sciences, philosophy, and history. Archivum Immunologiae et Therapiae Experimentalis. [Online] 56 (6), 363–371.

Bearfield, Domonic A. & Warren S. Eller (2008). Writing a Literature Review: The Art of Scientific Literature. Handbook of Research Methods in Public Administration, 61–72. Boca Raton: CRC Press.

Berger, A. A. (2016). Media and communication research methods : an introduction to qualitative and quantitative approaches (Fourth edition.). Los Angeles: SAGE.

Bietz, M., Patrick, K., & Bloss, C. (2019). Data Donation as a Model for Citizen Science Health Research. Citizen Science : Theory and Practice, 4(1). https://doi.org/10.5334/cstp.178

Birch, K., & Bronson, K. (2022). Big Tech. Science as Culture, 31(1), 1–14. https://doi.org/10.1080/09505431.2022.2036118

Björk, B.-C. & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. Journal of informetrics. [Online] 7 (4), 914–923.

Bol, N., Strycharz, J., Helberger, N., van de Velde, B., & de Vreese, C. (2020). Vulnerability in a tracked society: Combining tracking and survey data to understand who gets

targeted with what content. New Media & Society, 22(11), 1996–2017. https://doi.org/10.1177/1461444820924631

Booth, A., Sutton, A., Papaioannou, D., & Booth, A. (2016). Systematic approaches to a successful literature review (Second edition.). London ;: SAGE Publications Ltd.

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. New Media & Society, 22(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Campbell, E. (2017). "Apparently being a self-obsessed CT is now academically lauded": Experiencing twitter trolling of autoethnographers. Forum, Qualitative Social Research, 18(3). https://doi.org/10.17169/fqs-18.3.2819

Chang, H., Hernandez, K.-A. C., & Ngunjiri, F. W. (2016). Collaborative autoethnography. London ;: Routledge. https://doi.org/10.4324/9781315432137

Chavez, P., Szathmary, Z. & Evans, S.J. (2014). Are your iCloud photographs STILL at risk? Apple unable to offer any assurances after theft of naked celebrity images which is deemed so serious the FBI are investigating. MailOnline. Saatavilla: <https://www.dailymail.co.uk/news/article-2739215/Jennifer-Lawrence-victim-hacker-leaks-slew-graphic-nude-photos-Oscar-winning-actress.html> (linkki tarkistettu 19.1.2023)

Colley, A., Pfleging, B., Alt, F., & Häkkilä, J. (2020). Exploring public wearable display of wellness tracker data. International Journal of Human-Computer Studies, 138, 102408–. https://doi.org/10.1016/j.ijhcs.2020.102408

Couldry, N., & Hepp, A. (2016). The Mediated Construction of Reality. (1st ed.). Oxford: Polity Press.

Couldry, N., & Mejias, U. A. (2019). The costs of connection : how data is colonizing human life and appropriating it for capitalism. Stanford, California: Stanford University Press. https://doi.org/10.1515/9781503609754

Denshire, S. (2014). On auto-ethnography. Current Sociology, 62(6), 831–850. https://doi.org/10.1177/0011392114533339

D'Onfro J. (2018). We sat in on an internal Google meeting where they talked about changing the search algorithm — here's what we learned. Saatavilla: <https://www.cnbc.com/2018/09/17/google-tests-changes-to-its-search-algorithm-how-search-works.html> (linkki tarkistettu 3.4.2023)

Doss, A. F. (2020). Cyber privacy : who has your data and why you should care. Dallas, TX: BenBella Books, Inc.

Efron, S. E., & Ravid, R. (2019). Writing the literature review : a practical guide. New York ;: Guilford Press.

Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 12(1), Article 10.

Ellis, C., & Adams, T. E. (2020). Practicing Autoethnography and Living the Autoethnographic Life. In The Oxford Handbook of Qualitative Research (pp. 359–396). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190847388.013.21

Eriksson, P., & Koistinen, K. (2005). Monenlainen tapaustutkimus. Helsinki: Kuluttajatutkimuskeskus.

Esguerra, R. (2009). Google CEO Eric Schmidt Dismisses the Importance of Privacy. Electronic Frontier Foundation. Saatavilla: <https://www.eff.org/deeplinks/2009/12/google-ceo-eric-schmidt-dismisses-privacy> (linkki tarkistettu 30.4.2023)

Euroopan parlamentin ja neuvoston asetus (EU) 2016/679 luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta. http://data.europa.eu/eli/reg/2016/679/2016-05-04 (linkki tarkistettu 5.1.2022)

Flynn, S. (2021). Celebrity Data Breaches: Why No Star Is Safe From Cybercriminals. MUO. Saatavilla: < https://www.businessinsider.com/google-manipulates-search-results-report-2019-11?r=US&IR=T> (linkki tarkistettu 25.5.2023).

Gilbert, B. (2019). Google reportedly manipulates search results to hide controversial subjects and favor big business. Business Insider. Saatavilla: <https://www.eff.org/deeplinks/2009/12/google-ceo-eric-schmidt-dismisses-privacy> (linkki tarkistettu 30.4.2023)

Gusenbauer, M. & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research synthesis methods. [Online] 11 (2), 181–217.

Haasio, A. (2013). Netin pimeä puoli. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hammersley, M. (2018). What is ethnography? Can it survive? Should it? Ethnography and education. [Online] 13 (1), 1–17.

Harari, Y. N. (2017). Dataism Is Our New God. New Perspectives Quarterly, 34(2), 36–43. https://doi.org/10.1111/npqu.12080

Helles, R., Lomborg, S., & Lai, S. S. (2020). Infrastructures of tracking: Mapping the ecology of third-party services across top sites in the EU. New Media & Society, 22(11), 1957–1975. https://doi.org/10.1177/1461444820932868

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. International Journal of Human-Computer Studies, 143, 102498–. https://doi.org/10.1016/j.ijhcs.2020.102498

Holman Jones, S. L., Adams, T. E., & Ellis, C. (2016). Handbook of autoethnography. London ;: Routledge.

Hyvärinen, M., Nikander, P., Ruusuvuori, J., & Aho, A. L. (2017). Tutkimushaastattelun käsikirja. Tampere: Vastapaino.

Johnson, B. (2010). Privacy no longer a social norm, says Facebook founder. The Guardian. Saatavilla: < https://www.theguardian.com/technology/2010/jan/11/facebook-privacy> (linkki tarkistettu 30.4.2023)

Kao, G., Hong, J., Perusse, M., & Sheng, W. (2020). Turning Silicon into Gold The Strategies, Failures, and Evolution of the Tech Industry (1st ed. 2020.). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-5629-9

Kennedy, H., Oman, S., Taylor, M, Bates, J., Steedman, R. (2020). Public understanding and perceptions of data practices: a review of existing research. Review summary. https://livingwithdata.org/current-research/publications/ (linkki tarkistettu 9.4.2021)

Komulainen, M. (2018). Menesty digimarkkinoinnilla (1. painos.). Helsinki: Kauppakamari.

Kristensen, D. B., & Ruckenstein, M. (2018). Co-evolving with self-tracking technologies. New Media & Society, 20(10), 3624–3640. https://doi.org/10.1177/1461444818755650

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A. K., Hammerstrøm, K., & Sathe, N. (2017). Searching for studies: a guide to information retrieval for Campbell systematic reviews. Campbell Systematic Review, 13(1), 1–73. https://doi.org/10.4073/cmg.2016.1

König, R., Uphues, S., Vogt, V., & Kolany-Raiser, B. (2020). The tracked society: Interdisciplinary approaches on online tracking. New Media & Society, 22(11), 1945–1956. https://doi.org/10.1177/1461444820924629

Lapadat, J. C. (2017). Ethics in Autoethnography and Collaborative Autoethnography. Qualitative Inquiry, 23(8), 589–603. https://doi.org/10.1177/1077800417704462

Lindgren, J., Mokka, R., Neuvonen, A., Toponen, A., Liukas, L., & Hirvonen, I. S. (2019). Digitalisaatio : murroksen koko kuva. Helsinki: Tammi.

Lomborg, S., Thylstrup, N. B., & Schwartz, J. (2018). The temporal flows of self-tracking: Checking in, moving on, staying hooked. New Media & Society, 20(12), 4590–4607. https://doi.org/10.1177/1461444818778542

Louridas, P., & Pietiläinen, K. (2021). Algoritmit. Helsinki: Terra Cognita.

Mariani, M. M., Ek Styven, M., & Teulon, F. (2021). Explaining the intention to use digital personal data stores: An empirical study. Technological Forecasting & Social Change, 166, 120657–. https://doi.org/10.1016/j.techfore.2021.120657

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. London: John Murray.

Mejias, U. A. & Couldry, N. (2019). Datafication. Internet Policy Review, 8(4). DOI: 10.14763/2019.4.1428

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society, 23(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Obringer, R., Rachunok, B., Maia-Silva, D., Arbabzadeh, M., Nateghi, R., & Madani, K. (2021). The overlooked environmental footprint of increasing Internet use. Resources, Conservation and Recycling, 167, 105389–. https://doi.org/10.1016/j.resconrec.2020.105389

Pentland, A. (2012). Society's Nervous System: Building Effective Government, Energy, and Public Health Systems. Computer (Long Beach, Calif.), 45(1), 31–38. https://doi.org/10.1109/MC.2011.299

Pullen, J. (2017). Jennifer Lawrence Reveals Why She Didn't Sue Apple Over Her Nude Photo Leak. Fortune. Saatavilla: < https://fortune.com/2017/11/21/jennifer-lawrence-apple-lawsuit-nude-photo-leak/> (linkki tarkistettu 19.1.2023).

Puusa, A., Juuti, P., & Aaltio, I. (2020). Laadullisen tutkimuksen näkökulmat ja menetelmät. Helsinki: Gaudeamus.

Rovira, C., Codina, L., & Lopezosa, C. (2021). Language bias in the google scholar ranking algorithm. Future Internet, 13(2), 1–17. https://doi.org/10.3390/fi13020031

Rumrill, Phillip D.; Fitzgerald, Shawn M., Merchant, William R. (2010). Using scoping literature reviews as a means of understanding and interpreting existing literature. Work 35 (2010) 399–404. DOI: 10.3233/WOR-2010-0998

Schufrin, M., Reynolds, S. L., Kuijper, A., & Kohlhammer, J. (2021). A Visualization Interface to Improve the Transparency of Collected Personal Data on the Internet. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1840–1849. https://doi.org/10.1109/TVCG.2020.3028946

Seale, C. (2012). Researching society and culture (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Silverman, D. (2020). Qualitative Research (5E. ed.). London: Sage Publications Ltd.

Stark, L. (2020). The emotive politics of digital mood tracking. New Media & Society, 22(11), 2039–2057. https://doi.org/10.1177/1461444820924624

Taiabul Haque, S. M., Wright, M., & Scielzo, S. (2014). Hierarchy of users′ web passwords: Perceptions, practices and susceptibilities. International Journal of Human-Computer Studies, 72(12), 860–874. https://doi.org/10.1016/j.ijhcs.2014.07.007

Taylor, L. (2016). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning. D, Society & Space, 34(2), 319–336. https://doi.org/10.1177/0263775815608851

Tsekoura, T-M. & Panagopoulou, F. (2020). 'GDPR: a critical review of the practical, ethical and constitutional aspects one year after it entered into force', Int. J. Human Right and Constitutional Studies, Vol. 7, No. 1, pp. 35-51.

Uotinen, J. (2010). Digital Television and the Machine That Goes "PING!": Autoethnography as a Method for Cultural Studies of Technology. Journal for Cultural Research, 14(2), 161–175. https://doi.org/10.1080/14797580903481306

van Dijck, J. (2014). Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. Surveillance & Society, 12(2), 197–208. doi:10.24908/ss.v12i2.4776

van Dijck, J., Poell, T., & de Waal, M. (2018). The Platform Society. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Voutilainen, T. (2020). Digitaalisten palvelujen sääntely. Helsinki: Alma Talent.

Vänskä, R., Härkönen, T., Suomalainen, K. (2020). Ihmisistä kerätty data uppoaa monimutkaisiin verkostoihin. https://www.sitra.fi/artikkelit/ihmisista-keratty-data-uppoaa-monimutkaisiin-verkostoihin/ (linkki tarkistettu 7.4.2021)

WebXray. https://webxray.org/ (linkki tarkistettu 14.4.2023)

Your Europe. Yksityisyyden suoja verkossa. https://europa.eu/youreurope/business/dealing-with-customers/data-protection/online-privacy/index_fi.htm (linkki tarkistettu 10.1.2023)

Zimerman, M. (2010). Technology and privacy erosion in United States libraries: a personal viewpoint. New Library World, 111(1/2), 7–15. https://doi.org/10.1108/03074801011015649

Zuboff, S. (2019) The Age of Surveillance Capitalism. London, UK: Profile Books.

Liitteet

Liite 1. Artikkelikatsauksen viimeisessä vaiheessa mukana olleet artikkelit

AlSabah, M., Oligeri, G., & Riley, R. (2018). Your culture is in your password: An analysis of a demographically-diverse password dataset. Computers & Security, 77, 427–441. https://doi.org/10.1016/j.cose.2018.03.014Gan, D., & Jenkins, L. R. (2015). Social Networking Privacy—Who's Stalking You? Future Internet, 7(1), 67–93. https://doi.org/10.3390/fi7010067

Ameen, N., Tarhini, A., Shah, M. H., Madichie, N., Paul, J., & Choudrie, J. (2021). Keeping customers' data secure: A cross-cultural study of cybersecurity compliance among the Gen-Mobile workforce. Computers in Human Behavior, 114, 106531–. https://doi.org/10.1016/j.chb.2020.106531

Barth, S., de Jong, M. D. T., Junger, M., Hartel, P. H., & Roppelt, J. C. (2019). Putting the privacy paradox to the test: Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources. Telematics and Informatics, 41, 55–69. https://doi.org/10.1016/j.tele.2019.03.003

Bol, N., Strycharz, J., Helberger, N., van de Velde, B., & de Vreese, C. . (2020). Vulnerability in a tracked society: Combining tracking and survey data to understand who gets targeted with what content. New Media & Society, 22(11), 1996–2017. https://doi.org/10.1177/1461444820924631

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. New Media & Society, 22(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Gan, D., & Jenkins, L. R. (2015). Social Networking Privacy—Who's Stalking You? Future Internet, 7(1), 67–93. https://doi.org/10.3390/fi7010067

Greenleaf, G., & Park, W. (2014). South Korea's innovations in data privacy principles: Asian comparisons. The Computer Law and Security Report, 30(5), 492–505. https://doi.org/10.1016/j.clsr.2014.07.011

Haynes, D., Bawden, D., & Robinson, L. (2016). A regulatory model for personal data on social networking services in the UK. International Journal of Information Management, 36(6), 872–882. https://doi.org/10.1016/j.ijinfomgt.2016.05.012

Helles, R., Lomborg, S., & Lai, S. S. (2020). Infrastructures of tracking: Mapping the ecology of third-party services across top sites in the EU. New Media & Society, 22(11), 1957–1975. https://doi.org/10.1177/1461444820932868

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. International Journal of Human-Computer Studies, 143, 102498–. https://doi.org/10.1016/j.ijhcs.2020.102498

Hintikka, K. (2022). Asiantuntijahaastattelu 26.1.2022 (sähköposti).

König, R., Uphues, S., Vogt, V., & Kolany-Raiser, B. (2020). The tracked society: Interdisciplinary approaches on online tracking. New Media & Society, 22(11), 1945–1956. https://doi.org/10.1177/1461444820924629

Leering, A., van de Wijngaert, L., & Nikou, S. (2022). More honour'd in the breach: predicting non-compliant behaviour through individual, situational and habitual factors. Behaviour & Information Technology, 41(3), 519–534. https://doi.org/10.1080/0144929X.2020.1822444

Maris, E., Libert, T., & Henrichsen, J. R. (2020). Tracking sex: The implications of widespread sexual data leakage and tracking on porn websites. New Media & Society, 22(11), 2018–2038. https://doi.org/10.1177/1461444820924632

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society, 23(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Olivero, M. A., Bertolino, A., Domínguez-Mayo, F. J., Escalona, M. J., & Matteucci, I. (2020). Digital persona portrayal: Identifying pluridentity vulnerabilities in digital life. Journal of Information Security and Applications, 52, 102492–. https://doi.org/10.1016/j.jisa.2020.102492

Ong, R. (2012). Data protection in Malaysia and Hong Kong: One step forward, two steps back? The Computer Law and Security Report, 28(4), 429–437. https://doi.org/10.1016/j.clsr.2012.05.002

Paramarta, V., Jihad, M., Dharma, A., Hapsari I. C., Sandhyaduhita, P. I. & Hidayanto A. N. (2018). Impact of User Awareness, Trust, and Privacy Concerns on Sharing Personal Information on Social Media: Facebook, Twitter, and Instagram. 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2018, pp. 271-276, doi: 10.1109/ICACSIS.2018.8618220.

Rosso, M., Nasir, A., & Farhadloo, M. (2020). Chilling effects and the stock market response to the Snowden revelations. New Media & Society, 22(11), 1976–1995. https://doi.org/10.1177/1461444820924619

Schufrin, M., Reynolds, S. L., Kuijper, A., & Kohlhammer, J. (2021). A Visualization Interface to Improve the Transparency of Collected Personal Data on the Internet. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1840–1849. https://doi.org/10.1109/TVCG.2020.3028946

Singh, P. (2021). Aadhaar and data privacy: biometric identification and anxieties of recognition in India. Information, Communication & Society, 24(7), 978–993. https://doi.org/10.1080/1369118X.2019.1668459

Sparks, H., Collins, F. L., & Kearns, R. (2016). Reflecting on the risks and ethical dilemmas of digital research. Geoforum, 77, 40–46. https://doi.org/10.1016/j.geoforum.2016.09.019

Swart, I. P., Grobler, M. M., & Irwin, B. (2013). Visualization of a data leak. 2013 Information Security for South Africa, 1–8. IEEE. https://doi.org/10.1109/ISSA.2013.6641046

Taiabul Haque, S. M., Wright, M., & Scielzo, S. (2014). Hierarchy of users′ web passwords: Perceptions, practices and susceptibilities. International Journal of Human-Computer Studies, 72(12), 860–874. https://doi.org/10.1016/j.ijhcs.2014.07.007

Taylor, L. (2016). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning. D, Society & Space, 34(2), 319–336. https://doi.org/10.1177/0263775815608851

Wang, X., Wang, X., Liu, Z., Chang, W., Hou, Y., & Zhao, Z. (2022). Too generous to be fair? Experiments on the interplay of what, when, and how in data breach recovery of the hotel industry. Tourism Management (1982), 88, 104420–. https://doi.org/10.1016/j.tourman.2021.104420

Yu, W., & Shen, F. (2021). The relationship between online political participation and privacy protection: evidence from 10 Asian societies of different levels of cybersecurity. Behaviour & Information Technology, ahead-of-print(ahead-of-print), 1–16. https://doi.org/10.1080/0144929X.2021.1953597

Zhou, C., Li, K., & Zhang, X. (2022). Why do I take deviant disclosure behavior on internet platforms? An explanation based on the neutralization theory. Information Processing & Management, 59(1), 102785–. https://doi.org/10.1016/j.ipm.2021.102785

Liite 2. Kirjallisuuskatsauksen artikkelien koonti

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
1. (#1.13)	Swart, I.P.; Grobler, M.M.; Irwin B. (2013), Etelä-Afrikka	Mitä dataa ja missä muodossa internetiin vuosi 2013 Etelä-Afrikan hakkeritapauksessa? Miten vuodetun datan visualisointi vaikuttaa sen kokemiseen?	Vuonna 2013 Etelä-Afrikassa eri yritysten palvelimilta verkkoon vuotanutta henkilökohtaista dataa (700 000 yksikköä sisältäen mm. käyttäjänimi ja salasanoja, sivilisääty- ja pankkitietoja, henkilökohtaisia viestejä). Data teemoiteltiin ja aineiston käyttö perusteltiin sen vapaalla saatavuudella verkosta, vaikka sinne vuotaminen oli rikollista alkuperää.	Valtavan datamäärän käsittely visualisointia varten haastavaa. Yksilön tunnetasolla arviota ei voitu suorittaa (tutkimuskysymys 2), mutta visualisointi auttoi työryhmää datavuotojen uhkien määrittelemisessä. Henkilökohtaisten tietojen säilömistä koskevaa lainsäädäntöä tulisi uudistaa	Eettisesti kyseenalaista, aineisto rikollista alkuperää
2. (#2.6)	Maris, Elena; Libert, Timothy; Henrichsen, Jennifer R (2020), Yhdysvallat	Mitä ja miten paljon käyttäjädataa pornosivustoilta vuotaa kolmansille osapuolille, mitä kolmannet osapuolet ovat, voiko seksuaaliset kiinnostuksenkohteet vuotaa kolmansille osapuolille?	22 484 verkkosivustoa, joiden verkko-osoite, sivun otsikko tai kuvaus sisälsivät sanan "porn". Sivustoista kaavittiin myös 3856 yksityisyyskäytäntö (privacy policy) ilmoitusta. Data kaavittiin webXray ohjelmiston avulla.	Pornosivustoilla yhdistyy erityisen yksityinen data, löyhät datansuojaustoimet ja riski hakkerointien aiheuttamille vuodoille. Seksuaalista suuntautumista tai valtavirrasta poikkeavia mielenkiinnonkohteita olettavaa dataa vuotaa kolmansille osapuolille. Uhka erityisesti marginalisoiduille ihmisryhmille. Tietoja kerätään, mutta niiden käyttö epäselvää. Yksityisyydensuojalausekkeet kautta linjan	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
				vaikeaselkoisia eikä kaikilta osin saatavilla.	
3. (#2.8)	Valentinus Paramarta; Muhammad Jihad; Ardhian Dharma; Ika Chandra Hapsari; Puspa Indahati Sandhyaduhita; Achmad Nizar Hidayanto (2018), Indonesia	Miten tietoisuus sosiaalisen median alustojen vuodoista vaikuttaa käyttäjiin?	340 sosiaalisen median käyttäjää. Verkkokysely, vastattavissa 2kk. Kyselyä välitettiin sähköpostilla ja pikaviestipalvelun kautta indonesialaisille henkilöille. 92.1% vastaajista vähintään kandidaattitutkintoa parhaillaan suorittavia henkilöitä.	Tietoisuus yksityisyyteen liittyvistä ongelmista lisää yksityisyyttä suojaavien toimintojen käyttämistä ja rajoittaa käyttäjien itse palveluihin lisäämien tietojen määrää merkittävästi.	Kyllä
4. (#5.5)	Gan, Diane; Jenkins, Lily R. (2015), Englanti	Mitä kaikkea tietoa Twitterin käyttäjästä voidaan saada pelkästään yhden (sijaintimerkityn) twiitin kautta?	Esitutkimus 90 sattumanvaraisella Twitter-tilillä joista kerättiin tilastollista dataa, varsinaisessa näytteessä 3 henkilöä joilta suostumus. Paikannustiedon sisältävät twiitit julkista dataa, kaavintoja tehty vapaasti saatavilla olevilla mining-ohjelmistoilla.	Pystyttiin seuraamaan kuljettuja reittejä, selvittämään työajat ja kodin ja työpaikan sijainti. Yhden some-kanavan avulla helppo löytää muutkin varsinkin jos profiilikuva on sama. Pelkän geodatan perusteella olisi ollut haastavaa, mutta yhdistelemällä geodata twiittien sisältöön paikkojen nimeäminen oli varsin nopeaa ja helppoa. Kerätty data riittäisi identiteettivarkauteen.	Kyllä, nykyään Twitter julkaisee sijainnin vain jos se on erikseen laitettu päälle käyttäjän toimesta. Käytetyt louhintatyökalut vanhentuneita.

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
5. (#5.7)	Schufrin, Marija; Steven Lamarr Reynolds; Kuijper, Arjan; Kohlhammer, Jorn (2021), Saksa	Miten GDPR:n avulla saatavilla olevista datavuodoista voitaisiin muotoilla käyttäjälle selkeitä?	GDPR:n perusteella saatu data Facebookista, Googlesta, Twitteristä ja Instagramista. Tutkimusryhmän itse kehittämän prototyyppiohjelmiston (TransparencyVis) koekäyttö 37 saksalaisen vapaaehtoisen henkilökohtaisella datalla. 30 ikäluokassa 20-34, 12 opiskelijaa.	Datan visualisointi auttaa tavallisia käyttäjiä ymmärtämään ja käsittelemään oman datansa käyttöä internetissä. 45% ilmoitti muuttaneensa yksityisyyteen liittyvää verkkokäyttäytymistään. Jo ennen visualisointia omasta yksityisyydestään välinpitämättömien ryhmässä ei kuitenkaan ollut merkittävää muutosta. Koekäytön perusteella palvelu julkaistiin vapaasti käytettäväksi.	Kyllä
6. (#5.8)	AlSabah, Mashael; Oligeri, Gabriele; Riley, Ryan (2018), Qatar	Miten kulttuuri ja kieli vaikuttavat salasanojen valintaan?	Aineisto 400 000 Qatarilaisen pankin asiakkaan tietoja sisältävä Lähi-Idässä 04/2016 tapahtunut datavuoto. Asiakkaat mm. arabeja, filippiiniläisiä, intialaisia ja pakistanilaisia. Aineisto sisälsi tietoja salasanoista, nimistä, kansalaisuuksista, osoitteista, salasanojen palautuskysymyksistä ja -vastauksista. Aineisto vuodettu sosiaaliseen mediaan ja	Salasanan valintaan vaikuttiniin henkilön demografinen asema kuin kielellinen tausta. Pankkisivuston salasanat olivat selkeästi vahvempia kuin verrokkiryhmänä olleiden aikaisempien vuotojen salasanat. Salasanan vahvuuden arviointiin käytetyt menetelmät ovat englantipainotteisia, eivätkä huomioi esimerkiksi arabiankielisiä yleisiä salasanoja.	Eettisesti kyseenalaista, aineisto rikollista alkuperää

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
			kerätty sieltä tutkimuskäyttöön. Teemoittelu, kielianalyysi.		
7. (#6.3)	Hayley Sparks; Francis L.Collins; Robin Kearns (2016), Uusi-Seelanti	Millaisia eettisiä riskejä ja dilemmoja digitaaliseen tutkimustyöhön sisältyy? Mitä tapahtuu, kun arkaluontoista tutkimusdataa vuotaa?	Tohtoritutkintoon (PhD) tähtäävän tutkimuksen aikana tapahtunut datavuoto, jossa tutkimukseen osallistuneiden henkilöiden päiväkirjatyyppisiä merkintöjä vuosi koko ryhmän luettavaksi.	Datavuodosta huolimatta tutkimukseen osallistuneet henkilöt jatkoivat osallistumista. Tutkija voi pahoin, eli käytetyillä menetelmillä voi epäonnistuessaan olla odottamattomia vaikutuksia tohtorikoulutettavan hyvinvointiin. Työkaluja tutkijoiden henkiseen hyvinvointiin pitäisi olla paremmin saatavilla. Digitaalisten tutkimusmetodien käyttöön tulisi kiinnittää enenevissä määrin huomiota tutkijoiden koulutuksessa.	Eettisesti kyseenalaista, vaatisi uuden tutkimuksen ohteina olevien henkilöiden tietojen vuotamisen.

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
8. (#6.4)	Miguel Angel Olivero; Antonia Bertolino; Francisco José Domínguez-Mayo; María José Escalona; Ilaria Matteucci (2020), Espanja	Mitä tietoja (dataa) kohdehenkilöistä löytyi julkisena internetistä, mitä informaatiota voidaan luoda yhdistelemällä näitä tietoja ja voiko yksilölähtöinen verkkopresenssi johtaa yksilön haavoittuvuuteen ?	17 datavuodon uhrin tiedot, lähteenä verkkokyselypanelin tietovuoto. Verkkokaavinta (kyselytutkimuksen tietovuodosta)	Salasanojen turvakysymyksiin voidaan saada vastauksia sosiaalisen median julkisten tietojen avulla. Muita tietoja esimm. ikä ja syntymäaika (onnentoivotuksista), sukupuoli, kansalaisuus, osasta myös työpaikka, koulutus, puolelta käyttäjistä myös kasvot. Tiedot riittävät identiteettivarkauteen. Eri verkkoalustojen identiteettejä käsiteltiin SoS (System of systems) -järjestelmänä, jonka tietoturvaheikkouksia kartoitettiin.	Kyllä, aineisto lähde tosin vaatii eettistä tarkastelua. Turvakysymykset vaihtuneet useissa palveluissa puhelin jne. varmennuksiin.
9. (#6.9)	Rebecca Ong (2012), Kiina (Hong Kong)	Miten Malesian ja Hong Konging datasuojalait eroavat muiden maiden vastaavista? Keskittyykö lainsäädäntö organisaatioiden toiminnan legitisoimiseen vai yksilön datan suojaamiseen?	Yksityisyys- ja datasuojalait ja -säädökset Malesiassa ja Hong Kongissa. Saatavilla vapaasti verkossa.	Malesian säädöksiin on varioitu OECD:n ohjeistuksia ja GDPR:n kohtia. Hong Kongin PDPO laahaa jäljessä. Molemmilla pyritään vastaamaan EU-direktiivien vaatimuksiin. Kumpikaan lainsäädäntö ei turvaa yksityisyyden suojaan liittyviä riskejä datan keräämisestä ja käytöstä yksilöihin kohdistuvaan	Kyllä

Järjesty sluku, tunniste	Tekijät, vuosi, taustaorgan isaation maa	Tutkimuskysym ys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettav uus 2022 (EU-maassa)
				profilointiin ja markkinointiin.	
10. (#6.12)	David Haynes; David Bawden; Lyn Robinson (2016), Englanti	Miten Englannin Data protection act 1998 toimii ja millainen on Englannin datasuojan toimintaympäristö erityisesti sosiaalisen median alustat huomioiden?	10 haastattelua asiantuntijoiden kanssa (sosiaalisen median alustat, sääntelytyötä tekevät, tutkijat) kasvotusten ja puhelimessa.	Palveluiden tarjoajien itsesääntely suuressa roolissa, parannuksia tarvitaan erityisesti yritysten vastuuseen ilmoittaa datavuodoista. Ehdotetaan sääntelymallia mukaellen Lessigin 2006 jaottelua, jossa sääntely kohdistetaan erikseen lakeihin, koodaukseen, normeihin ja markkinoihin.	Kyllä
11. (#6.14)	Cheng Zhou; Kai Li; Xiaofei Zhang (2022), Kiina	Mitkä tekijät saavat verkkopalveluiden käyttäjän antamaan vääriä henkilökohtaisia tietoja?	Kyselytutkimus, skenaariokysely verkossa, vastausaika 3 viikkoa, 306 validoitua vastausta. Kyselyn toteutti verkkokyselyitä tarjoava kiinalainen yritys. Vastaajista yli 50% vähintään alemman korkeakoulututkinnon suorittaneita, 64% ikäluokassa 25-30.	Käyttäjät käyttivät neutralisointiteknii koita hyöty-riski vaihdannan ajattelemisen sijaan. Viisi suurta persoonallisuuste kijää (neuroottisuus, ekstroversio, avoimuus uusille kokemuksille, sovinnollisuus ja tunnollisuus) vaikuttivat päätöksentekoon tietojen antamisessa, mutta muita vaikuttavia tekijöitä oli mm. turvallisuuden tunne tietovuotojen osalta ja annettavan tiedon laatu (esim. terveyttä	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
				koskevista tiedoista annettiin herkemmin väärää tietoa).	
12. (#6.17)	Graham Greenleaf; Whon-il Park (2014), Australia, Etelä-Korea	Miten Etelä-Korean Personal Information Privacy Act of 2011 eroaa muista vastaavista laeista?	Etelä-Korean Personal Information Privacy Act of 2011 (PIPA), OECD:n ja EU:n säädökset, muiden Aasian maiden lait ja yksityisyyden- ja datasuojaa koskevat säädökset. Saatavilla vapaasti verkossa.	Muihin Aasian maihin verrattuna Etelä-Korealla "innovatiivisin datasuojalaki", sisältää mm. pakollisen riskianalyysin potentiaalisesti vuotaville julkisen sektorin verkkojärjestelmille, tietojen poistamisen käyttäjän pyynnöstä ja käyttäjäsuostumusten ketjuttamisen kieltämisen. Käytäntöönpano ja valvonta vielä epäselvää.	Kyllä
13. (#6.19)	Xuhui Wang; Xuequn Wang; Zilong Liu; Wen Chang; Yuansi Hou; Zhihe Zhao (2022), Kiina, Australia, Englanti	Milloin ja miten hotellialalla voidaan ja kannattaa kompensoida datavuotojen uhreja?	Skenaario lähetettiin sähköpostitse, 1. koeryhmä luki ensin uutisen datavuodosta ja sitten viestin jossa tarjottiin kompensaatiota, 2. koeryhmä toisinpäin. 178 vastausta, vastaajat yliopisto-opiskelijoita.	Arvointi hankalaa, koska yksilöiden kokemukset datavuotojen haitoista vaihtelevat suuresti. Rahallinen korvaaminen haastavaa, koska vuodot yleensä suuria. Kompensaation aikaansaama hyödyn tunne asiakkaalle maksimoituu, jos kompensaatio on rahallinen ja tarjottu ennen kuin asiakas lukee vuodosta muista lähteistä.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
14. (#6.23)	Nisreen Ameen; Ali Tarhini; Mahmood Hussain Shah; Nnamdi Madichie; Justin Paul; Jyoti Choudrie (2021), Englanti, Oman, Yhdysvallat	Miten omien matkapuhelinten käytön aiheuttamia riskejä voitaisiin vähentää työpaikoilla erityisesti Gen-Mobile (18-35v.) työntekijöiden keskuudessa?	Kyselykaavakkeet jaettiin kasvotusten (vastausprosentit vaihteli 85-93% välillä). 1735 vastausta (UK, USA, Arabiemiraatit). Vastaajat kansainvälisten yritysten 18-35v. työntekijöitä, joilla oli käytössään älypuhelin	Huomattavia eroja maasta riippuen, paikallinen sääntely vaikutti paljon työntekijöiden asenteisiin ja käyttäytymiseen. Merkittävä tekijä kansallisista säädöksistä riippumatta oli kuitenkin tietoturvasta kiinnostunut ja innostava työympäristö, erityisesti esihenkilöiden osalta.	Kyllä
15. (#6.32)	Susanne Barth; Menno D.T. de Jong; Marianne Junger; Pieter H. Hartel; Janina C. Roppelt (2019), Alankomaat	Vaikuttaako yksityisyysparadoksiin käyttäjän teknologinen osaaminen, tietoisuus tietoturvasta ja/tai varallisuus?	66 tietojärjestelmätieteiden maisteritason kurssin opiskelijaa. Empiirinen koe. Osallistujat täyttivät taustatietokyselyt lomakkeina, jonka jälkeen heidän piti valita ja ladata mobiilisovellus omaan puhelimeensa. Osallistujat kirjoittivat myös arvostelun sovelluksesta, jossa kuvailivat lataus- ja päätöksentekoprosessia.	Huolimatta siitä, että koeryhmällä oli teknologinen osaaminen, varallisuutta ja tietoisuus mahdollisista tietoturvariskeistä, mobiilisovellusta valittaessa merkittävät tekijät olivat sovelluksen toimivuus, muotoilu ja hinta. Itsearvioinneissa todettiin erilaisten datan jakamiseen liittyvien suostumusten olevan merkitsevä tekijä, mutta käytännössä sillä ei ollut merkitystä.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
16. (#7.2)	Wenting Yu; Fei, Shen (2021), Kiina (Hong Kong)	Miten politiikkaan osallistuminen verkkoympäristössä vaikuttaa yksityisyyden suojaamiseen liittyvään käytökseen verkossa?	6691 kyselyvastausta eri puolilta Aasiaa. Kyselyn tuotti kansainvälinen kysely-yritys YouGov.	Henkilöt, jotka olivat poliittisesti aktiivisia verkossa kiinnittivät enemmän huomiota omaan yksityisyydensuojaansa. Ns. matalan verkkoturvallisuuden maissa motivaatio yksityisyyden suojaamiseen pyrkimisessä oli pelko mahdollisista vuodoista. Poliittisessa kontekstissa yksityisyydensuojan tulisi olla ihmisoikeus. Kiinaa ei voitu ottaa tutkimukseen mukaan, koska kyselydatan tuottaminen poliittisesta osallistumisesta mahdotonta.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
17. (#7.4)	Singh, Pawan (2019), Australia	Millaisia yksityisyydesuojaan kohdistuvia ongelmia biometristä tunnistusta hyödyntävällä Aadhar avustusjärjestelmällä on Intiassa?	Aineisto laki- ja asiakirjatekstejä yksityisyydensuojasta ja aktivistien nauhoittamia videodokumentointeja. Saatavilla vapaasti verkossa.	Biodatan (sormenjälki) käyttöä tunnistautumiseen perustellaan mm. korruption vähentämisellä, mutta se luo yhteiskuntaluokan, jonka valvonta on sallittua ja vaikeuttaa haavoittuvimmassa asemassa olevien ihmisten elämää. Valtiollinen järjestelmä ja tietokanta, johon vain köyhien on pakko rekisteröityä ja jakaa tietonsa saadakseen valtiolta tukia on syrjivä. Oikeuden yksityisyyteen tulee olla universaali yhteiskuntaluokasta riippumatta.	Kyllä
18. (#7.19)	Alex Leering; Lidwien van de Wijngaert; Shahrokh Nikou (2020), Alankomaat, Suomi, Ruotsi	Mitkä tekijät vaikuttavat työntekijöiden piittaamattomuuteen yrityksen tietoturvaohjeista?	Tapausepisodi (vignette) tutkimus, kysely lähetettiin sähköpostitse hollantilaisille valtionalaisen organisaation työntekijöille, 651 vastausta (18% vastaanottajista). Vastaukset analysoitiin SmartPLS ohjelmistolla.	Suuri osa yritysten datavuodoista johtuu työntekijöistä (human error). Ohjeiden seuraamiseen vaikutti aikapaine, tietojen arkaluontoisuus ja ympäröivät olosuhteet. Sen sijaan sosiaalisilla arvoilla ja uskomuksilla oli aikaisempia tutkimuksia vähäisempi merkitys.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
19. (#7.20)	Jonathan A. Obar; Anne Oeldorf-Hirsch (2018), Kanada, Yhdysvallat	Miten ihmiset lukevat yksityisyydensuojalausekkeita ja verkko-ohjelmistojen käyttöehtosopimuksia?	N=543 empiirinen koe. Osallistujat yhdysvaltalaisia viestinnän perustutkinto-opiskelijoita yliopistossa. Mitattiin aikaa, joka käytettiin fiktiivisen sosiaalisen median alustan yksityisyydensuojalausekkeen ja käyttöehtosopimuksen lukemiseen sekä reagointia niissä esitettyihin asioihin. Lisäksi itsearviointi.	Valtaosa koki yksityisyydensuojalausekkeet ja käyttöehtosopimukset ärsyttävinä. Kolme neljäsosaa eivät edes avanneet yksityisyydensuojalauseketta. 98% eivät lukeneet tai reagoineet kohtiin joissa hyväksymällä sopimukset esim. luovutti esikoislapsensa sovelluksen kehittäjille ja suostui kaikkien tietojensa jakamiseen kaikille tuleville työnantajilleen. Käyttäytyminen oli varsin hyvin linjassa itsearviointien kanssa.	Kyllä
20. (#SB1.1)	René König; Steffen Uphues; Verena Vogt; Barbara Kolany-Raiser (2020), Saksa	Mitä näkökulmia verkossa tapahtuvaan seurantaan liittyy tällä hetkellä? Millaisia haasteita datankeruuseen ja sen tutkimukseen liittyy?	Kirjallisuuskatsaus, datan keräämistä käsittelevän tieteellisen julkaisun teemanumeron aiheiden esittely.	Yksityisyys ja sen loukkaukset verkossa on monialainen ilmiö, jollaisena sitä tulee myös tutkia. Yksityisyyden määritteleminen on haastavaa. GDPR:n vaikutus vähäinen ja osin haitallinen (byrokratia, monimutkaiset lakitekstit käyttäjille). Kritisoivan tutkimuksen lisäksi tulisi pohtia miten dataa voisi käyttää yhteisen hyvän saavuttamiseksi.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
21. (#SB1.2)	Rasmus Helles; Stine Lomborg; Signe Sophus Lai (2020), Tanska	Miten verkkosivujen kolmansien osapuolten (third-party services, TPSs) käyttö eroaa eri EU-maiden, alueiden ja erityyppisten sivustoiden välillä?	4200 verkkosivua (top 150 eurooppalaista verkkosivua 28 EU maasta). Datakaavinta webXray-ohjelmistolla. Verkostoanalyysi.	Vaikka Google ja Facebook lähes monopoliin verrattavassa asemassa, kolmansien osapuolten sisäinen rakenne on huomattavasti monimutkaisempi. Pitkä häntä -ilmiö (longtail pattern), huomiota tulisi kiinnittää myös valtioiden virastoihin ym. vähemmän kaupallisiin kolmansiin osapuoliin. Eniten kolmansille osapuolille menee dataa keltaisen lehdistön ja arpajais/uhkapelisivujen kautta. Itä-Euroopassa kolmannet osapuolet usein venäläisiä sivustoja.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
22. (#SB1.3)	Mark Rosso; ABM Nasir; Mohsen Farhadloo (2020), Yhdysvallat, Kanada	Miten yksilöiden verkkohakukäyttäytyminen ja taloudelliset muuttujat (pörssikurssit) reagoivat Edward Snowdenin tietovuotoon?	Anonyymien verkkohakujen statistiikka ja 18 tietoturvayrityksen (julkiset) pörssikurssit rajatulta ajalta. Verkkohaku (DuckDuckGo [DDG)]hakukoneella, joka ei tallenna käyttäjähakujen sisältöjä eikä profiloi käyttäjiä eli näyttää kaikille käyttäjille samat hakutulokset. Segmentoitu aikasaraja-analyysi.	Itsesensuuria aiheuttava tukahduttava vaikutus (chilling effect) on olemassa, kun verkon käyttäjät muuttivat käytöstään Snowdenin tietovuodon paljastaman valtiollisen seurannan julkisuuteen tulon jälkeen. Tietoturvayritysten pörssikursseihin paljastus ei kuitenkaan vaikuttanut, mutta toisaalta paljastukset myös lisäsivät kysyntää yksityisyydensuojapalveluille, eivätkä vain horjuttaneet väestön uskoa omaan suojaukseensa.	Kyllä
23. (#SB1.4)	Nadine Bol; Joanna Strycharz; Natali Helberger; Bob van de Velde; Claes H de Vreese (2020), Alankomaat	Miten sosiodemografiset tekijät vaikuttavat sosiaalisen median (Facebook) käyttäjille näytettävään markkinointisisältöön huomioiden erityisesti haavoittuvissa asemissa olevat ihmisryhmät?	Vapaaehtoinen alankomaalainen koeryhmä asensi verkkoliikennettä mittaavan lisäosan verkkoselaimeensa, jonka avulla kartoitettiin Facebookin näyttämää sisältöä. Lisäksi osallistujat täyttivät kyselyn. Osallistujat pystyivät sammuttamaan seurannan milloin tahansa. 712 osallistujaa asensi seurantatyökalun. 97 osallistujan	Algoritmeilla kohdistetaan markkinointia tiettyihin käyttäjäryhmiin. Erityisesti erottui sukupuoli-stereotypisointi (miehille teknologiaa ja autoja, naisille kauneuden- ja terveydenhoitotuotteita) ja terveystuotteiden markkinoinnin kohdentaminen vanhempiin käyttäjiin, naisiin ja henkilöihin jotka kyselyn perusteella luottivat	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
			Facebook-dataa saatiin kerättyä ja niistä 80 täyttivät myös kyselyn.	verrokkejaan enemmän verkossa toimiviin yrityksiin.	
24. (#SB1.6)	Johannes Breuer; Libby Bishop; Katharina Kinder-Kurlanda (2020), Saksa	Miten digitaalisten jälkien dataa voidaan hankkia tutkimuskäyttöön, ja millaisia eettisiä ja käytännön haasteita siihen liittyy?	Oma aiempi tutkimus tapaustutkimuksena. Kirjallisuuskatsaus.	Dataa voidaan hankkia tutkimusryhmän omilla työkaluilla (ohjelmointirajapinnat, verkkokaavinta), yhteistyöllä yritysten kanssa tai ostamalla dataa markkinointitutkimus- tai datayrityksiltä. Vaihtoehtoisia tapoja esim. "datalahjoitukset" vapaaehtoisilta. Haasteina datan kerääminen ja julkaiseminen eettisesti kestävästi käytetystä hankintatavasta riippumatta, riittävät resurssit (aika, raha, tietotaito), datayritysten käyttöehdot, tekijänoikeudet.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
25. (#SB2.1)	S. M. Taiabul Haque; Matthew Wright; Shannon Scielzo (2014), Yhdysvallat	Miten verkkosivuston tyyppi vaikuttaa salasanan keksimiseen ja miten helposti salasanat ovat selvitettävissä?	80 opiskelijan erityyppisille sivustoille keksimät salasanat, kysely salasanojen keksimisestä ja koe, jossa hyökättiin uutissivustolle keksittyihin salasanoihin yrittäen selvittää pankkisivustolle keksitty salasana.	Kokeessa keksittyjä alemman tärkeystason (esim. uutissivustot) salasanoja käyttämällä voitiin selvittää yksi kolmasosa korkean tärkeystason (esim. pankkipalvelut) salasanoista. Samojen salasanojen lisäksi käytettiin samanlaisia tapoja koota salasana. Artikkelissa ei testattu salasanojen semanttista samankaltaisuutta. Alemman tärkeystason salasanojen datavuotoja voidaan käyttää muiden palveluiden salasanojen selvittämiseen.	Kyllä
26. (#SB2.2)	Joanne Hinds; Emma J. Williams; Adam N. Joinson (2020), Englanti	Miten Cambridge Analytica -skandaali vaikutti Facebook-profiileihin ja asenteisiin datan käytöstä?	Teemahaastattelu, 30 Facebookia käyttävää yliopisto-opiskelijaa tai yliopistolla työskentelevää henkilöä. Haastattelut pidettiin huhti-toukokuussa 2018 (Cambridge Analytican toiminta nousi julkisuuteen maaliskuussa 2018).	Skandaali ei muuttanut selauskäytäntöjä, haastatellut eivät muuttaneet tai perehtyneet omiin turvallisuusasetuksiinsa. Kokeeseen osallistujat vaikuttivat ajattelevansa olevansa immuuneja psykografisesti kohdennetulle mainonnalle.	Kyllä

Järjestysluku, tunniste	Tekijät, vuosi, taustaorganisaation maa	Tutkimuskysymys	Aineisto ja sen keruutapa, merkittävät tutkimusmetodit	Keskeiset huomiot ja tulokset	Toistettavuus 2022 (EU-maassa)
27. (#H.1)	Taylor, Linnet (2016), Alankomaat	Millaisia eettisiä ja menetelmällisiä haasteita liittyy tutkimuksiin, joissa tarkastellaan ihmisten liikkumista matkapuhelimista kerättävän datan avulla matalan toimeentulon maissa?	Tapaustutkimus, jossa aineistona operaattorin julkaisemasta matkapuhelimista Saharan eteläpuolisesta Afrikasta kerätystä datasta tehdyt tieteelliset julkaisut.	Riskinä kerätyn aineiston väärintulkinta, mikäli tutkijoilta puuttuu konteksti ja ymmärrys alueen kulttuurista, poliittisesta tilanteesta ja lainsäädännöstä. Harvaan asutuilla alueilla datan anonymisointi erittäin haastavaa. Tarve eettiselle ohjeistukselle datan julkaisusta tutkimuskäyttöön ja kansainvälisten suuryritysten itsesääntelyyn.	Ei, operaattorit eivät luovuta vastaavaa dataa tutkimuskäyttöön.