JULKAISUJA 7/2017

Anonymisointipalvelut
Tarve ja toteutusvaihtoehdot

LIIKENNE- JA VIESTINTÄMINISTERIÖ

LVM 1892–2017

Suomi Finland 100

Liikenne- ja viestintäministeriön

visio
Hyvinvointia ja kilpailukykyä hyvillä yhteyksillä

toiminta-ajatus
Liikenne- ja viestintäministeriö edistää väestön hyvinvointia ja elinkeinoelämän kilpailukykyä.
Huolehdimme toimivista, turvallisista ja edullisista yhteyksistä.

arvot
Rohkeus
Oikeudenmukaisuus
Yhteistyö

LIIKENNE- JA VIESTINTÄMINISTERIÖ

Julkaisun päivämäärä
3.4.2017

Julkaisun nimi
Anonymisointipalvelut. Tarve ja toteutusvaihtoehdot

Tekijät
Asta Bäck, VTT
Janne Keränen, VTT

Toimeksiantaja ja asettamispäivämäärä
Liikenne- ja viestintäministeriö

Julkaisusarjan nimi ja numero

Liikenne- ja viestintäministeriön julkaisuja 7/2017

ISSN (verkkojulkaisu) 1795-4045
ISBN (verkkojulkaisu) 978-952-243-503-3
URN http://urn.fi/URN:ISBN:978-952-243-503-3

Asiasanat
Anonymisointi, henkilötietolaki, liiketoiminta, henkilödata

Yhteyshenkilö
Taru Rastas

Tiivistelmä

Selvityksen tavoitteena oli muodostaa kokonaiskuva siitä, millaisia käytännön vaihtoehtoja anonymisointipalveluiksi on ja voisi olla, jotta erityisesti datan liiketoiminnallinen hyödyntäminen tulisi mahdolliseksi.

Taustaksi selvitettiin yritysten ja viranomaisten datan anonymisointiin liittyviä käytäntöjä, tarpeita, ja haasteita. Selvityksen aikana oltiin yhteydessä kaikkiaan 28 tahoon, joista 15 oli yrityksiä ja loput viranomaisia tai tutkimusorganisaatioita. Haastattelut osoittivat, ettei anonymisointi ole yrityksissä laajasti käytössä, eikä näin ollen vakiintuneita ratkaisuja anonymisoinnin tekemiseen juuri ole. Monet haastatelluista olivat kiinnostuneita anonymisoidun datan hyödyntämisestä ja he näkivät alueen tärkeänä ja merkitykseltään kasvavana. Anonymisoinnin riittävyyden arviointi koettiin vaikeaksi, mikä on tehnyt toimijat varovaisiksi anonymisoidun datan tarjoamisessa. On siis tarve esimerkeille ja ohjeille, jotka auttavat käytännön työssä ja edistävät datan hyödyntämistä.

Tutkimukseen osallistuneet yritykset pitivät anonymisoinnin riittävyyden tarkistamiseen liittyviä palveluita ja itse hoidettavaa anonymisointia tukevia asiantuntijapalveluita kiinnostavina. Datan anonymisoinnin teettäminen kolmannella osapuolella tuntui monesta vastaajasta vieraalta, ja herätti huolta aineiston tietoturvasta ja vastuista. Anonymisointipalvelu, jossa eri toimijoiden datoja pystyttäisiin yhdistämään ennen niiden anonymisointia, antaisi kuitenkin uusia mahdollisuuksia, jotka jäävät saavuttamatta yrityskohtaisessa anonymisoinnissa.

KOMMUNIKATIONS-
MINISTERIET

Publikationsdatum
3.4.2017

Publikation
Anonymiseringstjänster. Behov och alternativ för genomförande

Författare
Asta Bäck, VTT
Janne Keränen, VTT

Tillsatt av och datum
Kommunikationsministeriet

Publikationsseriens namn och nummer

Kommunikationsministeriets publikationer 7/2017

ISSN (webbpublikation) 1795-4045
ISBN (webbpublikation) 978-952-243-503-3
URN http://urn.fi/URN:ISBN:978-952-243-503-3

Ämnesord
Anonymisering, personuppgiftslagen, affärsverksamhet, persondata

Kontaktperson
Taru Rastas

Rapportens språk
finska

Sammandrag

Syftet med utredningen var att skapa en helhetsbild av vilka typer av alternativa anonymiseringstjänster det i praktiken finns och skulle kunna finnas, särskilt för att man ska kunna göra det möjligt att utnyttja data affärsmässigt.

Som bakgrund utreddes praxis, behov och utmaningar som hänför sig till anonymisering av företags och myndigheters data. Under utredningen kontaktades totalt 28 instanser, varav 15 var företag och resten myndigheter eller forskningsorganisationer. Intervjuerna visar att anonymisering inte används i någon större utsträckning i företagen och att det därför inte egentligen finns några etablerade lösningar för hur anonymisering går till. Många av de intervjuade var intresserade av att utnyttja anonymiserade data och de ansåg att anonymisering är ett viktigt område som ökar i betydelse. Det upplevdes svårt att bedöma huruvida anonymiseringen är tillräcklig, vilket har gjort aktörerna försiktiga när det gäller att erbjuda anonymiserade data. Det finns med andra ord ett behov av exempel och anvisningar som underlättar arbetet i praktiken och främjar utnyttjande av data.

De företag som deltog i undersökningen visade intresse för tjänster för kontroll av anonymiseringens tillräcklighet och experttjänster som stöder egenhändig anonymisering. Att låta en tredje part utföra dataanonymiseringen kändes främmande för många svarspersoner och väckte oro om dataskyddet av och ansvaret för materialet. En anonymiseringstjänst där olika aktörers data kan kombineras innan de anonymiseras skulle dock öppna upp nya möjligheter som uteblir vid företagsspecifik anonymisering.

MINISTRY OF TRANSPORT AND COMMUNICATIONS

Date
3.4.2017

Title of publication
Anonymisation services. Need and implementation alternatives

Author(s)
Asta Bäck, VTT
Janne Keränen, VTT

Commissioned by, date
Ministry of Transport and Communications

Publication series and number
Publications of the Ministry of Transport and Communications 7/2017

ISSN (online) 1795-4045
ISBN (online) 978-952-243-503-3
URN http://urn.fi/URN:ISBN:978-952-243-503-3

Keywords
Anonymisation, Personal Data Act, business, personal data

Contact person
Taru Rastas

Language of the report
Finnish

Abstract

The purpose of the report was to provide an overall view of the types of practical alternatives for anynomisation services there are and could be to enable the utilisation of data, especially in business activities.

As background to the report, companies' and authorities' practices, needs and challenges related to data anonymisation were studied. For this purpose, a total of 28 stakeholders were contacted, 15 of which were companies, while the remainder were authorities or research organisations. The interviews showed that anonymisation is not widely used in companies, and few established solutions for it thus exist. Many of the interviewees were interested in using anonymised data, and they saw this as an important area that is increasing in significance. The respondents found assessing the adequacy of anonymisation difficult, and as a result, the stakeholders are cautious about offering anonymised data. Consequently, examples and guidelines are needed that will support the practical work and promote data utilisation.

The participating companies expressed the most interest in services related to verifying the adequacy of anonymisation and expert services that support in-house anonymisation. Many respondents were uncomfortable about the idea of outsourcing data anonymisation to a third party, as it gives rise to concerns over the information security of the data and the parties' responsibilities. However, an anonymisation service that could combine data sourced from different stakeholders before the anonymisation process would provide new possibilities that cannot be achieved by anonymising the data of an individual company.

Esipuhe

Digitaalisuuden edistäminen on keskeinen hallitusohjelman tavoite. Erityisesti henkilötieto katsotaan digitaalisten palvelujen kehityksen kannalta arvokkaaksi. Henkilötiedon anonymisointitarpeita on tunnistettu kaikilla yhteiskunnan sektoreilla, jotta dataa saataisiin laajemmin ja yhdistellen hyödynnettäväksi. Henkilötietojen käsittely, anonymisointi ja hyödyntäminen datamassoina vaativat osaamista, menetelmiä ja prosessien hallintaa, joita tulee vahvistaa erityisesti liiketoiminnan tarpeisiin.

Selvityksen tavoitteena on tunnistaa liiketoiminnallisia tarpeita henkilötiedon anonymisoinnille. Samoin selvityksen avulla pyritään kuvaamaan toteutusvaihtoehtoja anonymisointipalveluille, joita yritykset voisivat datan laajempaan käyttöön ja anonymisoinnin haastavuuteen peilaten hankkia. Selvityksen tarkoitus on kuvata menetelmiä, prosessia ja käytäntöjä mahdollisille anonymisointiratkaisuille sekä tuoda esille yritysten haasteita datan anonymisoinnissa. Analyysin tuloksina esitetään suosituksia toimenpiteiksi, joilla voitaisiin edistää anonymisoidun datan käyttöä liiketoiminnassa.

Selvitys vahvistaa näkymää, että yrityksillä on tarpeita datan anonymisointiin. Toisaalta haastattelut kertovat sitä, että kolmansien osapuolten anonymisointipalveluille ei nähdä kysyntää. Anonymisointi tehdään nykyisin useimmiten itse omille aineistoille. Odotuksena oli, että tarpeita anonymisoinnille palveluna nostetaan esille datojen yhdistelyn kautta esimerkiksi yritysten välisessä yhteistyössä. Selvitys ei siten välttämättä peilaa tulevaisuuteen niitä mahdollisia liiketoiminnan tarpeita, joita keinoälyn, suurten datamassojen ja alustojen kehittymisen odotetaan vaativan eli eri toimijoiden datojen jakamista ja yhdistämistä arvoverkostoissa ja uusissa yhteistyön muodoissa.

Varovaisuus anonymisoinnin teettämisessä on ilmeinen syy palvelujen kysynnän vähäisyyteen. Ensinnäkin henkilötietoaineiston anonymisointipalvelun käyttäminen edellyttää rekisterinpitäjän ja henkilötietojen käsittelijän oikeudellisten suhteiden järjestämistä selkeällä tavalla. Toiseksi henkilötietodatan anonymisoinnissa peruuttamattomuuden varmistaminen on haaste ottaen huomioon datan loputtomat yhdistelyn mahdollisuudet. Siten anonymisoinnin pitää olla osa jatkuvaa tiedonhallinnan prosessia yrityksissä, jota yhtenäisinä käytänteinä ja osaamiseen panostaen tulee tukea.

VTT on laatinut tämän raportin liikenne- ja viestintäministeriön toimeksiannosta. Selvityksen tekemistä on ohjannut liikenne- ja viestintäministeriö ja Trafi. Selvityksessä esitetyt näkemykset, johtopäätökset ja ehdotukset ovat selvitysten toteuttajien, eivätkä välttämättä sellaisenaan heijasta liikenne- ja viestintäministeriön näkemyksiä. Selvitys tarjoaa yritysten näkemyksiin perustuvaa tietoa henkilötiedon hyödyntämisen mahdollisuuksista hallituksen Digitaalisen liiketoiminnan kasvuympäristö-kärkihankkeelle.

Helsingissä 3 päivänä huhtikuuta 2017

Taru Rastas

Viestintäneuvos

Tieto-osasto

Sisällysluettelo

1. Johdanto

1.1 Tausta

Yritysten ja yhteiskunnan toiminnan digitalisoituminen tuottaa valtavasti dataa. Datojen analysoinnin avulla saadaan ymmärrystä nykyisten palvelujen kehittämiseen ja uusien palvelujen luomiseen. Iso osa syntyvästä datasta liittyy henkilöihin. Tällaisen datan hyödyntämistä voidaan edistää anonymisoinnin avulla. Anonymisoitua dataa voidaan käyttää sekä ratkaisemaan monia yhteiskunnallisia haasteita, kuten terveyteen ja ympäristöön liittyviä haasteita, että kuluttajia palvelevan liiketoiminnan kehittämiseen. Liikkumista ja liikennevirtoja koskevat tiedot ovat yksi konkreettinen esimerkki datasta, jota voidaan anonymisoituna hyödyntää palvelujen toteuttamisessa ja kehittämisessä. Anonymisoitua dataa voidaan käyttää monipuolisesti liiketoiminnan eri toiminnoissa, kuten markkinapotentiaalin analysoinnissa ja strategisessa suunnittelussa.

Kun tietoaineistot sisältävät henkilöihin liittyviä tietoja, datojen käsittelyssä on noudatettava henkilötietolainsäädännön vaatimuksia. Vaatimukset suojelevat perusoikeutta yksityiselämän ja henkilötietojen suojaan ja asettavat reunaehdot data-aineiston hyödyntämiselle. Tietosuojaa koskevia periaatteita ei kuitenkaan sovelleta tietoihin, joita ei voi liittää yksittäisiin henkilöihin, joten henkilötietojen anonymisointi vapauttaa henkilötietojen käsittelyyn liittyvistä vaatimuksista. Tämä antaa uusia mahdollisuuksia datan hyödyntämiseen yritysten liiketoiminnassa, kuten esimerkiksi datojen pitkäaikaiseen säilyttämiseen ja isojen datavarantojen keräämisen. Yrityksen sisäisen käytön lisäksi anonymisoitua dataa voidaan myös jakaa tai myydä ulkopuolisille tahoille tai ostaa ulkopuolelta täydentämään omia aineistoja. Anonymisoitu data tarjoaa siis uutta liiketoimintapotentiaalia.

Anonymisoinnin tarve ja perustelut tulevat hyvin esille EU:n tietosuoja-asetuksen 26. resitaalissa (Euroopan unionin virallinen lehti L119): ”Tietosuojaperiaatteita ei tämän vuoksi pitäisi soveltaa anonyymeihin tietoihin.” Rekisterinpitäjän on otettava huomioon kaikki ”kohtuudella toteutettavissa” (Tietosuojatyöryhmä 2014) olevat keinot rekisteröidyn tunnistamiseksi / anonymisoinnin purkamiseksi, joita joko rekisterinpitäjä tai jokin kolmas osapuoli voi tunnistamiseen käyttää. EU:n tietosuoja-asetuksen 26. resitaali jatkuu seuraavasti: ”Eli tietoihin, jotka eivät liity tunnistettuun tai tunnistettavissa olevaan luonnolliseen henkilöön, tai henkilötietoihin, joiden tunnistettavuus on poistettu siten, ettei rekisteröidyn tunnistaminen ole tai ei ole enää mahdollista. Tämä asetus ei tämän vuoksi koske tällaisten anonyymien, muun muassa tilasto- tai tutkimustarkoituksia varten käytettävien tietojen käsittelyä.”

Anonymisointia on haastavaa suorittaa niin, että yksilöiden tunnistettavuus katoaa, mutta että data on vielä hyödyllistä analyysien ja oikeiden johtopäätösten tekemiseen. Anonymisoinnin vaikeustaso riippuu keskeisesti siitä, minkä tyyppisiä tietoja aineistossa on ja miten monia eri muuttujia se sisältää. Yhä laajemmat, yleisesti saatavissa olevat taustatiedot ja jatkuvasti kehittyvät tiedonlouhintatyökalut muodostavat haasteen luotettavalle anonymisoinnille.

1.2 Tavoitteet ja rajaukset

Tämän selvityksen tavoitteena oli muodostaa kokonaiskuva siitä, millaisia käytännön vaihtoehtoja anonymisointipalveluiksi on ja voisi olla erityisesti anonyymin datan liiketoiminnallisen hyödyntämisen mahdollistamiseksi. Taustaksi selvitettiin, millaisia datan anonymisointiin liittyvät käytännöt ovat, mitkä ovat niiden vahvuudet ja heikkoudet, ja millaisia tarpeita ja haasteita toimijat kokevat anonymisointiin liittyen erityisesti datan liiketoiminnallisen hyödyntämisen näkökulmasta.

Selvityksen tutkimuskysymykset oli tarkennettu seuraavasti:

- Millaisia anonymisointiratkaisuja on käytössä yrityksissä, ja mitkä ovat näiden vahvuudet ja heikkoudet?
- Millaisia anonymisointipalveluja ja työkaluja on yrityksille tarjolla?
- Millaisia datan anonymisointitarpeita on eri toimijoilla ja vastaavatko nykyiset palvelut tai ratkaisut näihin tarpeisiin?
- Miten anonymisointi vaikuttaa datan käytettävyyteen ja millaisilla käyttöluparatkaisuilla ja anonymisoinnin pysyvyyden varmistuksilla käytettävyyttä voidaan edistää?

Selvityksessä tarkasteltiin anonymisointia ja anonymisoidun datan käsittelyä pääasiassa tilanteessa, jossa datan tuottaja ja hyödyntäjä edustavat eri organisaatioita, eikä tarkastelun kohteena ollut datan käsittely yrityksen sisäisissä järjestelmissä.

1.3 Tutkimusmenetelmä

Tutkimuksen pääasiallisena menetelmänä olivat haastattelut. Pääosa haastatteluista tehtiin henkilökohtaisten tapaamisten muodossa. Haastatteluja tehtiin kaikkiaan 19. Haastatteluja täydennettiin lyhyellä kirjallisella kyselyllä, johon saatiin vastaukset 13 yrityksen tai viranomaisen edustajalta. Taulukko 1 kertoo haastateltujen ja kyselyyn osallistuneiden henkilöiden taustaorganisaatiot toimialueen tasolla. Haastatellut henkilöt edustivat organisaatioita, joilla on henkilötietolain tarkoittamaa henkilödataa ja joiden liiketoiminnassa datan hyödyntäminen on keskeistä. Toimialoista olivat edustettuina muun muassa terveydenhoito, mainonta, televiestintä, energia, finanssipalvelut, IT-palvelut, datapalvelut ja liikenne (Taulukko 1). Haastatelluista kahdeksan edusti suurta yritystä, seitsemän edusti pientä yritystä, ja loput olivat viranomaisia, virastoja tai tutkimusorganisaatioita. Haastatteluja täydennettiin kirjallisella kyselyllä. Kaikkiaan näkemyksiä ja tietoja saatiin 28 eri toimijalta.

Haastattelujen ja kyselyn tarkoitus oli selvittää datan anonymisointiin ja anonyymin datan hyödyntämiseen liittyviä tarpeita, käyttökohteita ja haasteita, sekä kerätä tietoa käytössä olevista anonymisointiratkaisuista ja niistä saaduista kokemuksista. Kirjallisuudesta ja verkosta haettiin esimerkkejä datan anonymisointiin käytettävistä ratkaisuista ja palveluista, myös kansainvälisesti.

Taulukko 1. Tutkimukseen haastatellut ja kyselyyn vastanneet henkilöt, yhteensä 28 henkilöä, edustivat 12 eri toimialan yrityksiä ja organisaatioita.

Toimiala	Lukumäärä
Data- ja tietopalvelut	5
IT-palvelut ja ohjelmistot	4
Tutkimus ja kehittäminen	4
Viranomainen tai virasto	4
Teleoperaattori	3
Media ja mainonta	2
Arkistointi	1
Energia	1
Finanssipalvelut	1
Liikenne	1
Sairaanhoito	1
Teollisuuden etujärjestö	1

Selvityksen tulokset esitellään tässä raportissa seuraavasti: luvussa 2 esitellään anonymisointiin liittyvät keskeiset käsitteet ja menetelmät; luvussa 3 annetaan esimerkkejä haastatteluissa esiin tulleista anonymisointikäytännöistä; luku 4 esittää kyselyn ja haastattelujen tuloksia esteiden ja kaivattujen palvelujen osalta ja kuvaa anonymisointiin liittyviä tarpeita neljän erityyppisen tietoaineiston osalta; luvussa 5 esitellään kolme mahdollista tapaa edistää anonymisoinnin käyttöä; ja luvussa 6 esitetään vastaukset tutkimuskysymyksiin ja johtopäätökset.

2. Käsitteet ja menetelmät

2.1 Käsitteiden määrittely

Anonymisointi tarkoittaa prosessia, jossa data-aineistoa muokataan niin, etteivät yksilöt ole siitä tunnistettavissa edes hyödyntämällä epäsuoria tunnisteita tai taustatietoa. Anonymisoidulla datalla tarkoitetaan yleensä mikrodataa, jolloin yhtä kohdetta koskeva tieto on muusta aineistosta erotettavissa olevana yksikkönä. Tällaista dataa voidaan käyttää data-analyysien tekemiseen. Jos data-aineistosta kerrotaan vain tilastollisia tunnuslukuja ja jakaumatietoja, yksilöiden erottaminen ei ole mahdollista, mutta aineiston hyödyntämismahdollisuudet analyyseihin ovat rajalliset.

Aineiston muuttujat jaetaan yksilöiden tunnistamisproblematiikan näkökulmasta kahteen ryhmään: suoriin ja epäsuoriin tunnisteisiin. *Suora tunniste* kertoo välittömästi tai hyvin helposti pääteltävissä olevalla tavalla kenen tiedoista on kyse. Näitä ovat esimerkiksi nimi, sosiaaliturvatunnus tai osoite. *Epäsuora tunniste*, esimerkiksi ikä tai ammatti, ei yksinään paljasta käyttäjää, mutta epäsuorien tunnisteiden yhdistelmästä voi muodostua nopeasti kokonaisuus, joka mahdollistaa yksilön tunnistamisen.

Suorat ja epäsuorat tunnisteet ovat aineistoissa tyypillisesti taustamuuttujia. Näiden lisäksi aineistossa on muita muuttujia, joista vähintään osa voidaan luokitella *sensitiiviseksi eli arkaluonteiseksi tiedoksi*, kuten esimerkiksi diagnosoidut sairaudet. Muuttujien luokittelu on tärkeä osa anonymisointiprosessia. Anonymisoinnin ensimmäisenä vaiheena valitaan, mitä muuttujia käsitellään epäsuorina tunnisteita ja mitä arkaluonteisina. Anonymisoinnissa aineisto muokataan niin, että tämä arkaluontoinen tieto ei pääse paljastumaan.

Taulukko 2. Kuvitteellinen esimerkki anonymisointia vaativasta raakadatasta.

Henkilötunnus	Postitoimi-paikka	Ikä	Sukupuoli	Ammatti	Paino	Verenpaine	Diabetes
111158-1234	02100	58	M	Kapellimestari	85	90 / 145	Kyllä
111186-1233	02100	30	N	Myyjä	75	80/123	Ei

Taulukko 2 näyttää yksinkertaisen kuvitteellisen esimerkin anonymisointia vaativasta aineistosta. Taulukon esimerkkiaineistossa suora tunniste on henkilötunnus, joka on anonymisoinnin yhteydessä poistettava. Arkaluontoiseksi tiedoksi tässä esimerkissä voidaan määritellä tieto diabeteksesta. Anonymisoinnin tavoitteena on estää tämän tiedon paljastuminen ulkopuoliselle. Jäljelle jäävistä muuttujista postitoimipaikka, ikä, sukupuoli, paino ja ammatti ovat epäsuoria tunnisteita, jotka yhdistelmänä voivat johtaa yksilön tunnistamiseen. Jos esimerkki olisi todellinen, niin harvinainen ammatti, kuten kapellimestari, yhdistettynä ikään ja tarkasti rajattuun asuinpaikkaan mahdollistaisi henkilöllisyyden selvittämisen hyvin suurella todennäköisyydellä.

Jotta tämäntapainen aineisto saadaan anonymisoitua, tietojen tarkkuutta vähennetään niin, ettei tunnistaminen ole mahdollista. Taulukon kuvitteellisessa esimerkissä ammatti on määritelty hyvin tarkasti, joten ensimmäinen vaihe on arvioida, onko ammatti datan tulevan käytön kannalta oleellinen tieto. Jos se ei ole, on yksinkertaisinta poistaa tieto kokonaan. Jos tiedon oletetaan olevan tarpeen, on pohdittava, millä tavoin ammatti voidaan karkeistaa niin, että siinä säilyy merkityksellistä tietoa ilman liikaa paljastavuutta. Tunnistettavuutta voidaan heikentää myös muiden muuttujien osalta ja avulla: postinumeron tarkkuutta voidaan vähentää näyttämällä esimerkiksi vain kaksi ensimmäistä numeroa. Ikä puolestaan voidaan esittää viiden tai kymmenen vuoden portaina.

Pseudonymisointi on eri asia kuin anonymisointi. Pseudonymisointi tarkoittaa suoran tunnisteen korvaamista muunnetulla arvolla, josta ei pysty päättelemään alkuperäistä tietoa. Pseudonymisointia käytetään tilanteissa, joissa aineiston sisältämiä tietoja halutaan yhdistellä toisiinsa, tai varaudutaan siihen, että analysoinnissa käytettävää dataa voi tarvita täydentää myöhemmin. Pseudonymisointi estää vain välittömän helpon tunnistamisen, joten se ei tee aineistosta anonyymia. Kuten yllä olevasta esimerkistä nähtiin, pelkkä sosiaaliturvatunnuksen poistaminen tai muuttaminen pseudonyymiksi ei juuri vaikeuta aineiston joidenkin kohteiden tunnistamista. Myös yrityksen sisällä pseudonymisoitua aineistoa on käsiteltävä kuin tunnisteellista aineistoa niin kauan kuin alkuperäiset tiedot ovat tallessa (Tietosuojatyöryhmä 2014. s.9).

2.2 Anonymisointimenetelmät

Kuten edellä ilmeni, anonymisointi tehdään, jotta yksilöt eivät ole tunnistettavissa anonymisoidusta datasta, vaikka apuna käytettäisiin epäsuoria tunnisteita ja taustatietoa. Yksilöiden tunnistamiseen ja yksilöihin liittyvien tietojen paljastumiseen liittyvät riskit luokitellaan kolmeen pääryhmään:

- Henkilöllisyyden paljastuminen – tämä voi tapahtua joko suoraan aineiston pohjalta tai yhdistämällä aineisto muuhun data-aineistoon tai taustatietoon.
- Attribuutin paljastuminen – tämä tarkoittaa, että aineistosta paljastuu tiettyä henkilöä tai henkilöryhmää koskevaa tietoa, vaikka henkilö ei suoranaisesti ole yksilöitävissä aineistosta.
- Aineistoon kuulumisen paljastuminen – tämä tarkoittaa, että voidaan suurella todennäköisyydellä päätellä, että tietyn henkilön tiedot ovat mukana data-aineistossa. Tämä voi puolestaan johtaa esimerkiksi attribuuttien paljastumiseen.

Yksityisyysmalli tarkoittaa aineiston muokkaamisessa käytettävää periaatetta, jolla kohteiden suojaus tehdään. Keskeinen yksityisyysmalli on *k-anonyymisyys,* joka tarkoittaa, että aineistossa on epäsuorien tunnisteiden osalta vähintään k havaintoa, joilla on samat epäsuorat tunnistearvot (Sweeney 2002).

Taulukon 1 esimerkissä k-anonyymisyys arvolla 5 tarkoittaa vaatimuksena, että anonyymissa aineistossa pitää olla vähintään viisi kapellimestaria, jotka asuvat samalla postinumeroalueella, ovat miehiä, sekä lisäksi ovat saman ikäisiä ja -painoisia. Nähdään, että iän ja painon ilmoittaminen 5 tai 10 vuoden portaina helpottaa anonymiteettivaatimuksen täyttämistä, samoin kuin asuinpaikan ilmoittaminen suurempina kokonaisuuksina. Esimerkki havainnollistaa myös sen, että mitä enemmän epäsuoria tunnisteita aineistoon sisältyy, sitä haastavammaksi yksityisyysmallin vaatimusten täyttäminen tulee.

k-anonymisyys-mallia on täydennetty l-diversiteetti ja t-läheisyys -mittareilla, jotka asettavat vaatimuksia sensitiivisiksi määritellyille attribuuteille. *l-diversiteetti* määrittelee, miten suuri hajonta kunkin, k-anonyymiyden avulla muodostettavan ryhmän sensitiivisten attribuuttien arvoissa tulee olla (Machanavajjhala ym. 2006). Jos kaikilla ryhmän jäsenillä on sama sensitiivisen attribuutin arvo, esimerkiksi tietty sairaus, ei arkaluontoisen tiedon paljastumista voi estää, vaikka ryhmän koko olisi hyvinkin suuri. *t-läheisyys* tiukentaa l-diversiteetin vaatimusta siltä osin, että sensitiivisen attribuutin arvojakauma ei saisi poiketa koko väestön vastaavien arvojen jakaumasta (Li ym. 2007).

Hieman toisenlaista lähestymistapaa edustaa *differentiaalinen yksityisyys (*Dwork 2006). Se liittyy anonymisoitujen data-aineistojen tuottamiseen kyselyiden pohjalta. Sen keskeisenä ideana on muokata aineisto sellaiseksi, että kyselytuloksiin ei vaikuta, oli mikä tahansa datajoukon havainto mukana aineistossa tai poissa siitä. Tällöin datasta ei kyselyiden avulla voi paljastua yksittäistä henkilöä koskevaa tietoa. Käytännössä tämä hoidetaan lisäämällä kohinaa menetelmän käyttämille parametreille valittujen arvojen edellyttämä määrä.

Anonymisointimenetelmien kehittämisessä ensisijaisena käyttökohteena on ollut yksittäisiä kohteita kuvaava rividata. Toinen tärkeä datatyyppi on liikkumista koskeva tieto, eli paikan ja ajanhetken kertovat datasarjat. Fyysisen liikkumisen lisäksi vastaavaa dataa syntyy liikkumisesta verkkopalvelujen sivuilla. Erityisesti sijaintipaikkaan liittyvä tieto on helposti yksilöivää ja ulkopuolisten henkilöiden tiedossa lisäten tunnistamisen todennäköisyyttä. Myös tämän tyyppisen tiedon anonymisoinnissa voidaan käyttää k-anonymisyys-yksityisyysmallia: paikan ja ajan tarkkuutta karkeistetaan niin, että identtisiä havaintoja on riittävän paljon, ja harvinaiset, yksilöivät havainnot jätetään aineistosta pois. Ratkaisuksi on myös ehdotettu alueiden määrittelemistä sensitiivisiksi ja ei-sensitiivisiksi, ja sijaintidata näytetään vain siltä osin, kun liikutaan ei-sensitiivisellä alueella (Cicek ym. 2014).

Eräs tapa anonymisoida paikkasidonnaista tietoa on karkeistaa se niin, että sijainti kerrotaan ihmisten alueittaisina lukumäärinä. Kun yksilöllisten havaintojen sijasta ilmaistaan vain kullakin alueella tietyllä aikavälillä olevien henkilöiden kokonaismäärä ja heistä mahdollisesti joitakin tilastollisia lisätietoja, käyttäjien yksityisyys säilyy. Näin anonymisoidun paikkasidonnaisen datan hyödyntämisestä löytyy esimerkki viitteestä Bogomolov ym. (2015).

2.3 Anonymisointiin liittyvää päätöksentekoa tukeva malli

Isossa Britanniassa julkaistiin vuonna 2016 anonymisointiin liittyvää päätöksentekoa tukeva malli ja sen sisältöä perusteleva ja taustoittavan osio (Elliot ym. 2016). Mallin kehittämisen taustaorganisaatio on anonymisointiin keskittyvä verkosto UK Anonymisation Network (UKAN)[1]. Mallin kehittämisen tavoitteena on ollut auttaa toimijoita tarkastelemaan datan anonymisointia kokonaisuutena, jossa tasapainollaan datan anonymisyyden ja hyödyllisyyden välillä. Anonymisointitapaa valitettaessa otetaan huomioon datan tietosisältö ja ennakoitavissa olevat käyttötilanteet ja niistä syntyvät reunaehdot. Raportin laatijoiden mukaan olisikin parempi puhua anonymisointiprosessin läpi käyneestä datasta eikä anonyymista datasta.

[1] http://ukanon.net/

Prosessi perustuu viiteen periaatteeseen:

1. Pelkästään dataa tarkastelemalla ei voi päättää, onko datan julkaiseminen turvallista. Datan lisäksi turvallisuuteen vaikuttaa se, keillä on pääsy dataan ja millä tavalla: luovutetaanko data sellaisenaan vai pääseekö sitä analysoimaan vain tietyssä tietoteknisessä ympäristössä. Tulee ottaa huomioon myös, millainen sopimus datan käyttäjän kanssa tehdään ja millaiset ovat käyttäjän ja käyttöympäristön asenteet.
2. Dataan tutustuminen ja datan tarkastelu ovat tärkeitä: datan jakamisen riskit arvioidaan datan sisältämien tietojen perusteella ja tämän pohjalta voidaan tehdä arvio ja päätökset siitä, mitä dataa, kenelle ja milloin voidaan jakaa.
3. Anonymisointiprosessin tehtävänä on tuottaa turvallista dataa mutta datan on oltava myös hyödyntämiskelpoista, eli anonymisoidun datan tulee heijastaa riittävän tarkasti taustalla olevaa aineistoa. Jos data ei ole käyttökelpoista, ei sitä kannata julkaista, sillä pahimmillaan vääristynyt data johtaa vääriin johtopäätöksiin.
4. Käyttökelpoista anonymisoitua dataa ei voi tuottaa ilman minkäänlaista riskiä. Anonymisointi onkin nähtävä riskinhallintatehtävänä. Tiedottaminen ulkopuolisten tahojen ja erityisesti datakohteiden ja yhteiskunnan suuntaan on tärkeää, jotta kaikki osapuolet osaisivat painottaa oikealla tavalla riskejä ja hyötyjä.
5. Anonymisoinnin tekemisessä käytettävien menetelmien ja käytäntöjen tulee olla oikeassa suhteessa riskiin ja riskin toteutumisen vaikutuksiin. Jos data on hyvin yksityiskohtaista ja sisältää arkaluontoista tietoa, pääsyn rajoittaminen dataan on perusteltua, ja päinvastoin, jos datassa ei ole arkaluontoista tietoa, niin esimerkiksi sen käytön rajoittaminen vain turvalliseen käyttöympäristöön vaikuttaa ylisuojaamiselta.

Ehdotetun mallin mukainen anonymisointiprosessi koostuu kaikkiaan kymmenestä vaiheesta. Alussa täsmennetään datalle kaavaillut käyttötilanteet ja hankitaan ymmärrys lakien asettamista velvoitteista, ja tutustutaan luovutettavan datan sisältöön. Tärkeä osa prosessia on datan jakamiseen liittyvien riskien arviointi ja riittävien, muttei liioiteltujen suojausmenetelmien valinta. Riskiarviointiin liittyy myös viestintä- ja kriisiviestintäsuunnitelman tekeminen.

Tärkein viesti, jonka mallin laatijat haluavat tuoda esiin on, ettei ole olemassa yhtä ainoaa toimintatapaa tai anonymisoinnin riittävyyden arvioimisen laskentakaavaa. Päätökset tarkoituksenmukaisesta anonymisoinnin tavasta ja tasosta tulee tehdä ottaen huomioon sekä datan sisältö että miten ja missä ympäristössä dataa tullaan käyttämään. UKANin julkaiseman raportin mukaan k-anonymiteettiin pohjautuva yksityisyysmalli on selkeä keino arvioida anonyymisyyttä, mutta tarkastelua ei tule rajoittaa pelkästään tähän näkökulmaan. Ongelmana on erityisesti se, että korkealle asetettu tavoite, eli k:lle annettu korkea arvo, johtaa tyypillisesti suureen informaatiohävikkiin ja aineiston hyödyllisyyden jyrkkään laskuun.

2.4 Anonymisointityökaluista

Anonymisoinnin vaatimaa aineiston muokkausta voidaan tehdä tavanomaisilla tilastollisilla ohjelmistoilla, mutta tarjolla on myös erityisohjelmistoja, joista annetaan kaksi esimerkkiä. Nämä esimerkit eivät ole suosituksia, vaan vain esimerkkejä anonymisointiin liittyvästä teknologiakehityksestä ja työkaluista.

μ-Argus[2] on avoimeen lähdekoodiin perustuva ohjelmistotyökalu, joka on kehitetty rividatan anonymisointiin tilastollisia menetelmiä käyttäen. Se kehitettiin alun perin eurooppalaisessa “Computational Aspects of Statistical Confidentiality” -projektissa. Ohjelmistoa voi käyttää interaktiivisessa ja automaattisessa tilassa. Ohjelma tuottaa lokitiedoston, johon tallentuvat datalle tehdyt muunnostoimenpiteet. Ohjelmisto on mm. Alankomaiden tilastokeskuksen käytössä.

ARX Data Anonymization Tool[3] on avoimeen koodiin perustuva työkalu, joka mahdollistaa datan anonymisoinnin (Prasser ym. 2014). Työkalu tukee useiden yksityisyysmallien käyttöä ja mahdollistaa eri ratkaisuvaihtoehtojen vertaamisen jäljelle jäävän tunnistamisriskin ja anonymisoidun datan informaatioarvon suhteen. Työskentely etenee kolmessa vaiheessa: konfigurointivaiheessa valitaan käytettävät yksityisyysmallit, määritellään mitä muuttujia tulee käsitellä epäsuorina tunnisteina ja mitä sensitiivisinä, ja annetaan pohjatiedot muuttujien karkeistamiselle; toisessa vaiheessa ohjelma tarjoaa joukon vaihtoehtoja anonymisoinnin tekemiselle; kolmannessa vaiheessa tarkastellaan valitun anonymisointiratkaisun tuottamaa lopputulosta datahävikkiä, uudelleentunnistusriskiä ja datan käyttöarvoa kuvaavien tunnuslukujen avulla. Vaiheiden välillä voi liikkua iteratiivisesti.

[2] http://neon.vb.cbs.nl/casc/mu.htm
[3] http://arx.deidentifier.org/

3. Suomalaisia anonymisointikäytäntöjä

Hankkeen puitteissa tehtyjen haastattelujen perusteella syntyi käsitys, että Suomessa anonymisoinnin tekee yleensä toimija itse omalle datalleen. Tampereen Yliopiston yhteydessä toimiva Tietoarkisto (FSD)[4] on ainoa hankkeen yhteydessä tunnistettu toimija, joka tekee anonymisointia ulkopuoliselle datalle. Heidän tehtävänään on tutkimusaineistojen anonymisointi, joten kyseessä ei ole varsinainen kaupallinen palvelu.

Tilastokeskus on yrityksiä koskevaa liiketaloudellista tietoa hallinnoiva viranomainen[5]. Näiden tietojen toissijainen hyödyntäminen edellyttää yritysten identiteetin suojaamista henkilöiden identiteetin suojaamista vastaavalla tavalla. Tilastokeskuksella on siten osaamista anonymisointien tekemiseen ja tätä palvelua voidaan tarjota myös ulkopuolisille tahoile.

Anonymisoinnissa käytäntönä on, että tunnistamisen estämiseksi yritys- ja henkilötietoja sisältävissä aineistoissa karkeistetaan muuttujia kuten ammatti, toimiala, asuinpaikka ja työpaikan sijainti. Yrityksen tulos pyöristetään ja liikevaihto ilmoitetaan vain suuruusluokkana, ei tarkkana arvona.

Tilastokeskus tarjoaa dataa ulkopuolisille tahoille kolmella tavalla. Nämä tavat eroavat toisistaan sekä sen suhteen, miten paljon dataa on prosessoitu kohteiden tunnistamisen estämiseksi, miten kattava aineisto voi olla, ja missä olosuhteissa ja millä edellytyksillä siihen pääsee käsiksi. Tavat ovat seuraavat:

1. Julkiset, vapaasti jaettavissa olevat aineistot (Public Use File (PUF))

    - On ensisijaisesti tarkoitettu opetuskäyttöön, ei varsinaiseen tutkimukseen.
    - Sisältää vain otoksen koko aineistosta.
    - Aineistoon on saatettu lisätä kohinaa tunnistamisen hankaloittamiseksi. Aineisto vastaa tilastollisilta ominaisuuksiltaan oikeaa dataa, mutta kaikki yksittäiset arvot eivät ole totuudenmukaisia.
    - Aineisto on lähellä synteettistä aineistoa, eli todellista aineistoa vastaavaa, laskennallisesti tuotettua aineistoa.

2. Tieteelliset aineistot (Scientific Use File (SUF))

    - Aineiston saajan on haettava käyttölupa.
    - Aineisto voidaan luovuttaa tutkijalle analysoitavaksi tutkijan omassa ympäristössä.
    - Annetaan vain otos koko aineistosta.

3. Etäkäyttöympäristön kautta tarjottava pseudonymisoitu aineisto

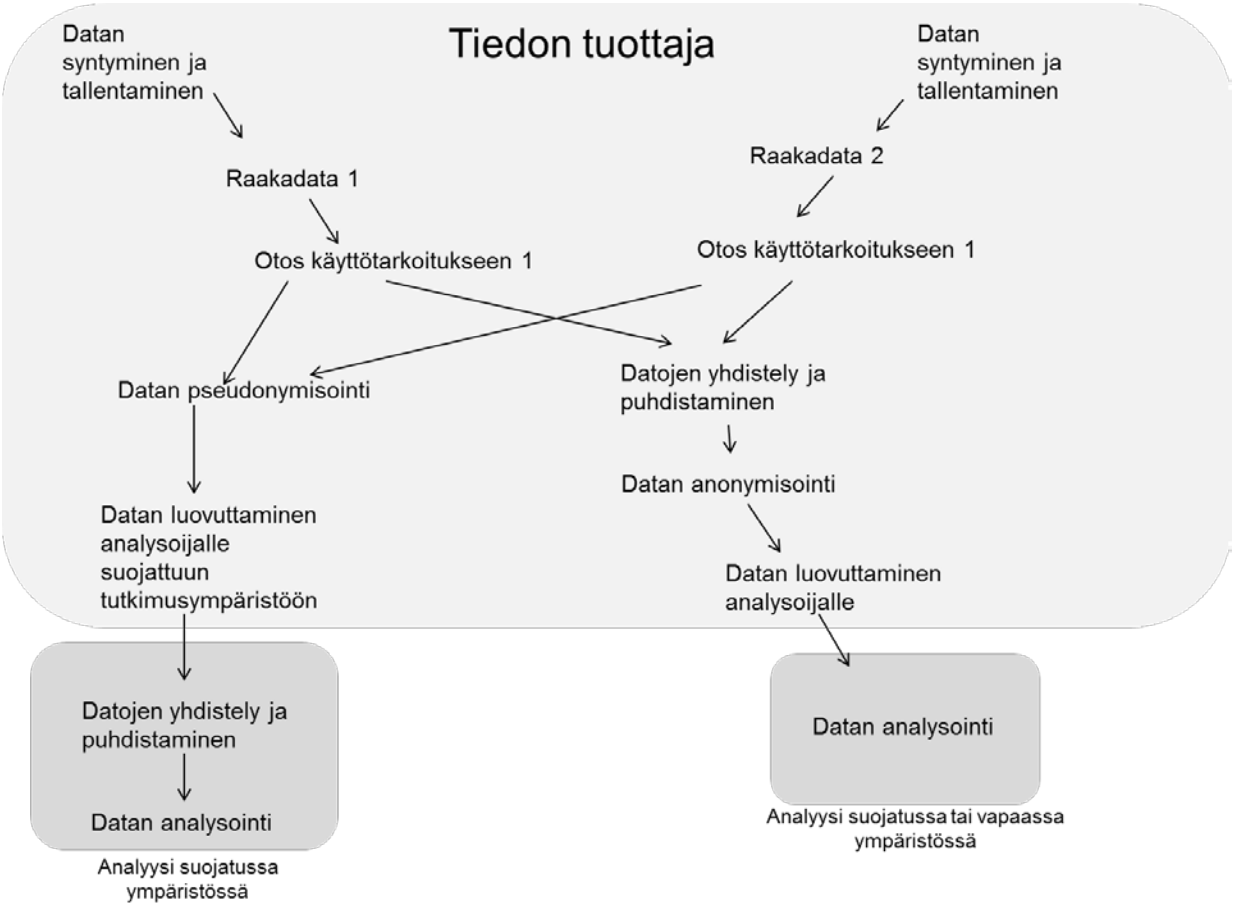
    - Suora tunnistaminen on estetty, mutta epäsuora tunnistaminen voi olla mahdollista.
    - Voi sisältää koko aineiston.
    - Käyttö turvallisessa etäkäyttöympäristössä, jossa tietojen siirtämistä ei sallita; analyysien tekemiseen on tarjolla ohjelmistoja, kuten SAS ja R.

[4] http://www.fsd.uta.fi/fi/
[5] https://tilastokeskus.fi/tup/mikroaineistot/aineistot.html

- Tietojen käyttäjä allekirjoittaa sopimuksen, jossa hän sitoutuu vain luvan mukaiseen aineiston käyttöön ja muihin aineiston käsittelyä ja tuhoamista koskeviin ehtoihin.

Tilastokeskus painottaa omassa toiminnassaan pseudonymisoinnin, sopimusten ja turvallisen ympäristön yhdistelmää aineistojen anonymisoinnin sijasta. Perusteluina tälle on, että pseudonymisointi ei heikennä aineiston laatua toisin kuin anonymisointi. Pseudonymisointi myös mahdollistaa, että aineistoa voidaan täydentää helposti lisäaineistolla, jos tutkimuksen kuluessa tähän nousee tarve.

Tilastokeskus pystyy tarjoamaan pseudonymisoidun aineiston tutkijoille anonymisoitua aineistoa edullisemmin. Kuva 1 esittää kahta eri työnkulkua sen mukaan, luovutetaanko data pseudonymisoituna vai anonymisoituna. Prosesseissa datan puhdistamiseen ja yhdistämiseen liittyvä työ sijoittuu eri toimijoiden tehtäväksi: koska pseudonymisoidussa aineistossa on tunnisteet, datan vastaanottaja pystyy yhdistämään ja puhdistamaan aineistoa, jolloin datan luovuttajalta vaadittu työpanos voi jäädä pieneksi. Anonymisoidun aineiston osalta datan luovuttajan tulee tehdä tämä työ. Mikäli samaa aineistoa tarvitaan toistuvasti, niin tiedon tuottajan tekemä aineiston puhdistaminen mahdollistaa kustannustehokkuuden parantamiseen.



Kuva 1. Kaavio datan käsittelyvaiheista ja suorittajista, kun data luovutetaan pseudonymisoituna (vasen puoli) tai anonymisoituna (oikea puoli).

Tilastokeskuksen asiantuntijat toivat haastattelussa esiin anonymisoinnin tekemiseen liittyviä haasteita: aineiston muuttujamäärän kasvaessa anonymisoinnin tekeminen tulee yhä vaikeammaksi. Käytännössä tämä johtaa siihen, ettei aineistoa välttämättä pystytä tekemään niin laajaksi kuin mihin mahdollisuuksia olisi. Yksi anonymisoinnissa käytössä oleva menetelmä on satunnaisvirheiden tuottaminen muuttujiin siten, että muuttujien tilastolliset ominaisuudet pysyvät ennallaan. Tämän tekeminen niin, että muuttujien väliset riippuvuudet pysyisivät ennallaan, on kuitenkin vaikeaa. Ellei anonymisoinnin tekemiseen ole monistettavaa lähestymistapaa, anonymisoinnin tekeminen vaatii paljon työtä, ja tulee siten kalliiksi.

Toisena esimerkkinä anonymisointia tekevästä viranomaisesta voidaan mainita *Trafi*[6], joka hallinnoi mm. ajoneuvoihin liittyvää tietoa. Ajoneuvotietoja tarjotaan ulkopuolisille kahden pääkanavan kautta: toinen on julkaiseminen avoimena datana ja toinen luovuttaminen hyödynnettäväksi käyttöluvan perusteella. Käyttöluvan ehdot ja käyttörajoitteisen tiedon sallitut käyttötarkoitukset on säädetty laissa. Avoimena datana julkaistavalle ajoneuvoaineistolle tehdään anonymisointi. Aineistoista poistetaan suoraksi tunnisteeksi tulkittu rekisterinumero, ja aineiston tietoja karkeistetaan mm. näyttämällä postinumerosta vain kolme ensimmäistä merkkiä ja valmistenumerosta 10 ensimmäistä merkkiä. Aineisto julkaistaan rivimuodossa, jolloin sitä voi analysoida ja hyödyntää monipuolisesti.

[6] http://www.trafi.fi/tietopalvelut/trafin_rekisterit

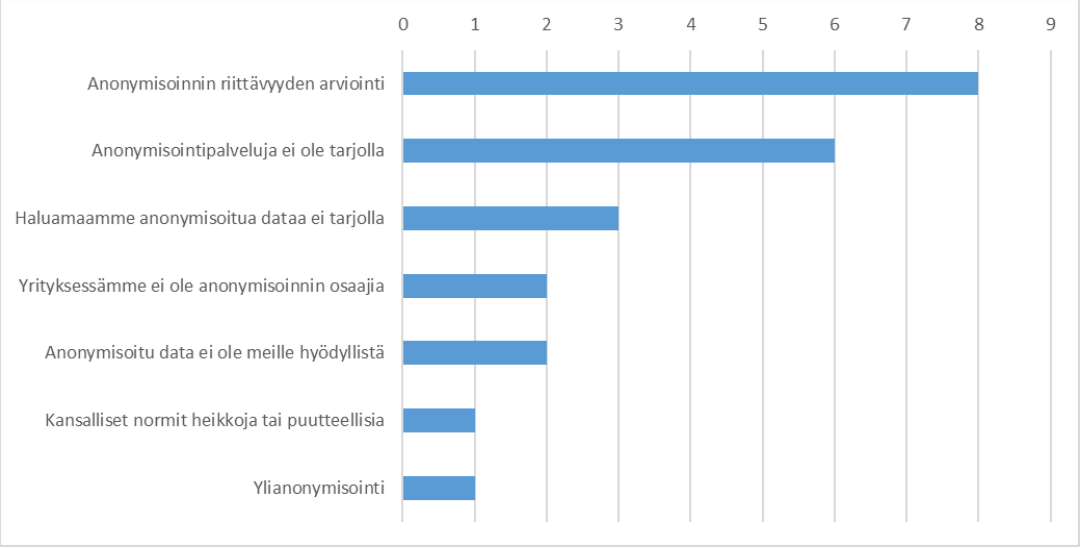
4. Anonymisointitarpeet

Tässä luvussa esitetään haastattelujen ja kyselyjen tuloksia siltä osin kuin ne liittyvät toivottuihin palveluihin ja eri tietotyyppeihin liittyviin erityiskysymyksiin. Ensimmäiseksi esitetään kyselyssä saadut vastaukset, jotka koskivat anonymisoinnin hyödyntämisen esteitä ja anonymisoinnin käytön edistämiseksi kaivattuja palveluita. Kyselyvastaukset saatiin 13 henkilöltä. Vastaukset olivat samansuuntaisia kuin haastatteluissa esiin nousseet näkemykset, joten kyselyn tulokset kuvastavat kaikkien haastateltujen näkemyksiä, vaikka kyselyyn vastanneiden määrä ei tätä suurempi ollutkaan. Kyselytulosten esittelyn jälkeen käydään läpi tärkeimpiin tietotyyppeihin liittyviä näkökulmia.

4.1 Esteet ja toivotut palvelut

Kyselyssä annettiin lista haastatteluissa esiin nousseita esteitä, ja vastaajia pyydettiin kertomaan, mitkä niistä he kokivat esteinä. Esteitä sai valita useamman kuin yksi ja tarpeen mukaan myös lisätä uusia. Kuva 2 esittää tulokset graafin muodossa. Anonymisoinnin riittävyyden arviointi koettiin suurimmaksi esteeksi. Lähes yhtä yleisesti koettiin, ettei tarvittavia anonymisointipalveluita ole tarjolla; kaksi vastaajaa myös koki, ettei heidän yrityksessään ole tarpeeksi osaamista anonymisoinnista.

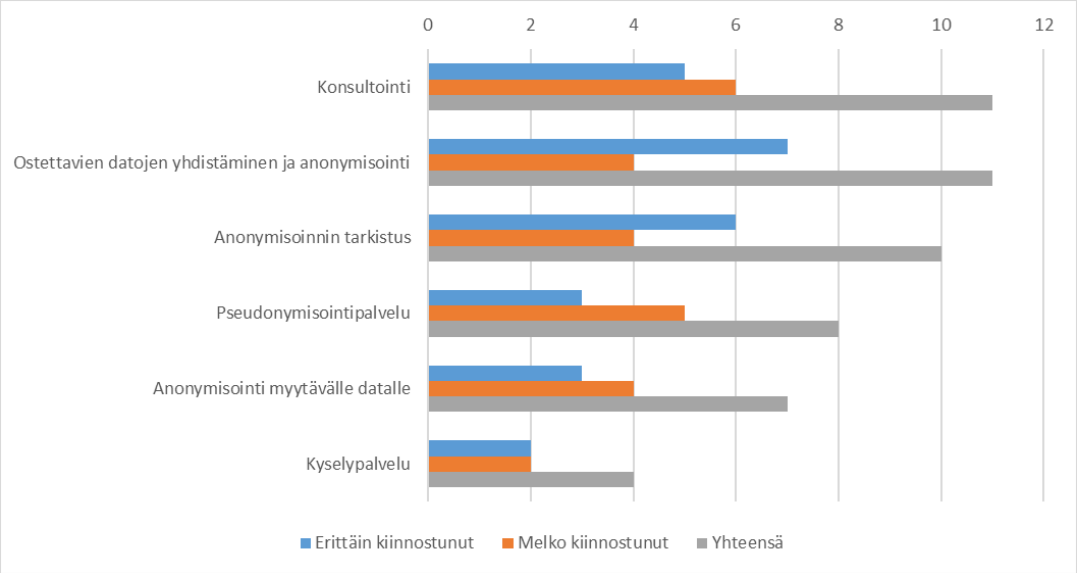
Datan hyödyntämisen osalta koettiin jossain määrin esteenä, ettei haluttua anonymisoitua dataa ole tarjolla. Yksi vastaaja mainitsi esteeksi ylianonymisoinnin. Hän tarkoitti tällä, että anonymisointikriteerit on jossain tapauksissa asetettu niin tiukoiksi, että anonymisoidun aineiston kattavuus jää heikoksi. Kaksi vastaajaa totesi, ettei anonymisoitu data ole yritykselle hyödyllistä.



Kuva 2. Kyselyyn vastanneiden 13 henkilön mainitsemat anonymisoinnin käytön esteet.

Kuva 3 kertoo, miten vastaajien kiinnostus kohdistui erityyppisiin palveluihin. Anonymisoinnin toteuttamista tukevat konsultointipalvelut kiinnostivat kahta vastaajaa lukuun ottamatta kaikkia. Sama tilanne oli myös suhteessa palveluun, jonka kautta voisi tilata eri tahojen hallinnoimaa dataa yhdistettäväksi ja anonymisoitavaksi. Myös anonymisoinnin riittävyyden tarkistaminen kiinnosti, mikä olikin odotettua, koska tämä mainittiin usein esteeksi anonymisoinnin käyttämiselle.

Palvelu, jonne voisi antaa omia datojaan anonymisoitaviksi ennen asiakkaille toimittamista, kiinnosti selvästi vähemmän kuin palvelu, jonka kautta dataa voisi ostaa. Kiinnostus pseudonymisointipalvelua kohtaan oli samaa luokkaa kuin oman myytävän datan anonymisointia tekevää palvelua kohtaan. Kyselyssä annetuista valmiista vaihtoehdoista pienintä kiinnostus oli datavarantoihin tehtäviä kyselyjä tarjoavaa palvelua kohtaan.



Kuva 3. Kyselylomakkeella tarjottujen, anonymisointiin liittyvien palvelujen kiinnostavuus (13 vastaajaa)

4.2 Terveys- ja hyvinvointitieto

Terveydenhuollon alueella syntyy ja kerätään paljon tietoa, jonka hyödyntäminen on sekä yhteiskunnallisesti tärkeää että liiketoimintanäkökulmasta kiinnostavaa. Potentiaalia on palveluiden kehittämisen saralla ja erilaisten hoitotoimenpiteiden ja lääkkeiden tehokkuuden arvioinnissa. Tietoja tuottavat julkiset ja yksityiset toimijat, sekä myös kansalaiset itse hyödyntäen terveyden ja aktiivisuuden seuraamiseen kehitettyjä laitteita ja sovelluksia.

Vuoden 2018 alussa on astumassa voimaan lakiuudistus sosiaali- ja terveystietojen tietoturvallisesta hyödyntämisestä (Hallitus 2016). Lailla tavoitellaan terveydenhuollon prosesseissa syntyvän tiedon entistä laajempaa hyödyntämistä. Hyväksytyiksi käyttökohteiksi on ehdotettu:

- tutkimus,
- kehittäminen ja innovaatiot,
- opetus,
- tietojohtaminen,
- ohjaus ja valvonta, ja
- suunnittelu, selvitykset ja tilastot

Sitran rahoittamassa ja vetämässä Isaacus-hankeessa[7] kehitetään uusia toimintamalleja ja työkaluja terveystiedon toissijaisen käytön helpottamiseksi. Isaacus-hankkeen edustajien haastattelussa ilmeni, ettei anonymisointi ole ollut hankkeessa erityisen keskeisessä roolissa, vaan tietojen toissijaiseen käyttöön liittyvässä työssä ensisijaisena ratkaisuna nähdään pseudonymisoinnin ja tietoturvallisen tutkimusympäristön yhdistelmä. Tämä vastaa nykytilannetta, jossa on tyypillistä, että aineistot luovutetaan tutkijoille pseudonymisoidussa muodossa.

Terveyden- ja hyvinvoinnin laitoksen edustajan haastattelun yhteydessä ilmeni, että jonkin verran toimitaan myös niin, että asiakkaalle tuotetaan asiakkaan haluamat kyselyvastaukset, eikä luovuteta dataa itse tehtävää analysointia varten, eli tällöin ei tarvita anonymisointia. Tätä mallia käytetään erityisesti teollisuuden suuntaan, kun kyseessä ei ole varsinainen tieteellinen tutkimus. Malli sopii, kun vastaanottajalla on täsmälliset kysymykset.

Kyselypohjaisen toimintamallin laajentaminen onkin yksi pohdinnassa oleva asia. Vaihtoehto antaa henkilötiedoille hyvän turvan, mutta vaatii selvitystä siitä, olisiko palvelu tarjottavissa riittävän ketterästi ja kilpailukykyiseen hintaan. Tällaisen palvelun tarjoaminen edellyttää sekä hyvää substanssiosaamista että tilastollisten menetelmien hallintaa. Palvelun toteuttaminen vaatii myös panostamista aineistojen puhdistamiseen, eli palvelun käynnistäminen vaatii investointeja.

Toinen ilmentymä tämän tyyppisestä palvelusta voisi olla mahdollisuus oman datan vertaamiseen laajempaan aineistoon. Tämä voisi kiinnostaa kansalaisia heidän itse keräämiensä terveys- ja hyvinvointitietojen osalta.

Lääketeollisuus hyödyntää terveydenhuoltosektorin rekistereihin kertyvää tietoa, ja rekisteritietoihin pohjautuvien tutkimusten tekeminen onkin kasvussa, koska vallalla on suuntaus hakea aiempaa enemmän reaalimaailman todisteita lääkkeiden vaikutuksista. Lääketeollisuus ry:n edustajan mukaan lääketeollisuus ei kovin usein tarvitse yksilötasolla olevaa aineistoa, vaan datan pohjalta lasketut aggregoidut tulokset usein riittävät, kunhan on varmuus siitä, että laskenta on tehty oikean datan pohjalta.

Lääketeollisuuden tutkimuksen kannalta on tärkeää, että dataan käsiksi pääsemisen edellyttämä lupaprosessi on nopea, ja että ennen kaikkea sen kesto on ennakoitavissa luotettavasti. Useat haastatellut henkilöt toivatkin esiin, että nykyisin lupa tarvitaan erikseen jokaiselta rekisterinpitäjältä, mikä tekee prosessista hankalan ja pitkäkestoisen. Lääketeollisuus ry:n edustajan mukaan Suomeen kaavailtu uusi laki sosiaali- ja terveystietojen tietoturvallisesta hyödyntämisestä ja Isaacus-hanke ovat oikean suuntaisia, sillä tavoitteena on lupaprosessin yksinkertaistaminen. Jos terveystietojen hyödyntämisen vaatima lupaprosessi saadaan sujuvaksi, kansainväliset toimijat voivat kasvattaa Suomessa tehtävien tutkimustensa määrää.

Avoimen innovaation hackathoneissa voidaan hyödyntää anonymisoitua dataa. Esimerkki tästä on Slush 2016 -tapahtuman yhteydessä järjestetyssä Junction hackatonissa tarjottu aktiviteettimittaustieto. Aineiston anonymisoinnin tehnyt henkilö kertoi haastattelussa anonymisoinnin yhteydessä tehdyistä päätöksistä seuraavasti. Data päädyttiin julkaisemaan summaten suoritukset päivä- ja viikkotasoille, eikä paljastamalla esimerkiksi yksittäisten urheilusuoritusten tarkkoja ajankohtia. Urheilulajikohtainen tieto annettiin vain, mikäli aineistossa oli kyseiselle lajille vähintään 30 harrastajaa. Kaikki loput lajit yhdistettiin. Tietoja ei myöskään annettu siitä, millä urheilulaitteella mittaukset oli tehty. Tietoja olisi ollut tarjolla

[7] http://www.sitra.fi/hyvinvointi/hyvinvointidata

laajemminkin, mutta anonymisoinnin riittävyyden varmistamiseksi päädyttiin suhteellisen suppeaan muuttujamäärään. Hackathonin osallistujat allekirjoittivat myös sopimuksen, jossa he sitoutuivat käyttämään dataa vain kyseisen hackathonin yhteydessä.

Aktiviteettiaineiston anonymisoinnin tehnyt henkilö näki anonymisoidun datan hyödyntämisen esteinä anonymisoinnin riittävyyden arvioinnin sekä anonymisointiosaamisen ja anonymisointiin liittyvien palvelujen puuttumisen. Palvelu, jossa datoja voisi yhdistää ja anonymisoida olisi hänen mukaansa kiinnostava; tämä palvelu voisi myös tuottaa tilastodataa. Myös anonymisointiin liittyvät konsulttipalvelut ja palvelut, joissa datan anonymisyyden voi tarkistaa, kiinnostivat haastateltua.

4.3 Liikenne ja liikkuminen

Liikkumiseen ja liikenteeseen liittyvää reaaliaikaista dataa tarvitaan esimerkiksi välittämään tietoa ruuhkista ja hetkellisistä esteistä. Pitemmän aikavälin tieto liikennevirroista puolestaan palvelee erilaisia suunnittelutehtäviä. Kun tieto kulkureiteistä on ihmisten tai ajoneuvojen tarkkuudella, syntyy henkilötietolain tarkoittamaa henkilödataa.

Ajoneuvojen mukana kulkevilla mobiililaitteilla kerätään tietoa liikenneolosuhteista ja tieverkoston kunnosta. Tietoa kerätään systemaattisimmin kuljetusyritysten kalustoa hyödyntäen. Tähän dataan liittyvä haaste on, että auton tietojen perusteella välittyy tieto myös auton omistajasta tai kuljettajasta. Anonymisointi on pulmallista erityisesti, kun kerätään tietoa harvaan asutuilla ja vähän liikennöidyillä alueilla.

Liikennepalveluita tarjoavan toimijan haastattelussa puolestaan ilmeni, että liikennepalvelujen tarjoajien kannalta olisi tärkeää saada tietoa ihmisten kulkureiteistä alusta loppuun. Kuljetuspalvelujen tarjoajat pystyvät seuraamaan käyttäjämääriä omien kuljetusvälineidensä osalta mutta todellisen liikkumistarpeen ymmärtämiseksi olisi hyödyllistä saada tieto koko kulkureitistä alusta loppuun. Kuluttajat voisivat luontevasti kerätä tällaista tietoa itse mobiilisovelluksilla ja halutessaan antaa se palveluntarjoajien hyödynnettäviksi.

Liikenteen datapalveluja tarjoavan yrityksen edustaja kertoi haastattelussa kokevansa, että anonymisoinnin riittävyyden arvioinnin vaikeus on este anonymisoinnin käytölle. Datan anonymisointiin liittyvät palvelut kiinnostivat tätä vastaajaa jonkin verran, vähiten ulkopuolinen datan anonymisointipalvelu.

Lupakäytäntöjen osalta on ollut ongelmia tapauksissa, joissa infrastruktuureihin liitettyjen kamerajärjestelmien keräämää dataa on haluttu hyödyntää. Kyseisissä tapauksissa kuvaus tapahtuu julkisella paikalla avoimesti, usein liikennetoimijan taholta, mutta niin, että videokuvaan tallentuu tunnistettavia kasvoja ja rekisterinumeroita. Ongelmana on, että tällaisen datan keräämisen ja käytön laillisuus on epäselvää eikä ennakkotapauksia ole.

Julkaistavissa kuvissa näkyvien henkilöiden kasvojen tai autojen rekisterinumeroiden käsittelemisessä voidaan hyödyntää konenäkömenetelmiä. Automaattiset algoritmit pystyvät tunnistamaan henkilöiden kasvoja ja autojen rekisteritunnuksia ja sumentamaan ne tunnistamattomiksi.

Liikkumiseen liittyvää tietoa kertyy myös teleoperaattoreille, mutta sitä on käsitelty erillisessä kappaleessa (katso kappale Mobiilidata).

4.4 Mobiilidata

Mobiiliverkkotoimijoille kertyy asiakkaistaan massiiviset määrät henkilöihin liittyvää dataa. Datan avulla ohjataan käytännön operatiivista toimintaa ja dataa analysoidaan asiakasymmärryksen saamiseksi sekä tuotteiden ja palveluiden kehittämiseksi. Esimerkiksi palvelutason takaaminen mobiiliverkoissa edellyttää datan tallennusta ja siitä tehtävää suorituskykyanalyysiä. Tämä järjestelmätason data anonymisoidaan aggregoinnin avulla palvelun laadun seuraamiseksi. Näin selviää, miten erityyppiset asiakasryhmät tai palvelut suoriutuvat mobiiliverkossa muihin verrattuna.

Mobiiliverkkotoimijat anonymisoivat dataa omaan käyttöönsä siten, että hakuvaiheessa tunnistetaan ja poistetaan datasta suoraan henkilöön liittyvät attribuutit. Pelkkä henkilötietojen pseudonymisointi ei riitä peruuttamattomaan anonymisointiin, sillä esimerkiksi laitteen ja sijainnin yhdistelmä voi mahdollistaa henkilön tunnistamisen. Yksi haastateltavista mainitsi, että mobiiliverkkotoimijat joutuvat tekemään jatkossa enenevissä määrin anonymisaatiota, sillä mobiilidataa joudutaan mahdollisesti julkaisemaan organisaation ulkopuolelle, esimerkiksi ihmisten liikkuvuusdataa valtion ja kaupunkien käyttöön. Mobiiliverkkolaitetoimittajat sisällyttävät usein anonymisoinnin järjestelmiinsä, joten operaattoreille on tarjolla valmiita työkaluja. Yksi haastateltava totesi myös että kolmansien yritysosapuolten osaamista anonymisointimenetelmien kehittämisessä on käytetty

Mobiiliverkkotoimijat eivät myy dataansa ulkopuolisille toimijoille, vaikka kiinnostusta on ilmennyt. Datan myyntiin liittyen on sekä lainsäädäntöön että kaupalliseen potentiaaliin liittyviä avoimia kysymyksiä. Suomessa viestintäsalaisuus on säädetty vahvaksi, mikä velvoittaa teleoperaattoreita suojaamaan datan hyvin ja tekee toimijat varovaisiksi datan toissijaisen käytön suhteen. Yksi operaattoreista vetosi viestintäsalaisuuden turvaamiseen myös ulkopuolisen anonymisointipalvelun käytön osalta, eli hän ei pitänyt mahdollisena datan luovuttamisesta ulkopuoliselle edes anonymisointia varten.

Kaupalliseen potentiaaliin liittyen datan tarjoamisen vaatimat investoinnit mietityttävät, ja asiassa etenemiseksi tarvittaisiin vahva näkemys riittävän markkinan olemassaolosta. Mobiilidatan käyttö populaatio- ja liikkumisanalyyseihin on yksi esimerkki liiketoiminta-alueesta, jossa mobiilidatan hyödyntämisen kilpailukykyä on verrattava muihin tapoihin saada vastaava tieto. Yksi haastatelluista, operaattoria edustavista henkilöistä totesi liiketoimintapotentiaalin arvioinnissa tärkeäksi näkökulmaksi myös kilpailulliset tekijät: onko realistista, että datan myynnistä saatavat tulot ovat suuremmat, kuin menetykset siitä, että kilpailijat saavat nykyistä laajempaa dataa käyttöönsä.

Mobiiliverkkotoimijat, samoin kuin muut kaupalliset yritykset, ostavat dataa erityisesti asiakasymmärryksen parantamiseksi ja markkinointi- ja kehitystoimenpiteiden tueksi, esimerkiksi ostokäyttäytymisluokitusten ja asiakasprofiilien tekemiseksi. Dataa ostetaan muun muassa Tilastokeskukselta ja mediataloilta; myös sosiaalisen median kanavista koostettua tietoa ostetaan. Suoratoistopalvelujen ja sosiaalisen median käyttäjädata auttaa asiakkaiden käyttäytymisen ymmärtämisessä ja kohderyhmien tunnistamisessa. Myös paikannukseen liittyvää ja sitä tukevaa kartta-aineistoa ostetaan. Toimijoita kiinnostaa myös ei-anonyymi data ja sen yhdistäminen esimerkiksi avoimeen dataan.

Mobiiliverkkotoiminnan sopimusehdot määrittelevät, mitä asiakasdatalle saa tehdä. Internetin, sosiaalisen median ja suoratoistopalvelujen kehitys on muuttanut sen, mitä asiakkaan dataa käytetään ja jaetaan kolmannen osapuolen kanssa. Kumppanuudet operaattoreiden ja mediatalojen välillä usein vaativat tietojen luovuttamisen anonyyminä kolmannelle osapuolelle (esimerkkinä suoratoistopalveluihin liittyvä

asiakaskäyttäytymisdata). Lupakäytännöissä epäselvyyttä aiheuttaa, mitkä asiakastiedot pitää rajata ulos datan hyödyntämisen yhteydessä.

4.5 Kuluttajaliiketoiminnassa syntyvä tieto

Kuluttajaliiketoiminnan piiristä nostetaan esimerkeiksi mainonta ja finanssiala.

Sähköisessä mainonnassa mainosten oikea kohdistaminen on keskeistä. Jos palvelua käytetään sisään kirjautuneena, yksittäisen käyttäjän seuraaminen on helppoa, ja kyse on selkeästi henkilötiedosta. Usein palveluita kuitenkin käytetään kirjautumatta, ja tällöin yksittäistä käyttäjää seurataan evästeiden avulla. Käytössä on myös mainosverkostojen omia tunnisteita, joiden toimintatapa vastaa evästeitä.

EU:n tietosuoja-asetuksen 30. resitaalissa todetaan seuraavaa: ”Luonnolliset henkilöt voidaan yhdistää heidän käyttämiensä laitteiden, sovellusten, työkalujen ja protokollien verkkotunnistetietoihin, kuten IP-osoitteisiin, evästeisiin tai muihin tunnisteisiin, esimerkiksi radiotaajuustunnisteisiin. Näin käyttäjästä voi jäädä jälkiä, joita voidaan käyttää luonnollisten henkilöiden profilointiin ja tunnistamiseen etenkin, kun niitä yhdistetään yksilöllisiin tunnisteisiin ja muihin palvelimille toimitettuihin tietoihin.”

Evästeiden ja IP-numeroihin perustuva selailutieto tulkitaan siis tietosuoja-asetuksen alaiseksi henkilötiedoksi. Evästeitä tarvitaan, kun halutaan yhdistää yksittäiset sivulataukset selailupoluiksi. Haastateltu media-alan edustaja totesi, että tämän tiedon osalta anonymisointi ei ole käyttökelpoinen ratkaisu, sillä ilman evästeitä pystytään käytännössä ainoastaan laskemaan sivulatauksien määriä. Sähköisen viestinnän tietosuoja-asetusta ollaan parhaillaan uudistamassa, ja siinä määritellään media-alan kannalta keskeisiä kysymyksiä evästeisiin, sijaintitietojen hyödyntämiseen ja sähköiseen suoramarkkinointiin liittyen. Komission ehdotus direktiiviksi julkaistiin tammikuussa 2017[8].

Haastateltu media-alan edustaja pohti, olisiko verkkomainonnan aineistojen hallintaan mahdollista perustaa pseudonymisointipalvelu, joka takaisi, että evästeet muutetaan kiistattomasti sellaisiksi pseudotunnisteiksi, joita mediatalo ei voi enää muuntaa takaisin alkuperäiseen muotoon. Näin pystyttäisiin selkeästi osoittamaan, etteivät tiedot ole enää yhdistettävissä nimettäviin henkilöihin, mutta tietojen hyödynnettävyys saataisiin säilytettyä. Pseudonymisoinnin hyödyntämistä mainosalalla anonymisoinnin sijasta ehdotetaan myös kirjoituksessa ”Pseudonymisation: What the Ad Tech Industry Needs to Know” (Rowntree 2016).

Evästeiden ohella toinen henkilön tunnistamiseen johtava tieto on käyttöpaikan sijainti. Verkko- ja mobiilikäytön yhteydessä paikka voidaan määrittää suhteellisen tarkasti, varsinkin jos käyttäjä hyödyntää paikkaan sidottuja täsmäpalveluita, kuten säätietoa. Mediatalon edustaja kertoi, että kiinnostusta mainonnan kohdistamista tukevaan datayhteistyöhön mediatalojen ja mainostajien väillä on paljon, ja varsinkin paikkatiedon pohjalta tehtävä kohdentaminen kiinnostaa mainostajia. Eri osapuolten evästepohjaisia datoja yhdistäviä hankkeita leimaa kuitenkin suuri varovaisuus, sillä pelätään, että tietoja yhdistettäessä kohteiden henkilöllisyys voi paljastua. Mediatalon kokemus datan jakamisesta onkin, että hankkeet vaativat perusteellisen suunnittelun, johon kummankin osapuolen asiakkaiden

[8] https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications

yksityisyyden huolehtimisesta vastaavien tiimien on osallistuttava, eivätkä tällaiset yhteistyöt etene kovinkaan nopeasti.

Hyvin perinteinen, mutta edelleen käytössä oleva esimerkki anonyymin datan käytöstä on alueellinen demografiatieto, jota on saatavissa esimerkiksi Tilastokeskukselta. Tätä käyttävät niin mediatalot kuin mainostajatkin rikastamaan tietoaan asiakas- ja käyttäjäkunnastaan.

Finanssialan edustajan vastauksessa tuli esiin kiinnostus datatuotteiden myyntiin tulevaisuudessa, erityisesti vuonna 2018 voimaan astuvan maksupalveludirektiivin, PSD2[9] myötä. Myös datan ostaminen kiinnosti vastaajaa, sekä reaaliaikaisena että pidemmän aikajakson aikana. Alueelliset, ikäryhmittäiset tai vastaavat koostetulokset eivät hänen mukaansa kuitenkaan ole yleensä riittäviä.

Finanssialan operatiivisessa toiminnassa tarvitaan useimmiten tunnistettuun käyttäjään liittyvää tietoa. Tämän johdosta vastaajaa askarrutti, voisiko datan pseudonymisointia hyödyntää tukemaan eri datalähteiden datojen yhdistämistä toisiinsa, ja näin saada hyödyllistä tietoa hyvän asiakaskokemuksen takaamiseksi.

Verkottuneessa toimintaympäristössä olisi käyttöä anonymisoidulle datalle, jota yhteistyökumppanit voisivat käyttää palveluita kehittäessään. Vastaaja kokikin, että olisi tarve palvelulle tai työkalusetille, joka tuottaisi aidon oloista testidataa yrityksen rajapintapalveluiden testiympäristöihin kehittäjäkumppanien käyttöön. Tämä voisi perustua joko hyvään anonymisointiin tai datan generointiin tarkan mallin perusteella.

Esteinä datan anonymisoinnille finanssialan vastaaja koki sekä anonymisoinnin riittävyyden arvioinnin että oman osaamisen puutteet. Palvelu, jossa voisi tarkistuttaa datan anonymisoinnin riittävyyden, ja anonymisointiin liittyvät konsultointipalvelut kiinnostivat. Myös palvelu, jonka kautta voisi tilata eri tahojen hallinnoimaa dataa yhdistettäväksi ja anonymisoitavaksi omaan käyttöön kiinnosti vastaajaa. Sitä vastoin palvelu, johon voisi luovuttaa omat aineistot anonymisoitaviksi, ei kiinnostanut vastaajaa juuri lainkaan, kuten ei myöskään muualla olevaan dataan kohdistuvat kyselypalvelut.

Yksi haastateltu liiketoimintainformaation parissa toimiva taho ilmaisi että lupakäytännöt on koettu hankaliksi tapauksissa, joissa on epäselvää, miten tietoja saa käyttää esimerkiksi verkkomainontaan liittyen. EU:n tietosuoja-asetukseen kaivataan myös kansallista yhtenäisyyttä: henkilötietojen käsittelyyn liittyvää sääntelyä on olemassa, mutta asetuksen käytännön sovellustapaa ei tiedä tällä hetkellä oikein kukaan, mikä aiheuttaa ongelmia liiketoiminnan suunnittelussa ja ennakoinnissa. Lupakäytäntöjen suhteen keskeisin haaste on tieto siitä, mikä on sallittua ja mikä ei. Lupaselvittelyt saattavat viedä paljon aikaa, vaikka data olisikin anonymisoitua.

[9] http://vm.fi/maksujarjestelmat

5. Anonymisointipalveluiden vaihtoehdoista

Tässä luvussa esitellään haastattelujen pohjalta nousseet kolme päävaihtoehtoa, joiden avulla anonymisoidun datan tuottamista ja hyödyntämistä voidaan edistää. Vaihtoehdot eivät ole toisensa poissulkevia, vaan niillä kaikilla voi olla rooli joillakin osa-alueilla. Vaihtoehdot ovat:

- Palvelu, jonne anonymisointia vaativa data voidaan luovuttaa anonymisoitavaksi,
- Yrityksissä ja organisaatioissa tehtävää anonymisointia tukevat asiantuntijapalvelut mukaan lukien anonymisoinnin riittävyyden arviointi, ja
- Palvelu, jonka kautta voi saada vastauksia anonymisointia vaativaan dataan kohdistettuihin kyselyihin.

5.1 Erillinen anonymisointipalvelu

Yksi haastattelujen ja kyselyjen yhteydessä esillä ollut vaihtoehto oli dataa tuottavista ja dataa hyödyntävistä toimijoista erillinen taho, jolle voisi toimittaa aineistot anonymisointia varten. Tällainen toimija voisi myös toimia aineistojen yhdistäjänä niin, että anonymisointi tehdään vasta, kun kaikki kulloinkin halutut datat on yhdistetty. Datan ei siten tarvitsisi olla peräisin yhdeltä organisaatiolta.

Ajatuksellisesti tällainen palvelu on ihanteellinen, mutta sen käytännön toteuttamisessa on joitakin haasteita. Anonymisointipalvelun tulee saada anonymisoitavat aineistot itselleen, ja mikäli aineistoja olisi yhdisteltävä ennen anonymisointia, mukana tulee olla yhdistämisen mahdollistavat tunnisteet. Esimerkiksi haastateltu teleoperaattorin edustaja ei pitänyt tätä mahdollisena, koska viestintäsalaisuuden turvaaminen voisi näin vaarantua.

Anonymisoinnin tekeminen vaatii ymmärrystä aihealueesta, johon data liittyy, sekä raakadatan sisällön että sen pohjalta tehtävien analyysien osalta. Tekijällä tulee olla tietoa saatavissa olevista tausta-aineistoista, joiden olemassaolo tulee ottaa huomioon tunnistusriskiä arvioitaessa. Luonnollisesti tarvitaan myös osaamista itse anonymisointimenetelmistä ja tilastotieteestä. Myös lainopillista osaamista tarvitaan anonymisyyden riittävyyden arviointiin. Palvelun tarjoajalla tulee olla osaamista monista eri sovellusalueista, tai vaihtoehtoisesti voisi olla monta toimijaa, joista kukin erikoistuisi jollekin sovellusalueelle. Koska anonymisointipalvelun tuottajan tulee käsitellä hyvin luottamuksellisia tietoja, sen itsensä toiminta tulee olla hyvin varmennettua ja ulkopuolisen arvioitsijan hyväksymää.

Kirjallisuudessa on esitetty malleja, miten anonymisointipalvelu voitaisiin järjestää. Eräs vaihtoehto, Privacy and Anonymization as a Service (PASS), on tarjota integroitu kokonaisuus yksityisyydenhallintapalveluita, johon on yhdistetty vaadittavat anonymisointityökalut, ja metodit yksityisyyden suojaamisen onnistumisen arvioimiseen (Heyrani-Nobari ym. 2010). Tämä kokonaisuus sijaitsisi pilviarkkitehtuurissa, ja Web -palvelut toimisivat kyseisen systeemin rajapintoina ja käyttöliittymänä. PASS käsittelee datan prosessointia projekteina, ja jokainen projekti on kokoelma datasettejä ja niitä koskevia prosessointiohjeita kuten algoritmeja tai työvuokokonaisuuksia. PASS referenssitoteutus on

jaettu datapalveluihin, käyttäjäpalveluihin, ja kontrollipalveluihin, jotka voidaan tavoittaa SOAP ja REST -protokollien avulla.

Ulkomainen esimerkki datan anonymisointipalvelun tarjoajasta on CROS NT[10], joka on kliiniseen dataan keskittyvä tutkimusorganisaatio. CROS NT auttaa kliinisiä kokeita tekeviä yrityksiä Euroopassa käsittelemällä niiden kliiniset tiedot luotettavasti ja jäljitettävissä olevalla tavalla poistamalla potilaiden tunnistetietoja. CROSS NT auttaa toteuttamaan kliinisen datan alan standardeja ja antaa tukea sekä aineistojen jäljitettävyyden että analysoinnin saralla. CROSS NT pyrkii auttamaan yrityksiä siinä, miten parhaiten valmistautua saattamaan kliiniset aineistot julkisesti saataville EU:n tietojen avoimuutta koskevan lainsäädännön vaatimusten mukaan. CROS NT tarjoaa teknologisia työkaluja kliinisen datan käsittelyyn, mukaan lukien visualisointityökalut, analyysityökalut ja datan tallennuspalvelut.

5.2 Anonymisoinnin konsultointi- ja sertifiointipalvelut

Anonymisoinnin riittävyyden arviointi nousi haastatteluissa useimmin koetuksi ongelmaksi ja esteeksi. Palvelut, joista saa tukea anonymisoinnin suorittamiseen ja anonymisointitulosten riittävyyden arviointiin, vastaisivat tähän ongelmaan. Tässä tapauksessa datoja ei tarvitse luovuttaa ulkopuolisille, vaan datan tuottaja tekisi anonymisoinnin, ja ulkopuolinen taho vain arvioisi anonymisoinnin tuloksen riittävyyden. Taho, joka olisi erikoistunut datan anonymisointiin liittyviin kysymyksiin ja anonymisyyden riittävyyden arviointiin, vähentäisi anonyymin datan tarjoamiseen liittyvää epävarmuutta ja edistäisi anonyymin datan tarjontaa. Haittana on, että mikäli jokainen osapuoli anonymisoi omat datansa, yhtä yritystä laajempi datojen yhdistely ei ole mahdollista.

Anonymisointien tekemistä ja halukkuutta anonymisoitujen aineistojen myymiseen ja julkaisemiseen edistäisi se, että tarjolla olisi selkeitä esimerkkejä siitä, millainen taso anonymisoinnissa on riittävä, ja tarjolla olisi ohjeistusta anonymisointituloksen arviointiin. Montaa osapuolta kiinnostavien kysymysten ja data-aineistojen pohjalta tuotetut esimerkit vähentäisivät pelkoja ja tuottaisivat konkreettista apua anonymisoinnin tekijöille.

Tieto teollisuuden hyvistä käytänteistä anonymisoinnin tekemisessä olisi myös avuksi. Selkeät mallit ja pelisäännöt siitä, miten anonymisointi tulee tehdä, takaisi myös dataa myyville yrityksille reilut, yhteneväiset toimintamallit.

Jossain määrin anonymisointiin liittyviä konsultointipalveluita on tarjolla Suomessakin. Esimerkiksi kaksi haastatelluista datapalvelujen tarjoajista kertoi tekevänsä, tai ainakin pystyvänsä tarjoamaan yksityisyyden hallintaan tai anonymisointiin liittyvää konsultointia. Myös Tilastokeskuksella on asiaan liittyvää osaamista.

Esimerkkinä varsinaisesta anonymisoinnin sertifiointipalvelusta voidaan mainita Kanadalainen Privacy Analytics, joka keskittyy anonymisointiin liittyvän uudelleentunnistusriskin arviointiin ja anonymisointiprosessien sertifiointiin. Sertifioinnin yhteydessä konsultit käyvät läpi aineiston ja anonymisointiprosessin ja varmistavat, että prosessi on linjassa lainsäädännön kanssa. Arviointi on riskipohjainen ja pyrkii maksimoimaan anonymisoidun datan arvon.

Isossa Britanniassa toimiva UKAN on julkaissut anonymisoinnin vaatimaa päätöksentekoa tukevan prosessimallin, joka esiteltiin lyhyesti luvussa 2.3. Mallin kehittämisen tavoitteena on

[10] http://crosnt.com/

ollut auttaa toimijoita tarkastelemaan datan anonymisointia kokonaisuutena, ja löytämään kussakin käyttötapauksessa tasapaino datan anonymisyyden ja hyödynnettävyyden välillä. UKAN tarjoaa myös anonymisointiin liittyvää tietoa ja järjestää halukkaille työpajoja ja klinikoita; UKANin sivuilta löytyy myös tietoja alueen palvelutarjonnasta Isossa Britanniassa[11]. Tarjolla olevat palvelut ovat pääosin asiantuntija- ja konsultointipalveluita.

5.3 Kyselyihin perustuvat palvelut

Kolmas esiin noussut vaihtoehto ovat kyselyihin perustuvat palvelut. Tässä mallissa dataa ei luovuteta, vaan tarjotaan mahdollisuus kyselyjen ja vertailujen tekemiseen. Tällaisen palvelun tekeminen vaatii, että datan omistaja jalostaa datan muotoon, jonka pohjalta kyselyihin voi vastata nopeasti. Vaikka tämän tyyppisessä palvelussa ei suoraan luovuteta dataa ulkopuolisille, myös kyselyjen antamia tuloksia pitää seurata, jotta niiden avulla ei paljastu yksilöiviä tietoja. Palvelun luominen vaatii investointeja, joten se sopii tapauksiin, joissa on nähtävissä toistuvaa kysyntää, ja dataa hallitseva taho näkee tämän kiinnostavana uutena liiketoiminta-alueena. Palvelu voisi periaatteessa myös yhdistää useamman toimijan datoja yhden rajapinnan taakse.

Aircloak[12] on esimerkki kansainvälisestä anonymisointiratkaisun tarjoajasta. Heidän teknologiansa lupaa pääsyn reaaliaikaiseen dataan anonyymilla tavalla täyttäen EU:n tietosuojasäädösten (GDPR) vaatimukset. Aircloakin ratkaisu ei perustu pseudonymisointiin tai k-anonymiteettiin, vaan se toimii dynaamisena välikätenä muokaten sensitiiviseen dataan kohdistuvat tietokantakutsut anonymiteetin varmistavalla tavalla. Aircloak antaa esimerkkeinä soveltuvista käyttökohteista älykkään kaupungin / liikenteen, lääketieteellisen tutkimuksen, ja geolokaatiodatan liiketoiminnallisen hyödyntämisen.

[11] http://ukanon.net/external-resources/
[12] https://www.aircloak.com/

6. Tulokset ja johtopäätökset

Selvitykselle oli asetettu neljä tutkimuskysymystä, ja seuraavassa annetaan niihin vastaukset kootun aineiston pohjalta. Tulokset perustuvat tutkimuksen aikana haastateltujen tahojen esittämiin näkemyksiin ja vastauksiin.

Ensimmäinen kysymys koski sitä*, millaisia anonymisointiratkaisuja yrityksissä on käytössä*, ja mitkä ovat näiden *vahvuudet ja heikkoudet*. Selvityksen aikana saatiin tietoja kaikkiaan 28 eri tahon edustajilta. Haastattelut osoittivat, ettei anonymisointi ole yrityksissä laajasti käytössä, joten vain muutamia esimerkkejä anonymisoinnin käytännön toteuttamisesta tuli esiin.

Anonymisoinnin käytön esimerkit hyödynsivät yleisesti tunnettuja tapoja heikentää aineiston havaintojen tunnistettavuutta. Keinoina mainittiin muuttujien arvojen karkeistaminen tarkoista arvoista suuruusluokiksi, tarkkojen kellonaikojen poistaminen, harvinaisten havaintoarvojen yhdistäminen yhdeksi yhdistelmäarvoksi, ja sijaintitiedon karkeistaminen.

Tilastokeskuksella oli haastatelluista eniten kokemusta anonymisoinnin tekemisestä ja myös anonymisoinnin haasteista. Anonymisoinnin tekeminen on vaativaa ja se muuttuu sitä vaativammaksi, mitä enemmän aineistossa on muuttujia. Aineiston hyödyllisyys laskee nopeasti sen mukaan, mitä enemmän muuttujien arvoja joudutaan karkeistamaan. Myös aineiston laajuutta, joko muuttujien tai havaintojen lukumääränä mitattuna, voidaan joutua rajoittamaan, jotta anonymisyys saadaan riittävälle tasolle. Hyödyntämisen kannalta parhaaseen lopputulokseen päästään, kun anonymisointi tehdään ottaen huomioon datan tuleva käyttötarkoitus ja siinä keskeisimmät muuttujat. Haittoina ovat vaadittava työpanos ja sen mukana tulevat kustannukset, ellei ratkaisua päästä monistamaan.

Toinen kysymys koski sitä, *millaisia anonymisointipalveluja ja työkaluja* yrityksille on tarjolla. Palveluita, joissa anonymisoidaan asiakkaiden dataa, ei haastattelussa suoranaisesti tullut esiin. Tilastokeskus pystyy tarjoamaan anonymisointia myös muihin kuin omiin aineistoihin, samoin kuin anonymisointiin liittyvää konsultointia. Myös yrityspuolelta tuli esiin muutama toimija, jolla on anonymisointiin liittyvää osaamista ja jotka voivat tarjota siihen liittyviä asiantuntijapalveluita.

Anonymisoinnin tekemiseen on kehitetty avoimen lähdekoodin ohjelmistoja, joista on annettu muutama esimerkki myös tässä raportissa. Aineiston anonymisointia voi sinänsä tehdä millä tahansa datan käsittelyyn tarkoitetulla ohjelmistolla, mutta nimenomaan anonymisointiin kehitetyt ohjelmistot sisältävät toimintoja vaihtoehtojen ja niiden tuottamien tulosten helppoon vertailemiseen. Myös kaupallisia palveluita on tarjolla kansainvälisesti, ja niistä on annettu esimerkkejä tässä raportissa.

Kolmas kysymys liittyi eri toimijoiden *datan anonymisointitarpeisiin* ja siihen, *vastaavatko nykyiset palvelut ja ratkaisut näihin tarpeisiin*. Datan anonymisointi herätti monessa haastatellussa henkilössä kiinnostusta, ja he näkivät alueen tärkeänä ja merkitykseltään kasvavana. Konkreettisten käyttöesimerkkien antaminen tuotti kuitenkin vaikeuksia.

Selkeä, esiin noussut käyttöalue anonyymille datalle oli käyttö avoimen innovaation yhteydessä arvoverkostoissa ja hackatoneissa. Anonymisoidun datan vaihtoehtona voisi myös olla palvelu, jossa generoidaan todellista dataa edustavaa keinotekoista dataa. Näin luodun datan tulisi kuvata riittävällä tarkkuudella todellisia havaintoja tai tapahtumia, jolloin sitä voisi käyttää datapohjaisia palveluita kehitettäessä. Tämä on analoginen esimerkki

Tilastokeskuksen käytännölle: julkinen avoin data on tarkoitettu opiskelijoille käytettäväksi analyysimenetelmien opettelemisen yhteydessä, mutta varsinaiseen tutkimukseen suositellaan tutkimusluvan takana olevaa pseudonymisoitua aineistoa.

Datan anonymisoinnin teettäminen kolmannella osapuolella tuntui monesta vastaajasta vieraalta, ja joissakin tapauksissa yrityksillä oli selvä kanta, etteivät he voi luovuttaa aineistojaan ulkopuoliselle taholle edes anonymisontia varten. Datan alkuperäisellä omistajataholla on viimekäden vastuu datojen lainmukaisesta käsittelystä, ja siksi dataa ei haluta luovuttaa. Eräs haastateltu myös huomautti, ettei ulkopuolista tahoa uskallettaisi käyttää, ellei sen toiminta ole hyvin valvottua ja sertifioitua.

Ulkopuolista anonymisointipalvelua kiinnostavampana yritykset pitivät anonymisoinnin riittävyyden tarkistamiseen liittyviä ja itse hoidettavaa anonymisointia tukevia asiantuntijapalveluita. Kaivataan selkeitä esimerkkejä ja vertailukohtia, joita voi käyttää anonymisointia tehtäessä ja anonymisoinnin riittävyyttä arvioitaessa. Yksi vastaajista ehdotti kansallisten työkalujen, normien ja prosessien kehittämistä anonymisoinnin tason arvioinnin tueksi. Esimerkkien ja toimintamallien avulla yrityksiä voidaan rohkaista anonyymin datan tarjoamisessa, mikä osaltaan edistäisi datapohjaisen liiketoiminnan laajentumista. Selkeät esimerkit ja toimintamallit loisivat myös yhteiset pelisäännöt, ja takaisivat yhtäläiset edellytykset dataa tarjoavien yritysten keskinäisessä kilpailussa. Esimerkkejä tarvitaan erityyppisistä data-aineistoista niiden erilaisen problematiikan johdosta; tässä raportissa esillä on ollut muun muassa terveys- ja hyvinvointitiedot, sijaintitiedot, kuvat ja videot, verkon selailutiedot ja ostotapahtumia koskevat tiedot, joihin kaikkiin liittyy omia erityispiirteitä.

Vaikeus anonymisyyden riittävyyden arvioinnissa heijastuu tällä hetkellä varovaisuutena aineistojen hyödyntämisessä oman yrityksen ulkopuolella. Pelätään myös, että yhdistettäessä omia aineistoja yhteistyökumppanin aineistoihin, yksittäiset ihmiset voivat paljastua. Yksi vastaaja näki, että ylianonymisointi on este datapohjaiselle liiketoiminnalle. Hän tarkoitti, että anonymisyyden turvaamiseksi asetetaan niin korkeat vaatimukset, että dataa tulee tarjolle vain vähän. Anonymisyyden riittävyydelle ei ole olemassa yksinkertaista mittaria, vaan riittävyys tulee arvioida paitsi datasisällön osalta myös ottaen huomioon datan ennakoidut käyttötilanteet ja se, miten arkaluontoista tietoa käsitellään. Kyse on siis riskiarvioinnin tekemisestä.

Hajallaan olevien käyttäjäkohtaisteen tietojen kokoaminen laajasta datamassasta ei onnistu ilman tunnistetietoja, joten yhdistäminen tulisi tehdä ennen anonymisointia. Maantieteelliset alueet toimivat tällä hetkellä tapana jakaa anonyymia tietoa niin, että sillä pystytään rikastamaan omasta asiakas- tai käyttäjäkunnasta olevaa tietoa. Keskusteluissa nousi esiin ajatus, voisivatko pseudonymisointiratkaisut tukea data-aineistojen yhdistämistä niin, että pystyttäisiin tuottamaan käyttökelpoisempaa anonyymia dataa.

Anonyymin datan tarjonnan kasvu myös edellyttää, että yrityksillä on myönteinen näkemys datan hyödyntämisestä liiketoiminta-alueena. On voitettava pelko siitä, että data, vaikka se olisi anonymisoitua tai aggregoitua, hyödyntäisi kilpailijoita enemmän kuin mitä etuja on datan laajemman saatavuuden kautta itselle tuotettavissa.

Neljäs ja viimeinen kysymys oli, *miten anonymisointi vaikuttaa datan käytettävyyteen ja millaisilla käyttöluparatkaisuilla ja anonymisoinnin pysyvyyden varmistuksilla käytettävyyttä voidaan edistää*. Anonymisointi heikentää datan käytettävyyttä, kuten on tullut esille aiempien kysymysten yhteydessä. Pääongelmat liittyvät aineiston yksityiskohtaisuuden vähenemiseen vaaditun karkeistuksen ja muiden anonymisoinnin vaatimien toimenpiteiden takia.

Terveys- ja taloustietojen tutkimuskäytössä on pääsääntöisesti päädytty anonymisoinnin sijasta pseudonymisoidun aineiston käyttöön, jotta tulokset perustuvat paikkansapitävään dataan. Pseudonymisoinnin lisäksi tietoja suojataan käyttöluvan ja käyttötavan määrittelevän sopimuksen avulla ja enenevässä määrin rajaamalla aineiston käyttö tiettyyn, turvalliseen käyttöympäristöön. Myös yritysten välisessä datan hyödyntämisessä voidaan ottaa tästä mallia: anonyymi data ei tarkoita samaa kuin avoin data, vaan data voi olla maksullista, ja sen käyttöön voi liittyä ehtoja ja rajoituksia. Kunhan mahdollisten käyttölupien saaminen ja sopimuksia tekeminen on sujuvaa, ei niistä tule ylivoimaista estettä datan hyödyntämiselle.

Tässä raportin tulokset perustuvat suhteellisen pienen vastaajajoukon näkemyksiin. Vaikka vastauksissa toistuivat samat teemat, olisi ollut hyvä, jos vastauksia olisi saatu tätä laajemmalta joukolta. Haastattelut painottuivat tiedon tuottajiin, jotka tietysti itsekin voivat myös olla tiedon käyttäjiä, mutta olisi ollut hyvä, jos anonymisoidun tiedon potentiaalisia hyödyntäjiä olisi tavoitettu enemmän. Tätä voidaan pitää selvityksen merkittävimpänä rajoituksena. Asetettu aikataulu ja budjetti kuitenkin rajoittivat selvityksen laajuuden nyt käsillä olevaan.

Tulosten pohjalta voidaan esittää seuraavat johtopäätökset. Monet haastatelluista olivat kiinnostuneita anonymisoidun datan hyödyntämisestä, mutta anonymisoinnin riittävyyden arviointi koetaan vaikeaksi, ja tämä tekee toimijat varovaisiksi anonymisoidun datan tarjoamisessa. Jotta tämä este saataisiin poistettua, on tarpeen tuottaa esimerkkejä ja ohjeita, jotka auttavat käytännön työssä. Esimerkkejä tarvitaan erityyppisistä datoista, koska niiden anonymisointiin liittyy toisistaan poikkeavia haasteita.

Anonymisoidun datan tuottaminen vaatii usein paljonkin työtä, kuten datan puhdistamista ja jalostamista ennen anonymisointia. Tunnistamalla usein kysyttävät data-aineistot ja niihin kohdistuvat analyysitarpeet saadaan mahdollisuus kehittää toistettavissa olevat, tehokkaat anonymisointiprosessit. Pseudonymisoinnin hyödyntämistä ennen varsinaista anonymisointia kannattaa myös selvittää mahdollisuutena helpottaa data-aineistojen yhdistämiseen liittyvää problematiikkaa ja lisätä datan arvoa.

Anonymisointiin liittyvän palvelutarjonnan osalta kiinnostusta ja kysyntää on erityisesti anonymisointia tukeville asiantuntijapalveluille. Tämä on luontevaa, sillä eteneminen datan hyödyntämisessä vaatii, että toimijoilla on riittävä ymmärrys anonymisointiin liittyvistä kysymyksistä. Erillinen anonymisointipalvelu, jossa eri toimijoiden datoja pystyttäisiin yhdistämään ennen niiden anonymisointia, ei saanut haastatelluilta varauksetonta kannatusta. Tällainen palvelu antaisi kuitenkin uusia mahdollisuuksia datapohjaisen liiketoiminnan kehittämiseen, sellaisia, jotka jäävät saavuttamatta, mikäli rajaudutaan yrityskohtaiseen anonymisointiin.

7. Lähdeluettelo

Cicek, A. E., Nergiz, M. E & Saygin, Y. (2014) Ensuring location diversity in privacy-preserving spatio-temporal data publishing. The VLDB Journal (2014) 23:609-625.

Dwork, C. (2006) Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP) (2) (2006), 1–12.

Elliot, M., Mackey, E., O'Hara, K. & Tudor, C. (2016) The Anonymisation Decision - Making Framework. Published by UKAN, University of Manchester. 171 s. http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

Euroopan unionin virallinen lehti L119 (2016). http://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=FI

Hallitus (2016) Hallituksen esitys laiksi sosiaali- ja terveystietojen tietoturvallisesta hyödyntämisestä sekä eräiksi siihen liittyviksi laeiksi. http://stm.fi/documents/1271139/3091050/Luonnos-HE--Sote-tietojen-tietoturvallinen-hy%C3%B6dynt%C3%A4minen.PDF/7e6eb683-437f-4fbd-9684-82d8032a9d5b

Heyrani-Nobari G., Boucelma O., Bressan S. (2010) Privacy and Anonymization as a Service: PASS. In: Kitagawa H., Ishikawa Y., Li Q., Watanabe C. (eds) Database Systems for Advanced Applications. DASFAA 2010. Lecture Notes in Computer Science, vol 5982. Springer, Berlin, Heidelberg

Li, N., Li, T. & Venkatasubramanian, (2007) t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. 2007 IEEE 23rd International Conference on Data Engineering, 2007. 106-115.

Machanavajjhala, A, Gehrke, J., Kifer, D. & Venkitasubramaniam, M. (2006) ℓ-diversity: Privacy keyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), p. 24, 2006

Prasser, F, Kohlmayer, F., Lautenschlaeger, R. & Kuhn, K. A. (2014) ARX – A Comprehensive Tool for Anonymizing Biomedical Data. Proceedings of the AMIA 2014 Annual Symposium, November 2014, Washington D.C., USA

Rowntree, L. (2016) Pseudonymisation: What the Ad Tech Industry Needs to Know. Exchange Wire. https://www.exchangewire.com/blog/2016/10/17/pseudonymisation-ad-tech-industry-needs-know/

Sweeney, L. (2002) Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

Tietosuojatyöryhmä (2014) Lausunto 5/2014 anonymisointitekniikoista. 39 s. 0829/14/FI WP216. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_fi.pdf