

Kvantitatiivisen aineiston anonymisointikeinot

*Eliisa Haanpää, Tietoarkisto
Tutkimusaineistojen anonymisointi -seminaari
5.4.2017 Tampere*



TIETOARKISTO



Sisältö

Kvantitatiivinen aineisto

Anonymisointi yleisesti

Anonymisointi käytännössä

Anonymisoidessa huomioitavaa



Kvantitatiivinen eli määrällinen aineisto

- Kerätty käynti-, puhelin-, internet-kyselyinä strukturoidulla kyselylomakkeella
- Voi olla myös mittausdataa, asiakirja-aineistoja ja rekisteriaineistoja
- Havaintoyksikkö useimmiten henkilö (voi myös olla mm. ilmiö, maantieteellinen alue)
- Kvantidata = tallennettu havaintomatriiseiksi





Kvantitatiivinen eli määrällinen aineisto

Datan käsittely Tietoarkistossa

- Muuttujat tunnistetaan ja tarkistetaan
- Muuttujiin lisätään dokumentaatiota (selitteet)
- Tietosuojasyistä muuttujia voidaan poistaa tai luokitella uudelleen





Tietosuojasyistä muuttujia voidaan poistaa tai luokitella uudelleen **eli anonymisoida**





Muuttujia poistetaan ja luokitellaan, jotta

- aineisto on anonyymi
- aineistoa voidaan pitkäaikaissäilyttää tunnistettomana
- aineistoa voidaan käyttää uudelleen muihin tutkimuksiin



Anonymisoidessa on arvioitava

- Onko yksilö edelleen mahdollista erottaa joukosta?
 - Onko tietojen yhdistäminen yksilöön edelleen mahdollista?
 - Voidaanko yksilöä koskevat tiedot päätellä?
-
- Onko vastaajan tunnistaminen kohtuullisin keinoin mahdollista?

Lähde: Article 29 Data Protection Working Party: Opinion 05/2014 Anonymisation Techniques: https://cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf



Kenen jalat?





Anonymisoinnista

Tietoarkisto toivoo tutkijoiden anonymisoivan kvantitatiivisen aineiston ennen sen arkistointia

Anonymisointi tulee suunnitella aina aineistokohtaisesti

Keinoja kvantitatiivisten aineistojen anonymisoinnille löytyy Tietoarkiston [Aineistonhallinnan käsikirjasta](#)



Yleisimmät kvantitatiivisen aineiston anonymisointikeinot



Miksi aineistohallintaa ja jatkokäyttöä?

Tutkimusaineistojen aineistohallinnalla tarkoitetaan sitä, että tutkimusaineistot (data) ja niihin liittyvä kuvaileva tieto (metatieto, metadata) on luotu, tallennettu ja järjestetty siten, että aineiston säilyy käyttökuntoisena ja luotettavana ja että tietosuojat ja tietoturva on varmistettu aineiston koko elinkaaren ajan.

Yhteiskuntatieteellinen tietoaarkisto on edistänyt yhteiskuntatutkimuksen aineistojen avoimuutta.

[Lue lisää](#)

Miksi aineistohallintaa ja jatkokäyttöä?

- Osa hyvää tieteellistä käytäntä
- Rahoittajat edellyttävät aineistohallintaa
- Jatkokäytön ja avoimen saatavuuden hyödyt
- Tutkimusaineiston elinkaari

Aineistohallinnan suunnittelu

- Aineistohallintasuunnitelma
- 1. Aineisto
- 2. Oikeudet
- 3. Tietosuojat ja tietoturva
- 4. Tiedostomuodot ja ohjelmistot
- 5. Aineiston käytön ja sisällön kuvaus
- 6. Elinkaari

Sopimukset ja oikeudet

- Tekijänoikeus
- Tutkijoiden välisistä sopimuksista
- Oikeuksien siirtymisestä sopiminen
- Tekijyyttä tieteellisissä julkaisuissa ja tutkimusaineistoissa

Tutkittavien informointi

- OSA I: Informoinnin merkitys
- OSA II: Informoinnin sisältö, muoto ja tutkittavan suostumus
- OSA III: Informoinnin osa-alueet
- Osa IV: Esimerkkejä informoinneista

Tiedostomuodot ja ohjelmistot

- Tiedostojen konvertointi
- Tutkimusaineiston digitointi
- Havaintomatriisit ja tilasto-ohjelmistot
- Tekstiaineistot
- Kuva-aineistot
- Ääni- ja äänikuvataallenteet

Kvantitatiivisen datan käsittely

- Havaintomatriisin tallennus
- Muuttujien nimet ja selitteet
- Muuttujien arvot ja niiden selitteet
- Muuttujien muunnokset
- Puuttuva tieto
- Painomuuttujat
- Syntaksin käyttö

Kvalitatiivisen datan käsittely

- Litterointi
- Aineiston sisältämien datatiedostojen organisointi
- Datatiedostojen nimeäminen
- Taustatietojen kirjaamistavat
- Lehtiaineistot

Tunnisteellisuus ja anonymisointi

- Mikä on henkilötieto?
- Milloin tieto on tunnistettavissa?
- Milloin tieto on anonymia?
- Tunnisteellisten aineistojen käsittely
- Anonymisoinnin lähtökohdat
- Kvantitatiivisen aineiston anonymisointi
- Kvalitatiivisen aineiston anonymisointi

Aineiston kuvailu ja metadata

- Metadatan tallennus
- Tutkimuksen toteutuksen kuvailu
- Aineistonkuvauksinstrumentti
- Tiedostojen kuvailu
- Muuttujien kuvailu
- Saatavuustiedot
- Kontekstittiedot ja paradatat
- Metadataformaattit

- Muuttujan poistaminen
- Arvojen uudelleen luokittelu sekä arvojen harkinnanvarainen luokittelu
- Tunnisteiden poistaminen avointen kysymysten vastauksista

www.fsd.uta.fi/aineistohallinta



Muuttujan poistaminen

Muuttujat, joissa on tunnisteita, käsitellään seuraavasti:

- **Suorat tunnisteet** (henkilötunnus, nimi, osoite, puhelinnumero, sähköpostiosoite jne.) poistetaan
- **Vahvat epäsuorat tunnisteet** (opiskelijatunnus, auton rekisteritunnus jne.) poistetaan
- **Epäsuorat tunnisteet** (kunta, maakunta, ikä, sukupuoli, ammatti jne.) arvioidaan, jonka jälkeen ne luokitellaan, karkeistetaan tai poistetaan tarvittaessa



Muuttujan arvojen luokittelu

- Avuksi luokitteluun: Tilastokeskuksen alue- ja henkilöluokitukset
- Luokittelutapoja:
 - Vastausten yhdistäminen luokkiin (esim. ikä: vuosista luokiksi)
 - Ääriarvojen poistaminen tai karkeistaminen (esim. yli 50 vuotiaat)
 - Selkeästi poikkeavien ja yksilöivien arvojen luokittelu Muu-vastaukseksi



Tunnisteet avovastauksissa

- Avoimet kysymykset saattavat sisältää vastaajaa tai kolmansia henkilöitä koskevia tunnisteita, joista vastaaja on kertonut omin sanoin
 - Yksittäisiä tunnisteita voidaan anonymisoida
 - Esim. Humppila > [Maaseutumainen kunta Kanta-Hämeessä]
- Tunnistamisriski tulee arvioida aina tapauskohtaisesti suhteessa aineiston aihepiiriin ja taustamuuttujiin



Mihin tulee kiinnittää huomiota?

Paljonko muuttujat antavat yksilöllistä tietoa vastaajasta?

Muun
muassa

- Työ, opiskelu
- Kotikaupunki/-kunta
- Ulkoiset/näkyvät vammat ja sairaudet
- Kieli tai synnyinmaa
- Vastuutehtävät, luottamustoimet

Mitä vastaus kertoo suhteessa muualta saataviin tietoihin?

Muun
muassa

- Verkossa julkaistu materiaali
- Julkiset asiakirjat
- Lehtijutut
- ”Yleisesti tiedossa oleva”



Mihin tulee kiinnittää huomiota? Esimerkki 1

Paljonko muuttujat antavat yksilöllistä tietoa vastaajasta?

Olin luonnonsuojeluyhdistyksen puheenjohtaja vuonna 2015

Mitä vastaus kertoo suhteessa muualta saataviin tietoihin?

Järjestöjen ja yhdistysten (nykyisten ja menneiden) hallitusten jäsenten nimet löytyvät usein järjestön/yhdistyksen verkkosivuilta tai julkisista asiakirjoista



Mihin tulee kiinnittää huomiota? Esimerkki 2

Paljonko muuttujat antavat yksilöllistä tietoa vastaajasta?

Työskentelen ratikkakuskina

Mitä vastaus kertoo suhteessa muualta saataviin tietoihin?

Raitiovaunuja on Suomessa tällä hetkellä ainoastaan Helsingissä, joten vastaaja työskentelee kyseisessä kaupungissa



Mihin tulee kiinnittää huomiota? Esimerkki 3

Paljonko muuttujat antavat yksilöllistä tietoa vastaajasta?

Mies kaupunginjohtajana Kainuun maakunnassa

Mitä vastaus kertoo suhteessa muualta saataviin tietoihin?

- Mies on Kajaanin kaupunginjohtaja, sillä Kainuussa on vain 2 kaupunkia, joista vain toisessa kaupunginjohtaja on mies
- Julkishallinnollisten tehtävien tiedot löytyvät usein julkisista asiakirjoista usean vuoden ajalta
- Kuntien ja kaupunkien määrät maakunnittain löytyvät verkosta



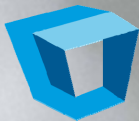
Mihin tulee kiinnittää huomiota? Esimerkki 4

Paljonko muuttujat antavat yksilöllistä tietoa vastaajasta?

Pääsin opiskelemaan sosiaalityötä vuonna 2012 ja kirjoitin pro gradu -tutkielmani ulkomaalaisten uhrikokemuksista

Mitä vastaus kertoo suhteessa muualta saataviin tietoihin?

- Opiskelijavalintojen tulokset saattavat olla julkaistuna verkossa
- Opinnäytetyöt löytyvät usein avoimesti verkosta



Tietoarkisto
www.fsd.uta.fi
[asiakaspalvelu.fsd \(a\) uta.fi](mailto:asiakaspalvelu.fsd@uta.fi)

Esityksen kuvat: Unsplash CC0 1.0