Anonymisoinnin perusteet kvanti- ja kvaliaineistoille

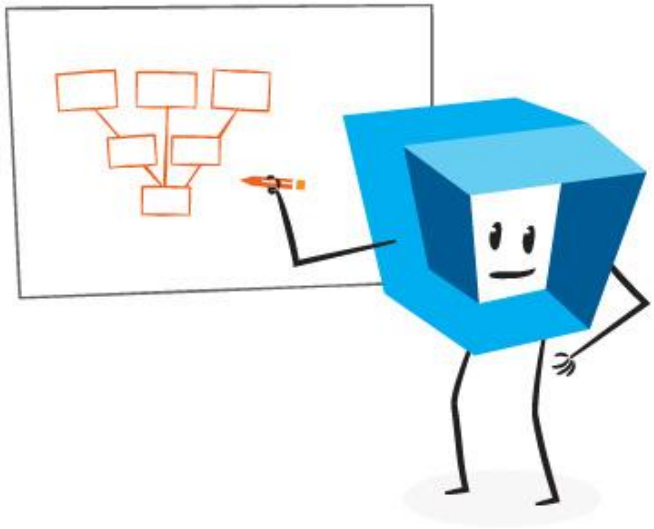
Annika Valaranta & Arja Kuula-Luumi

10.6.2020 Datakouluttamisen kesäpäivä - kouluttajilta kouluttajille

TIETOARKISTO

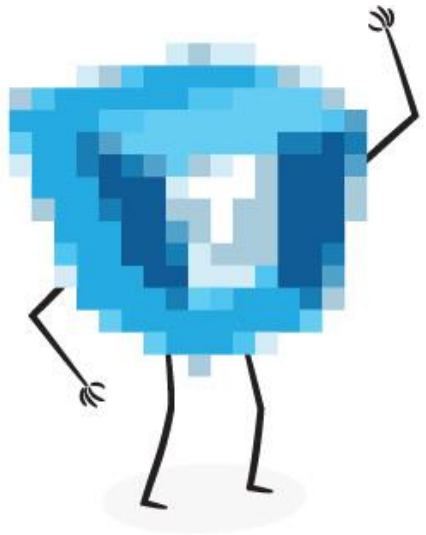
Työpajan sisältö

1. Keskeiset käsitteet
2. Anonymisoinnin suunnittelu
3. Lyhyt tauko
4. Vinkit kvaliaineistoille
5. Laadullinen anonymisointitehtävä pienryhmissä + purku
6. Vinkit kvantiaineistoille
7. Määrällinen anonymisointitehtävä pienryhmissä + purku



Keskeisiä käsitteitä

- Henkilötieto
- Anonyymi tieto
- Pseudonyymi tieto
- **Minimointi**
- Säilytyksen rajoittaminen



TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Minimointi

> *GDPR: henkilötietojen on oltava asianmukaisia ja olennaisia ja rajoitettuja siihen, mikä on tarpeellista suhteessa niihin tarkoituksiin, joita varten niitä käsitellään*

- Vältä turhien, yksityiskohtaisten tai tutkimuksen kannalta merkityksettömien henkilötietojen keräämistä

- Punnitse, mitä taustatietoja tutkittavasta tarvitaan ja millä tarkkuudella

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Konkreettiset vinkit minimointiin

- Kvantitatiivisille ja kvalitatiivisille aineistoille vinkit löytyvät Aineistonhallinnan käsikirjasta : **Minimointi eli miten kerätä aineisto niin, ettei se sisällä turhia tunnisteita?**
- https://www.fsd.tuni.fi/aineistonhallinta/fi/tunnisteellisuus-ja-anonymisointi.html#minimointi-eli-miten-kerata-aineisto-niin-ettei-se-sisalla-turhia-tunnisteita

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Taustatietojen luokittelu

- **Erilaisia taustatietoja ja taustamuuttujia**
- » Sukupuoli
  » Ikä
  » Paikkatiedot, synnyinmaa
  » Elämäntilanne, työmarkkina-asema
  » Suoritettu koulutus
  » Koulutusala
  » Tieteenala
  » Ammattiryhmä, ammattiasema
  » Sektori, toimiala
  » Tulotiedot, yhteiskuntaluokka
  » Uskonto, uskonnollisuus

- https://www.fsd.tuni.fi/aineistonhallinta/fi/dokumentit/taustatiedot-ja-taustamuuttujat.html

Minimointi haastattelujen keruuvaiheessa

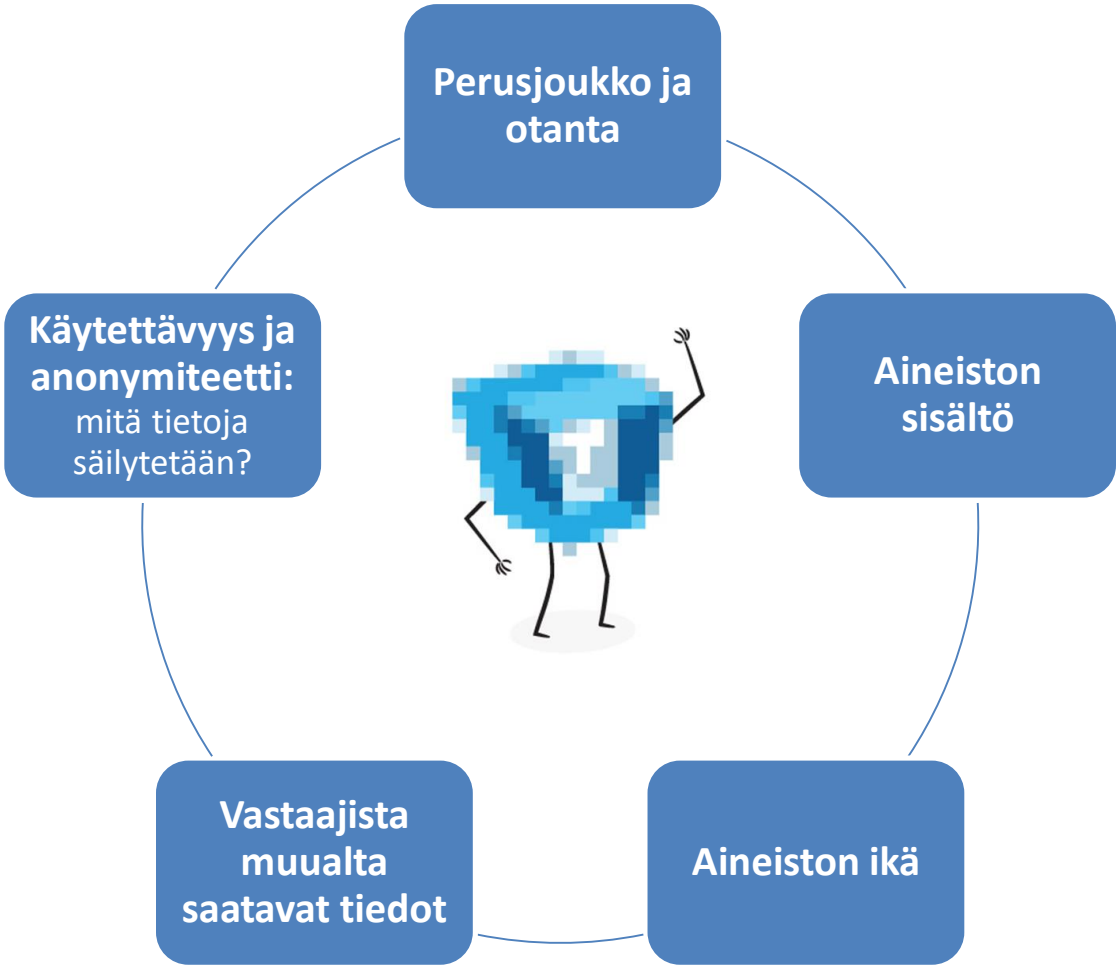
- Laadi taustatietolomake, johon keräät yksityiskohtaisten tietojen sijaan luokitellut tiedot
- Aloita haastattelun äänitallennus vasta kun yksilöidyistä taustatieto-osuudesta siirrytään tutkimuskysymyksiin
- Pyydä haastateltavaa välttämään erityisesti muita koskevien nimien ja muiden tarkkojen tietojen kertomista

Taustatietojen kirjaaminen haastattelussa

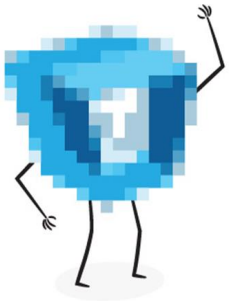
Tutkimus työpaikan sisäilmaongelmien vaikutuksesta työhyvinvointiin

- Haastattelun pvm
- Sukupuoli: Mies__ Nainen__
- Ikä. Haastateltavan Ikä kysytään täsmällisenä, mutta kirjataan taustatietolomakkeelle ikäryhmä:
  - Alle 20 vuotta
  - 20-24 vuotta
  - 25-29 vuotta
  - 30-34 vuotta
  - 35-39 vuotta
  - 40-44 vuotta
  - 45-49 vuotta
  - 50-54 vuotta
  - 55-59 vuotta
  - 60-64 vuotta
  - 65 tai enemmän
- Työpaikka: *Kysytään täsmällisesti, mutta kirjataan luokiteltuna.*
  - *”Tampereen Tammelan koulu”, kirjataan taustatietolomakkeelle ’koulu’*
  - *”Jyväskylän Coforen toimisto”, kirjataan taustatietolomakkeelle ’IT alan yritys’ jne.*
- Työpaikan sijainti:
  - Pääkaupunkiseutu
  - Muun yli 50.000 asukkaan kaupungin keskusta tai lähiö
  - Alle 50.000 asukkaan kaupungin keskusta tai lähiö
  - Taajama harvaan asutulla alueella kaupungissa tai muussa kunnassa
  - Harvaan asuttu alue
- Ammatti: *Voidaan kirjata täsmällisenä tietona, kun muut taustatiedot kirjataan karkeistettuina*

Mitä asioita tulee ottaa huomioon suunniteltaessa anonymisointia?



Milloin tieto on harvinainen? Pitääkö harvinainen tieto aina poistaa?



Kohdejoukko

N=50? 50000?

Bloggari?

Masennus?

Peppi-koira?

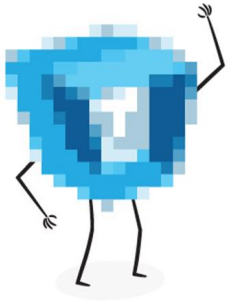
Kilpaurheilee?

Katoptrofobia? (peiliin katsomisen pelko)

Vastaajalla 1 katoptrofobia	Vastaajalla 2 masennus	Vastaaja 3 on bloggari	Vastaajalla 4 Peppi-koira	Vastaaja 5 kilpaurheilee

Pitääkö anonymisoida? Kysy:

1. Onko tieto on harvinainen kohdejoukossa?
2. Jos kyllä, kysy, voiko tiedon saada selville.
3. Jos kyllä, tieto pitää anonymisoida. Jos vastaus on ei, tietoa ei tarvitse anonymisoida.



Jos taustatietoja on paljon ja päätösten tekeminen tuntuu hankalalta…

1. Mieti erilaisia taustatietojen yhdistelmiä, voiko niiden perusteella henkilön tunnistaa kohdejoukossa.
2. Tabula rasa: ”Poista” kaikki tunnisteelliset epäsuorat tunnisteet aineistosta. Lisää niitä sitten yksitellen tutkimusaiheen kannalta tärkeysjärjestyksessä ja pohdi tunnisteellisuutta.
3. Joskus pitää pudottaa toinen tärkeimmistä muuttujista, valitse siis se kaikista tärkein.

Anonymisointitapoja

Henkilönimien korvaaminen peitenimellä: Katri → [Saara].

Yleistävät menetelmät:

- kategorisointi esim. [taajama], [metallialan yritys], [yläaste], [kauppakeskus]
- luokittelu (esim. ikän luokittelu 5 vuoden välein). Luokittelussa pyritään jättämään yksityiskohtaisin luokitus, esim. ei luokitella ikää 10 vuoden välein, jos ei pakko.
- poistaminen: tietojen/muuttujan/tutkittavat

Sekoittavat menetelmät:

- esim. luokitellaan ikä +-2 vuotta, muutetaan tapahtuman päivämäärää tai vaihdetaan vastaajien tietoja keskenään. Käytettävä harkiten!

Anonymisoinnissa poistetaan aina suorat tunnisteet!

Poistettava tieto ei saa olla pääteltävissä!

Tauko 5 min!

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Käytännön anonymisointivinkkejä kvaliaineistoille (1)

- Suunnittele anonymisointi ennen käsittelyä anonymisoinnin yhdenmukaisuuden vuoksi. Kokeile anonymisointia aluksi pariin tiedostoon.
- Tee anonymisointi kopioon. Näin anonymisoinnin virheet voidaan vielä korjata.
- Luo työdokumentti, johon kirjataan esim. yhdenmukaisuutta vaativat kategorisoinnit tai keksityt nimet, esim: Haastattelu 1: Pekka=Matti, Utra=[kaupunginosa 1].

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Käytännön anonymisointivinkkejä kvaliaineistoille (2)

- Käytä anonymisointimerkintöihin jotain merkintää, kuten [hakasulkuja], jotta tiedetään mitä on muutettu ja mitä ei. Älä käytä ’katoavia’ muotoiluja, kuten *kursiivia* tai *värejä*.
- Käytä Wordin Find & replace –komentoa apuna esim. keksittyjen nimien muuttamisessa Anne → [Maria] ja tarkistamisessa! Mutta ole varovainen replace –komennon kanssa, sillä kaikki esim. kirjaimet Anna voivat sisältyä myös muihin kuin nimiin, kuten ”kannattaa”
- Kun anonymisointi on valmis, tuhoa pseudonyymiluettelot ja alkuperäiset tiedostot.

Kuvitteellinen aineistoesimerkki

- **Haastattelututkimus:** Eron vaikutukset lapsiin
  Kohdejoukko: Eron kokeneet henkilöt, joilla on eron aikaan ollut alaikäisiä lapsia.
- **Otanta:** Itsevalikoitunut otos. Haastateltavat löydetty eri puolilta Suomea eroyhdistysten sivuilla julkaistujen tutkimuskutsujen kautta.

- T: Kerro vähän alkuun itsestäsi ja liitostasi ennen eroa.

- H5: No niinku jo puhelimessa sanoin, olen ollut töissä sairaanhoitajaksi valmistuttuani Virtain terveyskeskuksessa, siis Keiturin sotehan tää nyt on. Jennan syntyessä vielä opiskelin Tampereen ammattikorkeakoulussa. Ville valmistui mua myöhemmin Tampereen yliopistosta ja aluksi asuttiin Sarvijaakonkadulla perheasunnossa, mikä oli tosi hyvä, kun mun äiti jäi just eläkkeelle. Tuskin olisi niin onnistunut, jos ei mummo olisi aina ollut apuna, siis asui siinä Kalevassa. Siinä vaiheessa ei riitoja kyllä ollut, yhtä hyörinää. Kai Villekin ajatteli, että yhdessä elämä edetään.

- T: Missä vaiheessa ero sitten tuli?

- H5: Päätös tehtiin yhdessä vuosi sitten ja muistan sen päivän tosi hyvin. 5.7.2018 ja jotenkin se iski lujasti, vaikka itsekään vaihtoehtoa nähny. Kirjotin faceen illalla, että tästä päivästä eteenpäin ei enää Villen kanssa ja semmosta kauhistelua, myös kannustusta tuli, mutta en sitten kyllä siellä jatkanut sitä vatvomista. Yhdessä oltiin oltu kahdeksan vuotta ja Jenna oli viisi. Ville viimeisteli gradua yksinäisistä köyhistä ja sitten ei kai itse halunnut olla itsekään yksinäinen edes opiskeluissa, kun tuli tämä kolmiodraama kuvaan, sulle puhelimessa selitinkin.

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Breakout rooms 12 min!

https://www.fsd.tuni.fi/files/kvalitatiivinen-aineistoesimerkki.pdf

- T: Kerro vähän alkuun itsestäsi ja liitostasi ennen eroa.

- H5: No niinku jo puhelimessa sanoin, olen ollut töissä sairaanhoitajaksi valmistuttuani [terveyskeskuksessa].[Sannin] syntyessä vielä opiskelin [-] ammattikorkeakoulussa. [Saku] valmistui mua myöhemmin [-] yliopistosta ja aluksi asuttiin [tietyn kadun] perheasunnossa, mikä oli tosi hyvä, kun mun äiti jäi just eläkkeelle. Tuskin olisi niin onnistunut, jos ei mummo olisi aina ollut apuna, siis asui siinä [lähellä]. Siinä vaiheessa ei riitoja kyllä ollut, yhtä hyörinää. Kai [Saku]kin ajatteli, että yhdessä elämä edetään.

- T: Missä vaiheessa ero sitten tuli?

- H5: Päätös tehtiin yhdessä vuosi sitten ja muistan sen päivän tosi hyvin.[kesällä 2018] ja jotenkin se iski lujasti, vaikka itsekään vaihtoehtoa nähny. Kirjotin faceen illalla, että tästä päivästä eteenpäin ei enää [Saku]n kanssa ja semmosta kauhistelua, myös kannustusta tuli, mutta en sitten kyllä siellä jatkanut sitä vatvomista. Yhdessä oltiin oltu kahdeksan vuotta ja [Sanni] oli viisi. [Saku] viimeisteli gradua [gradun aihe poistettu] ja sitten ei kai itse halunnut olla itsekään yksinäinen edes opiskeluissa, kun tuli tämä kolmiodraama kuvaan, sulle puhelimessa selitinkin.

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Käytännön anonymisointivinkkejä kvantiaineistoille: (1)

- Suunnittele anonymisointi huomioon ottaen kaikki aineiston muuttujat, ei vain nk. taustamuuttujat. Silmäile myös avomuuttujien vastausten sisältöjä.
- Tee anonymisointi kopioon.
- Anonymisoi ensin numeeriset muuttujat ja viimeisenä avomuuttujat, koska numeeristen muuttujien muokkaus vaikuttaa siihen, miten avomuuttujia anonymisoidaan.
  - Perusohje on, että avomuuttujat eivät saa sisältää tarkempia tietoja esim. iästä, mitä numeerisissa muuttujissa saa selville.
- Anonymisointipäätöksissä ota huomioon muuttujien tilastollinen käyttökelpoisuus ja tieteellinen kiinnostavuus. Esim. kuinka paljon avomuuttujia käytetään kvantitutkimuksessa?

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Käytännön anonymisointivinkkejä kvantiaineistoille: (2)

- Ole huolellinen harvinaisten arvojen poistamisessa/muokkaamisessa. Esim. yhtä harvinaista luokkaa ei voi laittaa yksistään SYSMISiin. Eli jos sysmissiin laittaa arvoja, tulee sinne luokitella muitakin harvinaisia erisuuruisia arvoja tai sysmis-arvoissa tulee jo olla alun perin havaintoja.
- Kirjaa tehdyt anonymisointitoimenpiteet esim. syntaksiin, jotta myöhemmin tiedetään, miten alkuperäistä aineistoa on muokattu.
- Käytä avovastausten anonymisointitoimenpiteisiin jotain merkintää, esimerkiksi [hakasulkuja], jotta tiedetään mitä on muutettu ja mitä ei. Älä käytä ’katoavia’ muotoiluja, kuten *kursiivia* tai *värejä*. On silti perusteltua tehdä anonymisointia ilman merkintöjä tilanteissa, missä merkinnästä voi päätellä poistetun asian.

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Breakout rooms
10 min

https://www.fsd.tuni.fi/files/kvantitatiivinen-aineistoesimerkki-lyhennetty.pdf

Anonymisoinnin suunnittelu: Voit kysyä:

1. Kohdejoukko ja otanta: mitä kertovat vastaajista? Voidaanko ryhmä määritellä tarkasti ulkoapäin?
2. Aineiston sisältö:
   a) Mitä suoria ja epäsuoria tunnisteita aineisto sisältää? Mitä aineiston tietoja yhdistelemällä henkilö saattaa olla tunnistettavissa?
   b) Sisältääkö aineisto kolmansiin henkilöihin liittyviä tietoja ja voiko niiden perusteella tunnistaa henkilöitä?
   c) Sisältääkö aineisto harvinaisia tai ainutlaatuisia tietoja?
   d) Ovatko aineiston tiedot sensitiivisiä?
3. Mitä tietoja tietoihin voi yhdistää ulkopuolelta?
4. Käytettävyys ja anonymiteetti: mitä säilytetään ja mitä ”uhrataan”?

Anonymisointiehdotuksia: kvantitatiivinen aineistoesimerkki

Säilytetään kuntamuuttuja 1

LUOKITELLAAN:

- **Syntymävuosi**: luokitellaan 3 vuoden välein
- **Talouden koko**: TOP-koodataan 4 tai yli
- **Sukupuoli** ja **kunta** jätetään sellaisenaan aineistoon

POISTETAAN:

- Synnyinmaa, ks. huomio alta.
- Harrastusavomuuttuja

HUOM. Voidaan tehdä myös niin, että **synnyinmaata ei poisteta vaan** luokitellaan 1 suomi, 2 muu, **JOS** pienimmissä kunnissa Hailuodossa ja Uuraisilla ei ole tässä aineistossa muualla kuin Suomessa syntyneitä.

Säilytetään kuntamuuttuja 2

LUOKITELLAAN:

- **Kunta:** pienimmät kunnat Hailuoto ja Uurainen yhdistetään
- **Syntymävuosi**: luokitellaan 3 vuoden välein
- **Talouden koko**: TOP-koodataan 4 tai yli
- **Sukupuoli** jätetään sellaisenaan aineistoon
- **Harrasteavomuuttuja:** suurimmat luokat jätetään sellaisenaan, pienimmät yhdistetään

POISTETAAN:

- Synnyinmaa, ks. huomio alta.

HUOM. Voidaan tehdä myös niin, että **synnyinmaata ei poisteta vaan** luokitellaan 1 suomi, 2 muu, **JOS** pienimmissä kunnissa Hailuodossa ja Uuraisilla ei ole tässä aineistossa muualla kuin Suomessa syntyneitä.

Anonymisointiehdotus 3: kvantitatiivinen aineistoesimerkki

Ei kuntamuuttujaa!

Kaikki muut tiedot voidaan jättää sellaisenaan paitsi:

1. Synnyinmaa luokitellaan niin, että sinne jää yleisimmät tai Suomi ja muu
2. Avoimesta harrastusmuuttujasta poistetaan lajikerho-maininnat, kilpaurheilun laji poistetaan, s-postiosoitteet, blogimaininnat, koirien ja ihmisten nimet, ja mainta kirurgisesta virheestä Taysissä.