

VODIČ ZA ANONIMIZACIJO V KVALITATIVNEM RAZISKOVANJU: TRANSKRIPTI INTERVJUJEV

Verzija 1.0

Datum izdaje: Marec 2020

Avtorji: Sergeja Masten, Sonja Bezjak,
Irena Vipavc Brvar, Janez Štebe

Ta dokument je objavljen pod licenco
Creative Commons Attribution-ShareAlike
4.0.



UL, FAKULTETA ZA DRUŽBENE VEDE, KARDELJEVA PLOŠČAD 5 , 1000 LJUBLJANA

tel: +386 01 5805 292 // **e-mail:** arhiv.podatkov@fdv.uni-lj.si // **web:**
<http://www.adp.fdv.uni-lj.si/> // **Twitter:** @Arhivpodatkov // **Facebook:** Arhiv
družboslovnih podatkov

KAZALO

1. KDAJ ANONIMIZIRATI?	3
2. NAČRTOVANJE ANONIMIZACIJE	4
3. PRED PRIČETKOM ANONIMIZACIJE	5
4. POSTOPEK ANONIMIZACIJE	6
4.1 KAJ IN KAKO ANONIMIZIRATI?	6
4.1.1 Še nekaj tehničnih priporočil	6
4.2 PRVI KORAK ANONIMIZACIJE: NADOMEŠČANJE NEPOSREDNIH IDENTIFIKATORJEV	7
4.3 NADALJNE METODE ANONIMIZACIJE	8
4.4 ŠE NEKAJ PRAKTIČNIH NASVETOV	9
4.5 TABELA IDENTIFIKATORJEV IN METODE ANONIMIZACIJE	10
5. ARHIVIRANJE PODATKOV V ADP	12
6. DODATNI VIRI	13
7. UPORABLJENI VIRI	14

ANONIMIZACIJA V KVALITATIVNEM RAZISKOVANJU

Anonimizacija v raziskovalne namene pomeni takšno obdelavo osebnih podatkov, da zmanjša možnost razkritja posameznika - udeleženca raziskave, ter hkrati ohrani uporabnost podatkov za različne namene. Z njo se na svoji profesionalni poti srečuje skoraj vsi družboslovci in družboslovke, ki v svoje raziskovanje vključujejo ljudi.

Anonimizacija je neločljivi del etičnega ravnanja z raziskovalnimi podatki. Če ne prej, se z njo srečamo, ko se namenimo podatke deliti z drugimi. Z vidika sekundarne rabe je ključno vprašanje, kako podatke ohraniti čim bolj bogate in povečati potencial za prihodnjo rabo, pri tem pa poskrbeti za zadostno zaščito udeležencev raziskave, kar raziskovalcem predstavlja poseben izziv.

V Arhivu družboslovnih podatkov smo zbrali in predstavili nekaj ključnih priporočil in informacij, ki so lahko v pomoč pri anonimizaciji raziskovalnih podatkov. Vodič je razdeljen v smiselna poglavja, ki nas vodijo skozi celoten postopek anonimizacije, s ciljem, da se v procesu ne izgubimo ali nehote pozabimo na katero od pomembnih vprašanj. nadaljevanju so opisane minimalne zahteve za prevzem gradiva v ADP preko aplikacije DataVerse.

1. KDAJ ANONIMIZIRATI?

Vprašanje, ki naj raziskovalca vodi pri odločitvi za anonimizacijo je: **»Ali lahko raziskovalni podatki razkrijejo značilne in edinstvene informacije o udeležencu raziskave, ki bi mu škodovale?«** Če je odgovor pritrdilen, potem se moramo brez dvoma lotiti postopka anonimizacije.

Poleg raziskovalne etike nas k skrbnemu ravnanju z osebnimi podatki zavezuje tudi zakonodaja. Uredba EU o varstvu osebnih podatkov (GDPR), ki je v veljavo stopila 25. maja 2018, zahteva, da od udeležencev že pred začetkom raziskave pridobimo soglasje za obdelavo osebnih podatkov, iz katerega je razvidno, s kakšno obdelavo podatkov se udeleženec strinja. Velja omeniti še to, da Uredba ne zadeva obdelave podatkov, ki so anonimizirani do te mere, da posameznik ni (več) določljiv.

Omeniti velja tudi, da je na udeležencu, da sprejme odločitev, ali želi ostati v raziskavi imenovan ali ne. Če raziskovalna tema ni občutljiva in udeleženec oceni, da obdelava podatkov zanj ne predstavlja tveganja, se lahko strinja, da obdelujemo njegove osebne podatke brez, da bi jih anonimizirali. Na to možnost raziskovalci po naših izkušnjah pogosto ne pomislijo (npr. intervju z udeležencem kakšnega pomembnega zgodovinskega dogodka ali procesa). Prav tako ni potrebno anonimizirati podatkov, ki so del javne komunikacije ali zadevajo javne osebnosti.

Kadar bi z anonimizacijo izgubili preveč dragocenih podatkov in preveč okrnili njihovo uporabnost, velja razmišljati o alternativah. Ena možnost je nadzorovan omejen dostop do ne povsem anonimiziranih podatkov.

2. NAČRTOVANJE ANONIMIZACIJE

Anonimizacija je smiselna, ko omogoči zaščito udeležencev, hkrati pa z njo ne izgubimo preveč vrednih informacij. Da bomo zagotovili čim večjo uporabnost podatkovne datoteke, velja delovni proces skrbno načrtovati.

Podatke, ki jih bomo pridobili, moramo dobro poznati. Predvsem je potrebno imeti v mislih značilnosti in posebnosti konkretne raziskave (npr. starost in število udeležencev v raziskavi) in dejstvo, da so znotraj raziskave lahko podatkovne datoteke različno občutljive (npr. intervju z delodajalcem ali intervjuji z delavci). Dobro poznavanje nam bo v pomoč, da lažje ocenimo, koliko in kakšne podatke lahko delimo z drugimi.¹

Okvir za oceno tveganja razkritja posameznika lahko ustvarimo že pri načrtovanju anonimizacije, zato razmislimo o naslednjem:

- Kdo, kdaj in na kak način bo uporabljal podatke?
- Kje in kako bo poskrbljeno za njihovo varno fizično hrambo?
- O številu udeležencev v raziskavi: večje število sodelujočih pogosto zmanjša verjetnost razkritja posameznika.
- Kdaj je bila raziskava narejena? Če so podatki starejši, je manjša verjetnost, da bi sodelujoče lahko razkrili oziroma bi jim podatki lahko škodili.
- Kakšen je vzorčni okvir? Paziti moramo, da preprečimo nenamerno razkritje sodelujočih v raziskavi z informacijami, ki jih vpišemo v spremljajočo dokumentacijo (npr. sodelovali so študentje četrtega letnika sociologije na Fakulteti za družbene vede v študijskem letu 2017/18). To je sicer problematično le, če spremljajoče podatke objavimo skupaj s podatki.²

¹ Povzeto po finskem arhivu FSD: <https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>

² Povzeto po finskem arhivu FSD: <https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>

3. PRED PRIČETKOM ANONIMIZACIJE

Preden se lotimo anonimizacije, poskrbimo še za nekaj tehničnih zahtev³:

- Varno shranimo⁴ prvo, originalno verzijo transkripta, anonimizacijo izvedemo na kopiji.
- Pripravimo navodila za izvedbo postopka anonimizacije in dokumente za sprotno spremljanje poteka dela, saj je priporočljivo beležiti vse spremembe, ki jih bomo naredili kasneje med anonimizacijo.
- Če ocenjujemo, da bo v postopku anonimizacije sodelovalo več raziskovalcev, pripravimo dokument, v katerem bomo zbirali informacij o tem, kdo in kdaj bo izvajal postopek.
- Neanonimizirane dokumente varno hranimo ločeno od anonimiziranih, prav tako dokumente, ki bi jih lahko kdo uporabil za de-anonimizacijo.

³ Povzeto po Arhivu Timescapes: <https://timescapes-archive.leeds.ac.uk/wp-content/uploads/sites/47/2018/04/Timescapes-Anonymisation-Guidelines-18Aug08-in-use.pdf>

⁴ Varno hrambo podatkov omogočajo različni programi, iz Arhiva družboslovnih podatkov priporočamo uporabo VeraCrypt. Več informacij o varni hrambi in etičnem ravnanju z raziskovalnimi podatki je na voljo na spletni strani dogodka, ki smo ga organizirali spomladi 2018.

4. POSTOPEK ANONIMIZACIJE

Velja poudariti, da se anonimizacija lahko nanaša na vse vrste podatkov: ankete, zvočne posnetke in slike, podatke z družbenih omrežij, sodnih spisov in drugo.

V nadaljevanju smo pripravili priporočila za besedilne podatke (npr. transkripti intervjujev).

4.1 KAJ IN KAKO ANONIMIZIRATI?⁵

Prvi korak anonimizacije je opredelitev posrednih in neposrednih identifikatorjev; to so vse informacije, ki bi lahko razkrile sodelujoče v raziskavi. Identificiramo jih lahko na začetku, v praksi pa se pogosto zgodi, da identifikatorje najdemo tudi med samim postopkom.

Neposredni identifikatorji so npr. ime in priimek, naslov, e-poštni naslov, ipd. Posebno pozornost moramo nameniti tudi posrednim identifikatorjem, tj. posebnostim, po katerih udeleženec izstopa in na prvi pogled ne delujejo kot identifikatorji; npr. veliko število otrok, etnična, narodna ali spolna pripadnost, zaposlitveni sektor, prostočasne dejavnosti ali kaj drugega. Značilen primer je proučevanje družin. Če bi le-te raziskovali v Sloveniji, bi bile velike družine z več kot pet otroki precej prepoznavne, zlasti, če bi informacijo o številu otrok spremljala informacija o kraju bivanja. Posredni identifikatorji so tudi lastnosti, ki posamično ne razkrivajo posameznika, lahko pa se ga razkrije s povezovanjem posameznih lastnosti.

O tem, kateri indikatorji prinašajo tveganje razkritja, se mora odločiti vsak raziskovalec sam, saj najbolje pozna svojo proučevano populacijo ter namene in cilje raziskave.

4.1.1 Še nekaj tehničnih priporočil

Prvi korak anonimizacije je opredelitev posrednih in neposrednih identifikatorjev; to so vse informacije, ki bi lahko razkrile sodelujoče v raziskavi. Identificiramo jih lahko na začetku, v praksi pa se pogosto zgodi, da identifikatorje najdemo tudi med samim postopkom.

Neposredni identifikatorji so npr. ime in priimek, naslov, e-poštni naslov, ipd. Posebno pozornost moramo nameniti tudi posrednim identifikatorjem, tj. posebnostim, po katerih udeleženec izstopa in na

⁵ Povzeto po Arhivu Timescapes: <https://timescapes-archive.leeds.ac.uk/wp-content/uploads/sites/47/2018/04/Timescapes-Anonymisation-Guidelines-18Aug08-in-use.pdf>

prvi pogled ne deluje kot identifikatorji; npr. veliko število otrok, etnična, narodna ali spolna pripadnost, zaposlitveni sektor, prostočasne dejavnosti ali kaj drugega. Značilen primer je proučevanje družin. Če bi le-te raziskovali v Sloveniji, bi bile velike družine z več kot pet otroki precej prepoznavne, zlasti, če bi informacijo o številu otrok spremljala informacija o kraju bivanja. Posredni identifikatorji so tudi lastnosti, ki posamično ne razkrivajo posameznika, lahko pa se ga razkrije s povezovanjem posameznih lastnosti.

O tem, kateri indikatorji prinašajo tveganje razkritja, se mora odločiti vsak raziskovalec sam, saj najbolje pozna svojo proučevano populacijo ter namene in cilje raziskave.

4.2 PRVI KORAK ANONIMIZACIJE: NADOMEŠČANJE NEPOSREDNIH IDENTIFIKATORJEV⁶

Prvi korak anonimizacije je vedno detekcija in nadomeščanje neposrednih identifikatorjev. Pri tem se srečujemo s tipičnimi situacijami, najpogostejše so:

Imena oseb: namesto resničnih imen lahko uporabimo izmišljena imena ali pa ime nadomestimo z drugimi lastnostmi. Npr. če je oseba po poklicu zdravnica, lahko namesto imena uporabimo naziv poklica. Osebno ime lahko spremenimo tudi z začetnicama imen, npr. Ana Novak v »AN«, vendar bolj priporočamo uporabo izmišljenih imen, saj s tem ohranimo koherenco podatkov. Pri tem predlagamo uporabo istovrstnih imen, npr. slovenskih imen ne spreminjamo v angleška, saj lahko s tem spremenimo konotacijo vsebine. Več o tem v poglavju o psevdonimizaciji.

Imena naselij/držav: namesto imen naselij lahko navedemo višjo geografsko enoto - npr. občino, regijo ali pa zamenjamo z nadpomenko, npr. Španija z [evropska država], odvisno seveda, kako pomembna je informacija o naselju za razumevanje podatkov. Pozorni moramo biti tudi na anonimizacijo naselij, kjer živijo druge osebe, ki jih respondent tekom pogovora omeni, saj bi le-to lahko razkrilo respondenta ali posameznika, o katerem govori.

Ime/naziv delovnega mesta, izobraževalne institucije: prav tako namesto natančnega opisa delovnega mesta ali šole, zapišemo bolj splošno ime; npr. ime trgovske verige Mercator zamenjamo z [živilska trgovina].

⁶ Povzeto po Arhivu Timescapes: <https://timescapes-archive.leeds.ac.uk/wp-content/uploads/sites/47/2018/04/Timescapes-Anonymisation-Guidelines-18Aug08-in-use.pdf>

4.3 NADALJNE METODE ANONIMIZACIJE⁷

Na voljo imamo različne metode anonimizacije. Lastnosti, s katerimi tvegamo razkritje posameznika, lahko psevdonimiziramo, odstranimo, kategoriziramo oziroma klasificiramo. V praksi se pogosto zgodi, da izvedemo različne kombinacije anonimiziranja, saj je to odvisno od podatkov, ki jih imamo. Ves čas pa imejmo v mislih, da želimo ohraniti čim več podatkov – zakriti jih želimo le toliko, kolikor je potrebno.

Zamenjava osebnih imen s psevdonimi – psevdonimizacija

Zamenjava originalnih imen s psevdonimi je najbolj razširjena in znana oblika anonimizacije kvalitativnih podatkov.

Opozoriti želimo še na posebne okoliščine v povezavi z evropsko Splošno uredbo o varstvu osebnih podatkov, ki v 4. členu opredeli še en pojem, to je »psevdoanonimizacija«. V primeru »psevdoanonimizacije« smo podatke sicer res psevdonimizirali, vendar smo ločeno shranili ključ, s katerim po potrebi lahko pridemo do identitete udeležencev. To je potrebno npr., ko izvajamo longitudinalno raziskavo in želimo z istimi udeleženci govoriti večkrat, npr. v daljšem časovnem razponu. V takih primerih moramo poskrbeti za tehnične in organizacijske ukrepe, s katerimi zaščitimo informacije, ki bi nepooblaščenim osebam razkrile identiteto in povzročile morebitno škodo.

Ponovimo, da so podatki anonimizirani lahko rečemo šele takrat, ko smo uničili posredne in neposredne identifikatorje.

Kategorizacija/klasifikacija

V nekaterih primerih moramo podatke anonimizirati z metodo kategorizacije ali klasifikacije. To pomeni, da ustvarimo smiselne skupine oz. kategorije, s katerimi opišemo lastnosti posameznika.

Kategorije določimo na podlagi tega, kakšno vrednost imajo ti podatki. Ime in priimek lahko zamenjamo s širšim poimenovanjem, npr. z vlogo, ki jo igra oseba v nekem odnosu: »prijateljica«, »sošolka«, »sosed« in podobno. Če podatkovna datoteka vključuje samo eno šolo, npr. Gimnazijo Bežigrad, njeno ime lahko zamenjamo z »gimnazija v Ljubljani«.

Če smo intervjuvali 44-letno profesorico sociologije na Fakulteti za družbene vede v Ljubljani, ki je poročena in ima dva otroke, starosti 9 in 11 let, starost respondentke smiselno uvrstimo v starostno skupino, prav tako naredimo za otroke, zaposlitveni status pa posplošimo.

⁷ Povzeto po finskem arhivu FSD: <https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>

Primer anonimizacije:

Spol: ženski

Starost: 40–50 let

Delovno mesto: Univerza v Ljubljani

Poklic: profesorica družboslovnih ved

Status: poročena, vsaj en otrok

Pozorni moramo biti, da podatki ne razkrijejo preveč. Če bi bile ženske v akademskem prostoru slabo zastopane (kot na primer v preteklosti), bi že podatek o tem, da gre za žensko, predstavljal tveganje za razkritje identitete.

Odstranitev občutljivih podatkov

Odstranitev podatkov naj bo skrajna odločitev, saj je cilj, da podatkovno datoteko čim manj okrnemo. Podatke odstranimo le, če podatek resnično odkriva informacije, ki bi sodelujočemu lahko škodili, ustrezno zamenjava ali nadpomenka pa ne bi bili ustrezna rešitev.

Ko anonimiziramo, podatka ne odstranimo oz. izbrišemo kar tako, temveč na tistem mestu besedo zamenjamo z besedno zvezo »zakrita vrednost«, ki jo postavimo v oglati oklepaj. S tem ohranimo sledljivost sprememb zase in za sekundarne uporabnike.

Dodaten ukrep

Če v postopku ugotovimo, da bi z anonimizacijo preveč posegli v vsebino, lahko delno anonimizirane ali neanonimizirane podatke predamo v režim omejenega nadzorovanega dostopa.

4.4 ŠE NEKAJ PRAKTIČNIH NASVETOV⁸

Preverimo, če so v izvorniku za isto osebo uporabili več poimenovanj, npr. vzdevek, samo ime ali samo priimek, idr., saj moramo v tem primeru anonimizirati vse.

Pri anonimizaciji uporabimo orodji »find/najdi« in »replace/zamenjaj« (če sta na voljo v našem urejevalniku besedila), kar bo olajšalo in pohitrilo spremembe na več mestih hkrati.

Ko anonimiziramo, bodimo temeljiti in natančni.

⁸ Povzeto po Arhivu Timescapes: <https://timescapes-archive.leeds.ac.uk/wp-content/uploads/sites/47/2018/04/Timescapes-Anonymisation-Guidelines-18Aug08-in-use.pdf>

4.5 TABELA IDENTIFIKATORJEV IN METODE ANONIMIZACIJE⁹

V finskem arhivu družboslovnih podatkov so za raziskovalce pripravili tabelo identifikatorjev, ki je lahko v pomoč pri odločitvi glede metode anonimizacije.

Metodo priporočajo v odvisnosti od tega, kakšno tveganje na splošno predstavlja identifikator. Končno presojo mora narediti vsak raziskovalec sam, glede na naravo podatkov in/ali dogovor iz soglasja, pridobljenega od udeleženca.

Identifikatorji, označeni z *, so bolj občutljivi.

identifikator	direktni identifikator	močan posredni identifikator	posredni identifikator	priporočena metoda anonimizacije
ID sodelujoče osebe	x			odstranitev
polno osebno ime	x			odstranitev/zamenjava (psevdonim)
starost			x	kategorizacija
letnica rojstva		x		kategorizacija
spol			x	
spolna usmerjenost*			x	(odstranitev)
poklic		x	x	kategorizacija
zakonski stan			x	
sestava gospodinjstva			x	
stopnja izobrazbe			x	
e-naslov	x	x		odstranitev
telefonska številka		x		odstranitev
poštna številka			x	odstranitev/kategorizacija
območje bivanja/del mesta			x	kategorizacija
občina			x	kategorizacija
regija			x	kategorizacija
avdio posnetek	x			odstranitev
video posnetek	x			odstranitev
fotografija	x			odstranitev
zaposlitveni status			x	
zaposlitveni sektor			x	

⁹ Povzeto po finskem arhivu FSD: <https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>

delovno mesto/samozaposlen		x	x	kategorizacija
nacionalnost				(kategorizacija)
materni jezik				kategorizacija
registrska številka vozila		x		odstranitev
številka bančnega računa		x		odstranitev
IP naslov		x		odstranitev
ID študenta		x		odstranitev
informacije o zdravstvenem stanju*		x	x	kategorizacija/odstranitev
pripadnost etnični skupini*		x	x	kategorizacija/odstranitev
članstvo v sindikatu*			x	kategorizacija
prejemniki socialne pomoči*			x	kategorizacija/odstranitev

5. ARHIVIRANJE PODATKOV V ADP

V ADP skrbimo za varno hrambo raziskovalnih podatkov, in sicer z ustrezno zaščito podatkov ter z različno strogim omejevanjem dostopa do podatkov.

V prvi fazi prevzema raziskovalnih podatkov in povezanega gradiva ocenimo tveganja za razkritje udeležencev raziskave. Raziskovalca izprašamo o podatkovni datoteki, vrsti osebnih podatkov ter o tematici raziskave. Prav tako preverimo, če je raziskovalec od sodelujočih pridobil soglasje za udeležbo pri raziskavi in za nadaljnjo obdelavo podatkov. Skupaj z raziskovalcem določimo stopnjo občutljivosti podatkov in mu po potrebi svetujemo v zvezi z anonimizacijo.

Raziskovalec, dajalec podatkov ali arhivistka v ADP po potrebi pripravijo različne verzije mikropodatkov z različnimi stopnjami zaščite za različne tipe dostopa. V ADP lahko hranimo neanonimizirano in anonimizirano podatkovno datoteko, za vrsto dostopa do podatkov pa se dogovorimo z dajalcem glede na že omenjene lastnosti podatkov in da zmanjšamo s tem povezana tveganja.

V ADP so na voljo štiri vrste dostopa do podatkov, s katerimi nadziramo uporabo (občutljivih) podatkov:

- Odprti dostop do neobčutljivih ali polno zaščiteneh (to je anonimiziranih) mikropodatkov za vse uporabnike (brez registracije v ADP).
- Dostop do polno zaščiteneh oz. anonimiziranih mikropodatkov za vse registrirane uporabnike (največkrat je to v obliki t.i. PUF - Public Use Files).
- Dostop do manj zaščiteneh (delno anonimiziranih) mikropodatkov (SUF - Scientific Use Files) za registrirane uporabnike in izpolnjenim obrazcom za dostop do gradiva na zahtevo; za namen analize jih lahko prenesejo k sebi.
- Dostop do nezaščiteneh občutljivih mikropodatkov (t.i. ScUF - Secure Use Files) za registrirane uporabnike s statusom raziskovalca (šifro ARRS) in izpolnjenim obrazcom za dostop do gradiva na zahtevo; analizira jih lahko le v varni sobi ADP.

Več informacij je na voljo v [Politikah ADP](#).¹⁰ Za dodatna vprašanja smo raziskovalcem na voljo tudi v ADP, kjer nudimo strokovne nasvete in drugo potrebno podporo.

¹⁰ Dostopno na: https://www.adp.fdv.uni-lj.si/media/publikacije/Politika_ADP.pdf

6. DODATNI VIRI

Raziskovalcem priporočamo, da več praktičnih primerov in napotkov za ustrezno anonimizacijo (in pripravo kvalitativnih podatkov) poiščejo še v:

- [finskem arhivu družboslovnih podatkov](#),
- angleškem arhivu [Timescapes](#),
- spletnem učbeniku o ravnanju z raziskovalnimi podatki [Data Management Expert Guide](#) konzorcija CESSDA,
- priporočilih revije American Behavioral Scientist, npr. v članku [Qualitative Data Sharing: Data Repositories and Academic Libraries as Key Partners in Addressing Challenges](#),
- angleškem združenju za anonimizacijo The UK Anonymisation Network (UKAN); npr. v dokumentu [The Anonymisation Decision-Making Framework](#).



7. UPORABLJENI VIRI

CESSDA. (b. d.). Data Management Expert Guide. Dostopno prek: <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide>

Finnish Social Science Data Archive. (b.d.). Anonymisation and Personal Data. Dostopno prek: <http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html>

Timescapes. (2008). Timescapes Anonymisation Guidelines. Dostopno prek <https://timescapes-archive.leeds.ac.uk/wp-content/uploads/sites/47/2018/04/Timescapes-Anonymisation-Guidelines-18Aug08-in-use.pdf>