UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Sabina Bevc

Vloga masovnih podatkov v družboslovnem raziskovanju

Diplomsko delo

Ljubljana, 2017

UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Sabina Bevc
Mentor: red. prof. dr. Vasja Vehovar

Vloga masovnih podatkov v družboslovnem raziskovanju

Diplomsko delo

Ljubljana, 2017

Hvala ti za ves nov svet.

Vloga masovnih podatkov v družboslovnem raziskovanju

V družboslovnem raziskovanju se vse pogosteje omenja pojem »big data« oziroma masovni podatki. Kljub temu pogosto ni jasno, kaj točno to pomeni in na kakšen način spreminja družboslovno raziskovanje. V diplomskem delu smo naredili pregled stanja na področju masovnih podatkov v odnosu do anketnega raziskovanja. Pregledali smo, kaj masovni podatki so, kako jih pridobimo in kako se uporabljajo. Prav tako smo raziskali vidike kakovosti in izzive pri delu z masovnimi podatki, posebej znanja, ki so potrebna za delo z njimi. Preučili smo tudi etične vidike pri zbiranju in analiziranju podatkov, zakonske okvire ter vidike prihajajočih sprememb. V empiričnem delu smo skušali pogledati v prihodnost in s pomočjo treh strokovnjakov, ki se ukvarjajo z zbiranjem in analizo podatkov, odgovorili na vprašanje, ali so masovni podatki le modna muha (angl. buzzword) ali pa bodo resno spremenili način družboslovnega raziskovanja. Delo smo zaključili z glavnimi ugotovitvami in priporočili za nadaljnje povezovanje masovnih podatkov z anketami.

Ključne besede: anketno raziskovanje, masovni podatki, big data, veliko podatkovje.

The role of big data in social research

In social sciences, the term »big data« is increasingly mentioned. Nevertheless, it is often not clear what exactly this means and in what way it changes social science research. In this diploma we conduct an overview of the situation in the field of big data in relation to surveys. We look at what big data are and how to obtain them, and consider how they can be combined with survey research. We also investigate the quality and the challenges when working with big data, examine the ethics and legal frameworks and the necessary regulations that will be introduced in the future. In the empirical part we try to look into the future with the help of three experts, who are dealing with data collection. The experts answered the question whether big data is only a buzzword, or will they seriously change the social science research. We conclude our work with the main findings and recommendations for further linking of big data with surveys.

Keywords: big data, massive data, survey research.

KAZALO

KAZALO SLIK

1 UVOD

Z razvojem novih tehnologij se v družboslovju ustvarja ogromno podatkov. V zadnjem času se je za posebej obsežne podatke začel uporabljati izraz »big data«, kar prevajamo kot masovni podatki. V diplomskem delu nas bo zanimalo, kako si s temi podatki lahko pomagamo pri družboslovnem raziskovanju, kakšna je njihova vloga pri analiziranju sodobne družbe in kako jih lahko povežemo z anketnim raziskovanjem. Zanimalo nas bo tudi, kaj masovni podatki sploh so, na kakšen način so proizvedeni in kako dostopamo do njih. Prav tako je pomembno vprašanje kakovosti ter pravnih in etičnih vidikov.

Posebej nas bo zanimalo, ali se masovni podatki in anketno raziskovanje dopolnjujejo oziroma kakšen je pravzaprav njihov odnos. Prav tako bomo poskušali odgovoriti na vprašanje prihodnosti masovnih podatkov.

Pregledali bomo tudi nekaj primerov praktične uporabe v trženjskem raziskovanju, saj lahko z masovnimi podatki odkrivamo vidike, ki jih z anketnim raziskovanjem veliko težje.

Delo bomo začeli v drugem poglavju s pregledom teorije, kjer bomo analizirali ozadje masovnih podatkov in njihov razvoj ter predstavili najbolj znano definicijo masovnih podatkov, ki sta jo zasnovala Beyer in Laney (2012). Širše bomo razdelali tudi vire masovnih podatkov in jih umestili v kontekst družboslovnega raziskovanja.

V tretjem poglavju se bomo osredotočili na izzive pri delu z masovnimi podatki in na njihovo zanesljivost. Posvetili se bomo etiki pri zbiranju in uporabi podatkov, pomembno vprašanje pa so tudi zakonski okviri, ki jih imamo oziroma nimamo. Posodobljena Splošna uredba EU o varstvu podatkov (General Data Protection Regulation – GDPR) namreč pride v veljavo maja 2018 in prinaša kar nekaj sprememb za raziskovalce.

Četrto poglavje smo namenili znanju, ki ga potrebujemo za delo z masovnimi podatki.

Najbolj pa nas bo seveda zanimal odnos anket in masovnih podatkov, kar bomo podrobneje opredelili v petem poglavju. Zanimalo nas bo, kako se anketni in masovni podatki dopolnjujejo, ter podali nekaj praktičnih primerov kombinirane uporabe v marketinškem in politološkem raziskovanju.

Šesto poglavje, empirični del, bo sestavljeno iz ekspertnih intervjujev s tremi strokovnjaki, ki se ukvarjajo z zbiranjem podatkov. To so dr. Ana Slavec, Boro Nikić in dr. Blaž Zupan. Predstavili bodo svoj pogled na prihodnost masovnih podatkov in anketnega raziskovanja.

V sklepnem delu, sedmem poglavju, bomo poskusili odgovoriti na vprašanje stanja uporabe masovnih podatkov in podali mnenje o nadaljnjih smereh družboslovnega raziskovanja.

Čeprav je na voljo veliko literature, smo ocenili, da je za pričujoče diplomsko delo v tem trenutku najbolj relevantno poglavje Callegara in Yanga (2017), »The Role of Surveys in the Era of »Big Data«, objavljeno v knjigi The Palgrave Handbook of Survey Research, v katerem avtorja poglobljeno obravnavata odnos masovnih podatkov in anketnega raziskovanja. Navedeno delo bomo v nadaljevanju pogosto citirali, nanj pa se bomo naslonili tudi pri strukturi in konceptualnih izhodiščih.

2 OPREDELITEV MASOVNIH PODATKOV

Najbolj znan izraz, ki je povezan s pojavom posebej obsežnih podatkov, je zagotovo angleški izraz »big data«. Pri izbiri primernega slovenskega prevoda smo naleteli na kar nekaj različic, med drugim masovni podatki, masivni podatki, veliki podatki, obilni podatki, velepodatki in veliko podatkovje. V pričujočem delu smo izbrali izraz masovni podatki, saj ga uporablja velika večina javnih slovenskih institucij in informacijsko tehnoloških podjetij, poleg tega pa ocenjujemo, da je ta izraz laiku najbolj razumljiv. Tudi iskalnik Google v povezavi z angleškim izrazom »big data« najpogosteje navaja prav poimenovanje »masovni podatki«.

2.1 Definicija masovnih podatkov

Definicija masovnih podatkov se stalno spreminja in je lahko zelo zapletena. Med najbolj citiranimi je opredelitev Beyerja in Laneyja (2012), ki masovne podatke pojasni takole:

- ko govorimo o *obsegu (Volume)*: to so tisti podatki, ki ne morejo biti obdelani s tradicionalnimi analitičnimi orodji;
- ko govorimo o *hitrosti (Velocity)*: to so tisti podatki, ki se ustvarjajo in so nam na voljo v realnem času (angl. real-time), kar pomeni v istem trenutku, kot so ustvarjeni;
- ko govorimo o *raznovrstnosti (Variety)*: to so zapleteni nabori podatkov, ki vsebujejo različne vire vsebine, na primer nestrukturirana besedila, slike, videe in druge vire podatkov.

V angleščini se izrazi, ki so zapisani poševno, začnejo s črko V, zato ponekod zasledimo uporabo definicije 3V po Gartnerju (Beyer in Laney 2012). Poleg navedenih treh

karakteristik pa nekateri avtorji dodajajo še *spremenljivost (Variability)* v smislu nekonsistentnosti njihovega pojavljanja, *verodostojnost (Veracity)* v smislu točnosti in kakovosti podatkov in *zapletenost (Complexity)*, ki se navezuje na možnost pravilnega povezovanja podatkovnih baz (Callegaro in Yang 2017).

Zanimiva je tudi definicija po Grovesu (2011), ki je za anketno raziskovanje posebej relevantna, saj primerja masovne podatke z anketnimi in pravi, da so ti načrtovani, medtem ko so masovni podatki organski in producirani s sistemi, ki samodejno beležijo transakcije in druge zapise.

2.2 Viri masovnih podatkov

Masovne podatke lahko z anketnimi primerjamo glede na izvor. V nadaljevanju bomo navedli primere virov masovnih podatkov po Callegaru in Yangu (2017) in na kratko opisali, v kakšen kontekst družboslovnega raziskovanja jih lahko umestimo.

- *Internetni podatki* – sem spadajo spletna besedila, videi ter glasovni podatki. Podrobneje te podatke delimo še na:
    - *podatke družbenih medijev* – podatki, ki vsebujejo besedila, videe in fotografije in so javno na voljo z rudarjenjem po družbenih omrežjih, kot je npr. Twitter. Ti podatki so tudi med najbolj analiziranimi masovnimi podatki, ko govorimo o merjenju javnega mnenja (Schober in drugi v Callegaro in Yang 2017);
    - *metapodatke, piškotke in analitike spletnih strani* – podatki, ki so ustvarjeni na spletnih straneh in z analitičnimi orodji, primer je Google Analytics. Ti podatki se večinoma uporabljajo za spletno oglaševanje, preučevanje obnašanja na strani in analizo delovanja spletne strani;
    - *internet stvari (Internet of Things – IoT)* – IoT se navezuje na katerokoli napravo, ki lahko komunicira z drugo napravo in pri tem uporablja internet kot skupni protokol prenosa informacij (Gershenfeld, Krikorian, Choen v Callegaro in Yang 2017). Z internetom je povezanih vse več naprav, pri čemer se ustvarja veliko sledi, zapisov in drugih podatkov. V tem okviru imamo še poseben sklop *podatkov o obnašanju*, ki jih dobimo iz npr. kamer, pametnih telefonov, pametnih ur in ostale tako imenovane nosljive (angl. wearable) tehnologije, ki spremljajo in zapisujejo podatke o fizični aktivnosti, lokaciji in splošnem telesnem zdravju uporabnika.

Podatke uporabnik pravzaprav zbira sam, vendar se zapisujejo oziroma so posredovani tudi snovalcem teh aplikacij in naprav (Callegaro in Yang 2017);

- *transakcijski podatki* – to so zapiski spletnih naročil, pošiljk, plačil, vračil, računov in aktivnosti kreditnih kartic (Ferguson v Callegaro in Yang 2017). Transakcijski podatki so obstajali še pred pojavom elektronskih podatkov, spremenil se je samo način njihovega generiranja. Ti podatki so na primer del orodij upravljanja odnosov s strankami (CRM), ki beležijo vsako interakcijo, ki jo ima potrošnik s podjetjem ali izdelkom (Callegaro in Yang 2017);
- *administrativni podatki* – so oblika masovnih podatkov, ki se zbirajo s strani uradnih institucij, kot so davčna uprava, šole in inštituti za varovanje zdravja. Te vrste podatkov se pogosto uporabljajo v namene statističnih analiz in se lahko povezujejo s podatki anketnih raziskav (Wallgren in Wallgren v Callegaro in Yang 2017);
- *podatkovne baze, namenjene komercialni uporabi* – vse več podjetij se ukvarja z zbiranjem in shranjevanjem podatkov o potrošnikih, ki jih z algoritmi in drugimi tehnikami uporabljajo za ustvarjanje oziroma razlago profilov potrošnikov in jih tako generirane prodajo naprej (Callegaro in Yang 2017).

Poseben izvor masovnih podatkov so pri anketnem raziskovanju lahko tudi tako imenovani *parapodatki*. To so pomožni procesni, administrativni podatki, ki nastanejo v procesu zbiranja anketnih podatkov (Groves in drugi 2004). Parapodatki so bili sprva vezani zgolj na podatke, ki so stranski proizvod računalniško podprtega anketiranja, s porastom spletnega anketiranja pa imamo sedaj dostop tudi do bolj zapletenih vrst podatkov, kot je premikanje miške na zaslonu. K parapodatkom med drugim prištevamo anketarjeve zapiske, atribute poskusov anketiranja, vedenje med intervjujem, podatke o vnosu podatkov, odzivni čas trajanja ankete, število poskusov navezave stika z anketirancem, napravo, s katero je spletna anketa opravljena, in celo način, na katerega se anketiranec odloča o odgovoru (Callegaro in drugi 2015).

Omeniti velja tudi pametno tehnologijo, ki je v velikem vzponu. Tu predvsem mislimo na pametne igrače, avtonomne avtomobilske sisteme in inteligentne osebne asistente, ki živijo v oblaku (na primer Alexa, ki jo je razvil Amazon). Tudi preko teh sistemov lahko pridobivamo masovne podatke.

3 KAKOVOST, ZANESLJIVOST, ZASEBNOST

V tem poglavju se bomo osredotočili na delo z masovnimi podatki in izzivi, s katerimi se raziskovalci srečujejo. Predvsem bomo podrobneje pregledali, kako kakovostni in zanesljivi so masovni podatki. Posvetili se bomo tudi pravnim vidikom, saj se vse pogosteje pojavlja vprašanje zasebnosti in meje uporabe teh podatkov.

3.1 Kakovost in zanesljivost

Callegaro in Yang (2017) pravita, da masovni podatki ne pomenijo nujno kakovosti brez napak. Pogosto jih spremlja »veliki šum« (angl. big noise), ki ga Waldherr in drugi (2016) razlagajo kot nepotrebne, nebistvene informacije v naboru podatkov. Ko govorimo o masovnih podatkih v odnosu do anket, pa raziskovalci upoštevajo koncept skupne anketne napake (Total Survey Error – TSE), in prav to združevanje dveh disciplin je obetaven način uporabe podatkov v novem, širšem smislu. TSE se »nanaša na kopičenje vseh napak, ki se lahko pojavijo med načrtovanjem, zbiranjem, obdelavo in analizo podatkov anketne raziskave. V tem kontekstu je anketna napaka (angl. survey error) opredeljena kot odstopanje anketnega odgovora od njegove osnovne prave vrednosti« (Japec in drugi 2015).

Oglejmo si najprej strukturo TSE:

- *napaka specifikacije* – do te napake pride, ko se koncept, ki ga meri anketno vprašanje, razlikuje od koncepta, ki bi moral biti merjen v anketi. Ko se to zgodi, se meri napačen konstrukt in posledično ocenjuje napačen parameter, kar lahko vodi do nepravih sklepov. Te napake so neredko posledica slabe komunikacije med raziskovalcem in oblikovalcem anketnega vprašalnika;
- *napaka merjenja* – je ena najpogostejših virov napake. Vsebuje napake, ki jih »povzročijo« respondenti, anketarji, anketna vprašanja in razni drugi dejavniki. Respondenti lahko namerno ali nenamerno podajo napačne odgovore, anketarji lahko povzročajo napake na več načinov, med drugim z načinom govora, videzom, namigovanjem odgovora respondentom, napačno razlago vprašanja, ponarejanjem podatkov in podobno. Anketni vprašalniki so lahko velik vir napake, če so narobe izdelani – dvoumna vprašanja in nejasna navodila so lahko velik problem;
- *napaka okvira* – ta napaka nastane v procesu izdelave, vzdrževanja in uporabe vzorčnega okvira za izbiro vzorca ankete. Vzorčni okvir pomeni seznam dela ciljne populacije ali uporabo kateregakoli drugega mehanizma za sestavljanje vzorca. Idealno bi vzorčni okvir vseboval vsakega člana ciljne populacije (brez dvojnikov),

enote, ki niso del ciljne populacije, pa bi bile odstranjene iz okvira. To pomeni, da morajo biti informacije, ki so uporabljene v vzorčnem okviru, točne in ažurirane. Na žalost se redko zgodi, da je vzorčni okvir popoln, kar lahko pripelje do napake populacijske nepokritosti;

- *napaka neodgovora* – je dokaj splošen vir napake, ki zajema tako enoto kot posamezno spremenljivko. Napaka neodgovora enote se pojavi, ko se enota v vzorcu (na primer gospodinjstvo) ne odzove na izpolnjevanje vprašalnika (na primer gospodinjstvo, ki noče sodelovati v osebnem terenskem intervjuju, ali vprašalnik, poslan po pošti, ki ni nikoli vrnjen). Neodgovor spremenljivke pa pomeni, da je vprašalnik izpolnjen samo delno, ker je bil intervju predčasno končan ali pa je bilo nekaj vprašanj preskočenih oziroma namenoma neodgovorjenih (na primer vprašanje o prihodkih ima pogosto visoko stopnjo neodgovora);
- *napaka v obdelavi podatkov* – ta napaka vsebuje napake v urejanju, vnosu in kodiranju podatkov, odstopanja uteži in napačen tabelarni prikaz podatkov. Ko govorimo o tabelarnem prikazu podatkov, moramo paziti na napake v celicah, saj lahko pride do napačnih rezultatov (Biemer 2010).

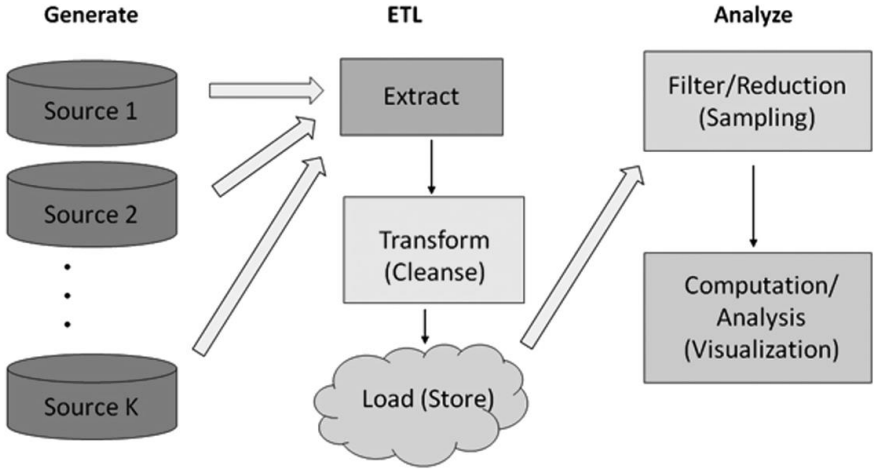
Zgornji koncept TSE, ki se uporablja v anketnem raziskovanju, so Japec in drugi (2015) implementirali tudi na masovne podatke in ga poimenovali »Big Data Total Error« (BDTE), kar lahko prevedemo kot *skupna napaka masovnih podatkov* oziroma skupna napaka v masovnih podatkih. Tako kot se za klasične ankete s TSE obravnavajo napake in se jim skuša izogniti že pred izvedbo, tudi za delo z masovnimi podatki potrebujemo koncept, ki bo raziskovalcem pomagal do pravih rezultatov in BDTE je bil razvit prav v ta namen. Vsebuje glavne tipe napak, ki so specifične za masovne podatke in lahko povzročijo dodatno dvoumnost in nezanesljivost v končnih rezultatih. Največ napak pri delu z masovnimi podatki se pojavi na naslednje tri načine:

- *generiranje podatkov* – proces, ki najbolj razlikuje masovne podatke in ankete. Pridobivanje podatkov je včasih črna skrinjica in napake se lahko najdejo v obliki manjkajočih ali nereprezentativnih podatkov, samodejnem izboru podatkov, slabem pokritju in podobno;
- *pridobivanje, preoblikovanje in nalaganje (Extract, Transform and Load – ETL)* – postopek, ko so podatki združeni v istem računalniškem okolju s procesom ekstrakcije (dostop do podatkov, razčlenjevanje in shranjevanje iz različnih virov), transformacije (kodiranje, ponovno kodiranje, urejanje) in nalaganja (integracija in

skladiščenje). Napake v postopku ETL so lahko v ujemanju, kodiranju, urejanju in čiščenju podatkov napak;

- *analiza in vizualizacija* – v zadnjem koraku se podatki pretvorijo v informacije skozi proces, ki vsebuje dve fazi. Prva faza je filtrirna stopnja (vzorčenje) oziroma zmanjšanje, kjer so izbrisane nezaželene lastnosti in vsebine. Funkcije se lahko združijo za izdelavo novih, podatkovni elementi pa se lahko stanjšajo ali vzorčijo, da so bolj obvladljivi. Druga faza je faza izračuna, analize in vizualizacije, v kateri se podatki analizirajo in vizualizirajo ter tako uporabijo za razlago in pridobivanje informacij (Japec in drugi 2015).

Slika 3.1: Načrt procesiranja masovnih podatkov



Vir: Biemer (v Japec in drugi 2015).

Ravno tako kot ankete lahko tudi proces generiranja podatkov za masovne podatke ustvari zmotne in nepopolne podatke. Poleg tega so lahko viri podatkov selektivni v smislu, da zbrani podatki morda ne predstavljajo natančno opredeljene populacije oziroma tiste, ki je reprezentativna za ciljno populacijo, ki nas zanima. To pomeni, »da napake pri generiranju podatkov vsebujejo nizko razmerje med signalom in šumom, nepopolne ali manjkajoče vrednosti, nenaključne selektivne vire in metapodatke, ki so pomanjkljivi, odsotni ali napačni« (Japec in drugi 2015).

Koncept skupne napake masovnih podatkov se v praksi še zelo malo uporablja, čeprav sta Japec in Biemer njegova velika zagovornika in spodbujata njegovo uporabo. V splošnem je na področju uporabe masovnih podatkov še ogromno prostora za izboljšave in odkrivanje novih načinov zagotavljanja kakovosti podatkov.

3.2 Zasebnost, etika in pravni vidiki

Kar se tiče pravnih vidikov, je zakonodaja glede uporabe masovnih podatkov še precej zastarela oziroma je sploh ni. Vsi vsakodnevno v digitalnem svetu proizvajamo ogromno podatkov, ki so potencialno uporabni za raziskovanje, vendar pa ostaja prisotno vprašanje, kdo je njihov lastnik. Je to tisti, ki jih proizvaja, organizacija, ki jih zbira, oseba, ki jih analizira, ali pa družba nasploh? Zakoni po svetu te podatke obravnavajo različno, nekateri kot »posest«, drugi kot informacijo, ponekod pa zakonov sploh še ni (Japec in drugi 2015). Zasebnost, etični in pravni vidiki za zbiranje, hrambo in analizo masovnih podatkov so še vedno pod velikim vprašajem in njihov status najbrž še kar nekaj časa ne bo znan.

Leta 2018 prihaja v uporabo posodobljena Splošna uredba EU o varstvu podatkov (General Data Protection Regulation – GDPR), ki prinaša kar nekaj sprememb na področju zbiranja in hrambe osebnih podatkov. Uredba je sprožila manjši preplah med podjetji, vendar pravzaprav prinaša zgolj večjo transparentnost, ki je pomembna predvsem za končne uporabnike oziroma potrošnike. Največ sprememb bo na petih področjih:

- *pridobitev oziroma posodobitev soglasja za zbiranje podatkov* – ko bodo podjetja zbirala podatke, bo potrebna natančnejša opredelitev namena in načina obdelave (na primer »zbrane podatke bomo uporabljali v trženjske namene« bo premalo, potrebna bo točna razlaga);
- *zbiranje podatkov* – ravno zaradi prve točke bo potreben temeljit razmislek, v katere namene se bodo podatki zbirali in kje jih bodo podjetja hranila;
- *izpis osebnih podatkov uporabnika* – na zahtevo uporabnika bo moralo podjetje posredovati vse podatke, ki jih bo o njem zbralo. Tu bo bolj problem sodelovanja podjetij – zelo pomembna je dobra, transparentna komunikacija vseh vpletenih;
- *obdelava podatkov* – za vse podatke, pri katerih ni bilo točno določeno, za kaj jih bodo podjetja uporabljala, bo treba pridobiti nova soglasja. Prav tako se bodo zbrani podatki lahko obdelovali zgolj na način, ki ga je sprejel uporabnik. Torej bo treba za vsako novo, drugačno obdelavo pridobiti novo soglasje;
- *izbris podatkov na zahtevo uporabnika* – podjetje mora izbrisati vse zbrane podatke, če tako zahteva uporabnik. Problem lahko nastane, če podjetja za obdelavo in uporabo uporabljajo orodja, ki niso del lastnega sistema. V tem primeru je lahko trajen izbris vprašljiv, saj nimamo zagotovljenega podatka, da je do izbrisa res prišlo (Informacijski pooblaščenec Republike Slovenije 2017).

Maja 2017 je na Pravni fakulteti Univerze v Ljubljana potekala Mednarodna znanstvena konferenca »Big Data: New Challenges for Law and Ethics«, ki se je posvečala predvsem zakonu in etiki zbiranja in uporabe podatkov. Velik del konference je bil namenjen uporabi masovnih podatkov pri delu policistov in kriminalistov, ki izvajajo nadzor kriminala in iščejo vzorce, s katerimi bi morda napovedali morebiten teroristični napad ali pa možnost ponovnih kriminalnih dejanj tistih, ki so izpuščeni iz zapora. V tem primeru so zbiranje podatkov, varnost in zasebnost zelo pomembni. Kar se tiče pravnih vidikov, torej niti najvišja avtoriteta ne more imeti povsem prostega dostopa do zbiranja, hrambe in uporabe masovnih podatkov. Govorci so tako predstavili kar nekaj realnih primerov reševanja in napovedovanja zločinov ter izpostavili nujnost ureditve zakonov glede vsega, kar se tiče masovnih podatkov. Za prihodnost v kriminalnem pravosodju napovedujejo povečano uporabo masovnih podatkov.

Poleg pravnih problemov obstaja dodatno vprašanje etike raziskovalcev in odgovornosti pri zbiranju in uporabi masovnih podatkov. Zook in drugi (2017) opozarjajo, da se je treba zavedati, da se podatki nanašajo na osebe in lahko zato naredijo veliko škodo, če niso pravilno uporabljeni. Velikokrat je način zbiranja podatkov nezakonit, zato je zelo pomembno, da je v raziskovalni ekipi ali podjetju kdo odgovoren za ozaveščanje in opominjanje raziskovalcev na etični pristop. Zelo hitro namreč lahko pride do nepopravljive škode, zato je priporočljivo razviti sistem, ki bi take dogodke preprečil.

4 DELO Z MASOVNIMI PODATKI

V tem poglavju bomo govorili o znanju, ki ga potrebujemo za delo z masovnimi podatki. Razložili bomo, kako se masovni podatki sploh pridobijo, in navedli orodja, s katerimi to počnemo. Preverili bomo tudi, katere spremembe je uporaba masovnih podatkov prinesla anketnemu raziskovanju.

4.1 Znanja za delo z masovnimi podatki

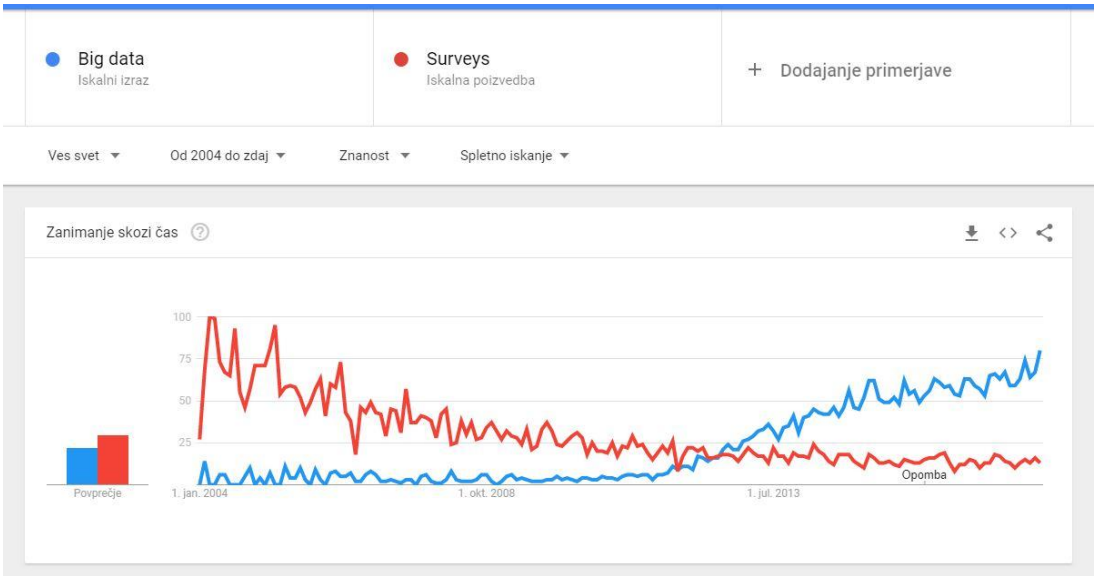
Za zbiranje, analizo in interpretacijo masovnih podatkov potrebujemo znanja, ki jih v družboslovnem raziskovanju navadno ne pridobimo oziroma se jih ne priučimo. Del teh znanj vsebuje delo s podatkovnimi bazami, programiranje, vizualizacijo podatkov in druge analitične tehnike, ki niso nujno del študijskega programa, ki sicer obravnava anketno raziskovanje (Callegaro in Yang 2017). V zadnjem času se veliko govori o poklicu »podatkovni znanstvenik«

(angl. data scientist), ki prav tako nima točno določene definicije, je pa to oseba, ki se ukvarja s podatki in obvlada tako statistiko in matematiko kot programiranje, zna podatke prečistiti, organizirati in tudi razumeti. Ker pa vemo, da je nemogoče, da en profil poklica zna vse, kar se tiče masovnih podatkov (predvsem zato, ker se ta panoga zelo hitro razvija), je nujno sodelovanje med različnimi znanji in poklici, saj le na ta način lahko res dobro delamo s podatki, ki jih potrebujemo za raziskovanje.

4.2 Orodja za delo z masovnimi podatki

Obstajajo številna orodja, ki nam lahko pomagajo pri delu z masovnimi podatki. Google Trends je na primer orodje, s katerim lahko izvemo indeks aktivnosti iskanja s ključnimi besedami ali kategorijami in indeks interesa za te ključne besede v nekem časovnem obdobju. Za primer vzemimo primerjavo iskanja ključnih besed »big data« in »surveys« po svetu od leta 2004 do danes. Na grafu (Slika 4.1) vidimo primerjavo indeksa iskanja, ki ponazarja porast iskanja izraza »big data«, preskok pa se je zgodil leta 2012. Na podoben način lahko uporabimo Google Trends za raziskovanje izbranih zanimanj populacije v časovnih obdobjih. Obstaja več študij, ki so že uporabile orodje Google Trends za napoved in orientacijo pri ocenjevanju anket, vendar ta metoda ni primerna za oceno specifičnih anketnih vprašanj, kot je recimo demografska analiza. Za primer lahko vzamemo odnos žensk in moških do recesije – tega ne moremo opredeliti zgolj s podatki iz orodja Google Trends, temveč potrebujemo poglobljen vpogled, ki ga lahko dosežemo zgolj z anketiranjem (samo z enim vprašanjem ne moremo oceniti odnosa po spolu, potrebnih je več namenskih anketnih vprašanj) (Callegaro in Yang 2017).

Slika 4.1: Graf orodja Google Trends, primerjava popularnosti iskalnih izrazov



Vir: Google Trends (2017).

Tu so še razna orodja za spremljanje družbenih omrežij, ki opravljajo dve pomembni nalogi, in sicer lociranje vsebine družbenih omrežij (preiščejo razna družbena omrežja) in avtomatizirano analizo vsebine zbranega besedila. Ta orodja se razlikujejo predvsem v tem, kako globoko, široko in daleč v čas sežejo z vsebino, ki jo sestavijo iz zbranih podatkov.

Vse bolj se uveljavlja tudi tekstovno rudarjenje (angl. data mining), ki deluje podobno kot besedilna analiza in je proces pridobivanja visokokakovostnih informacij iz besedila. Vsebina besedila se klasificira kot pozitivna, negativna ali nevtralna, v ta namen pa se uporabljajo dodatna orodja za kvalificiranje besedila, na primer slovarji ali leksikoni, ki določeno besedo opredeljujejo kot pozitivno ali negativno. Vse to pomaga raziskovalcem pri prepoznavanju vzorcev in podobno. Bolj relevantno za anketno raziskovanje je zagotovo mnenjsko rudarjenje (angl. *opinion mining*) in analiza stališč (angl. *sentiment analysis*), ki po podobnem principu analizira odnos oziroma čustveno reakcijo na nek dogodek, interakcijo, besedilo. Ta analiza se pogosto uporablja za analiziranje ocen (restavracij, potovanja, nakupne izkušnje ipd.), anketnih odgovorov, spletnih in družbenih medijev ...

V kontekstu anketnega raziskovanja se družbena omrežja uporabljajo tudi za predvolilne napovedi, pa vendar taki načini zaenkrat niso vedno najbolj natančni in zanesljivi. Pri uporabi teh orodij je še vedno precej metodoloških in tehničnih vidikov, na katere je treba biti pozoren, hkrati pa jih je treba še dodobra preučiti (Callegaro in Yang 2017).

Poleg naštetih orodij raziskovalci uporabljajo tudi tako imenovano »strganje podatkov« (angl. data scraping) za dopolnjevanje že pridobljenih podatkov z anketami. Strganje podatkov pomeni, da se skozi neko spletno stran sprehodi računalniški program in zbere vse informacije, ustvarjene s strani uporabnika. Programi za strganje podatkov so na voljo brezplačno in so dostopni vsem. Tak način dopolnjevanja podatkov, pridobljenih z anketami, uporabljajo tudi na Statističnem uradu Republike Slovenije, in sicer za raziskavo o prostih delovnih mestih – na straneh, ki objavljajo zaposlitvene oglase, aktivirajo program za strganje podatkov in s pridobljenim materialom potem ustvarijo končne analize.

4.3 Analitična orodja

Raziskovalci običajno porabijo precej časa za pripravo podatkov in za analizo, kar je navdahnilo kar nekaj zagonskih podjetij, ki so začela razvijati programsko opremo za delo z masovnimi podatki. Med najbolj znanimi je zagotovo podjetje Apache Hadoop, ki je razvilo odprtokodno platformo za delo z masovnimi podatki. Uporabljajo jo predvsem velika podjetja in organizacije, omogoča pa porazdeljeno obdelavo velikih nizov podatkov s skupinami računalnikov.

Med pogostimi odprtokodnimi programskimi jeziki, ki se uporabljajo za analizo masovnih podatkov, pa je tudi R. Ni samo programski jezik, ampak integriran paket programske opreme, ki deluje v okolju, v katerem se izvajajo statistične in grafične tehnike (linearni in nelinearni modeli, klasični statistični testi, združevanje podatkov, klasifikacija in drugo). Ker je razmeroma prijazen za uporabo, je med raziskovalci zelo razširjen. Čeprav je R statistično orodje, ima vse značilnosti programskega jezika, kot so Python, C+ in SAS (R-project 2017).

5 POVEZOVANJE ANKETNIH IN MASOVNIH PODATKOV

V tem poglavju si bomo ogledali, kako se lahko anketni in masovni podatki dopolnjujejo in kakšne so prednosti ali slabosti uporabe obeh virov podatkov za iskanje odgovorov. Japec in drugi (2015) v svojem poročilu AAPOR močno zagovarjajo sočasno uporabo masovnih podatkov in anket za čim bolj relevantne končne rezultate.

5.1 Spremembe v anketnem raziskovanju

Novi tipi podatkov in orodja za delo z njimi v raziskovalnem svetu povzročajo kar nekaj ugibanja, ali bodo masovni podatki sčasoma nadomestili ankete. Callegaro in Yang (2017) omenjata Bakerja, ki je podrobno razdelal poročilo ESOMAR (Global Market Research 2015) in ugotovil, da se manjša delež proračuna, porabljenega za ankete s strani tržno raziskovalnih podjetij skozi čas. Posebej izrazit je 6-odstotni padec proračuna za osebne terenske ankete, telefonske ankete in za poštne ankete v letih 2013 in 2014. Prav tako Baker (v Callegaro in Yang 2017) navaja podatek, da se je v tem času povečal proračun, porabljen za digitalno zbiranje in analizo podatkov, kar pomeni, da tržno raziskovalne agencije vse več denarja porabljajo za investicijo v informacije, ki jih dobijo z masovnimi podatki.

Callegaro in Yang (2017) se sicer strinjata s trendom, ki ga nakazuje ESOMAR, pa vendar menita, da v poročilu niso zajeli celotnega premika od tradicionalnih metod zbiranja podatkov k modernejši metodi. Tu ciljata predvsem na spletne in mobilne ankete, ki zamenjujejo metode osebnega terenskega pridobivanja podatkov. Poudarjata, da se vse bolj povečuje uporaba anket »naredi sam« (angl. *Do It Yourself – DIY*), ki jih omogočajo razne spletne platforme (pri nas je to 1KA) in orodja, ki jih razvijejo podjetja sama. Tako je bilo v Sloveniji zgolj na spletni platformi 1KA v letu 2017 1,5 milijona izpolnjenih vprašalnikov, skupno število registriranih uporabnikov pa je preseglo 45.000 (1KA 2017).

Tako bi lahko dejanski trend po Callegaru in Yangu (2017), kar se tiče anket v družboslovnem in trženjskem raziskovanju, opredelili približno takole:

- od metod klasičnega zbiranja podatkov k spletnim anketam,
- od spletnih anket k mobilnim,
- od zunanjih izvajalcev tržnih raziskav do internih, ki jih omogočajo spletne platforme za ankete »naredi sam«,
- od zunanjih izvajalcev tržnih raziskav do internih raziskav, popolno integriranih z notranjimi sistemi podjetja (Callegaro in Yang 2017).

5.2 Prednosti masovnih podatkov

Ljudje vse več časa prebijemo na spletu, kjer puščamo sledi in ogromno podatkov, ki veliko povedo o nas (morda več, kot bi si sami želeli), pa vendar raziskovalci včasih pozabljajo nanje

in se preveč osredotočijo zgolj na ankete. Eden večjih problemov v anketah predstavljajo vedenjska vprašanja, saj se moramo zanesti na spomin respondentov, vsekakor pa je prednost anket v prilagodljivosti vprašanj in pa v reprezentativnem vzorčnem okviru. Raziskovalcem je dana popolna kontrola s tem, ko lahko določijo reprezentativen vzorec in iz pridobljenih odgovorov sklepajo na populacijo, ki jih zanima. Poleg tega sta, kar se tiče verjetnostnih vzorcev, do potankosti razviti tako teorija kot praksa in raziskovalci točno vedo, kako zbrane podatke učinkovito uporabiti za odgovore na zastavljena raziskovalna vprašanja (Japec in drugi 2015). Masovni podatki po drugi strani ponujajo možnost večjega vzorca, ki je predvsem bolj podroben, kar se tiče prostora in časa ter podobnih skupin, vendar je tudi veliko bolj neobvladljiv. Poleg velikosti imajo masovni podatki tudi druge prednosti v primerjavi z anketami, največja je seveda to, da že obstajajo v neki obliki in da lahko načrtovanje in izvajanje pridobivanja podatkov izpustimo. Primarno zbiranje podatkov je namreč zelo drago in vzame veliko časa, upada pa tudi stopnja odgovora (angl. response rate), predvsem pri obsežnejših anketah (Japec in drugi 2015).

Z današnjimi tehnologijami je masovne podatke lažje obdelovati in analizirati kot v preteklosti, pozitivna je tudi dostopnost podatkov v realnem času (angl. real time), saj so ustvarjeni organsko. Japec in drugi (2015) v tem okviru posebej navajajo primere telefonskih klicev, brskanja po spletu, spletno nakupovanje in podobno. Lastnosti masovnih podatkov so najbolj privlačne za zasebni sektor, kjer lahko podjetja dnevno sprejemajo odločitve na podlagi podatkov uporabnikov, medtem ko si s klasičnim zbiranjem podatkov ta postopek lahko precej podaljšajo.

5.3 Združevanje masovnih podatkov in anket

Velika večina raziskovalcev se strinja, da bi za doseganje najbolj relevantnih rezultatov ankete in masovne podatke morali uporabljati kombinirano. Idealno bi bilo maksimalno izkoristiti prednosti obojih, tako lahko z masovnimi podatki spoznamo vedenje in »kaj«, z anketami pa »zakaj« (Callegaro in Yang 2017).

5.3.1 Kombiniranje v marketinškem raziskovanju

Primeri, ki jih kot dobro prakso izpostavljajo Japec in drugi (2015), temeljijo na natančni kombinaciji glede na zahteve določene situacije. Tako navajajo primer oglaševalca, ki stalno spremlja promet v trgovini in obseg prodaje v realnem času. Tradicionalni raziskovalni modeli,

ki anketirance sprašujejo po motivaciji in točki nakupa, lahko pomagajo trgovcem bolje ciljati določene kupce. Druga možnost pa je razširitev analitičnega načrta, tako da primarni podatki za spremljanje poslovanja postanejo promet v trgovini in obseg prodaje, ankete pa se uporabljajo za izvajanje globljega razumevanja trendov ali zaznavanje nepravilnosti, ki so odkrite v primarnem načinu spremljanja podatkov.

Trenutno zelo aktualna tema v trženju je izraz »growth hacking« oziroma pospeševanje rasti. Za to skrbijo strokovnjaki, ki se ukvarjajo s svetovanjem podjetjem, predvsem v digitalnem trženju. Gre za vprašanje kako spoznati potencialne kupce in na kakšen način jih prepričati v nakup izdelka ali storitve. Ponavadi ti strokovnjaki uporabljajo podatke, ki jih zberejo z orodji, kot sta Google Analytics in CRM podjetja. Tu pa se lahko hitro zatakne, saj lahko kljub temu ne dosežejo želenega cilja, zato še vedno anketirajo uporabnike. V tem okviru se kratke ankete z izbranim orodjem (zelo razširjena je uporaba orodja Hotjar) integrirajo na strani podjetja ali pa pošiljajo v obliki elektronske pošte na že zbrane naslove. Na ta način skušajo zbirati dodatne informacije o zadovoljstvu ali nezadovoljstvu, da bi dognali najboljši način za izboljšano poslovanje in večje prepoznavanje blagovne znamke.

5.3.2 Kombiniranje v političnem raziskovanju

Tudi v politiki najdemo dober primer kombiniranja podatkov anket in masovnih podatkov. Nickerson in Rogers (2014) sta odlično razdelala napovedno modeliranje ameriških političnih kampanj, ki za pravilno ciljanje oziroma nagovarjanje volivcev uporabljajo zelo veliko podatkov. Zberejo jih v treh kategorijah za vsakega državljana, ki je v bazi volivcev. Tako vedenjska kategorija napovednega modela zajema demografske podatke in preteklo obnašanje (na primer finančno podporo politične stranke ali udeležbo na shodu). Kategorija ocene podpore predvideva, kakšne so politične preference volivca – tu pa se ponovno vračajo k anketam. Ker je nemogoče vsakega volivca posebej vprašati, za katerega kandidata se bo odločil, raziskovalci sestavijo model, na podlagi katerega anketirajo določeno število volivcev, in rezultate posplošijo na populacijo. Tretja kategorija pa pokriva oceno odziva. Raziskovalci s terensko raziskavo, pri kateri vnaprej pripravijo nekaj različnih primerov kampanj, merijo odzive volivcev, da dobijo približno sliko vseh mogočih scenarijev odziva na kampanjo. Na ta način se lahko napovejo odzivi za vsako ciljno skupino, dobro pripravijo materiali kampanje in sprejemajo boljše odločitve med celotnimi aktivnostmi kampanje. Seveda je podatek, za koga kdo glasuje, anonimen, vendar pri analizi in napovedi raziskovalci uporabijo vsako podrobnost glede volivca, ki jo lahko dobijo kot javno dostopen podatek: ali je oseba sploh volila, v katerem

kraju živi, agregirani podatki popisa prebivalstva (povprečni dohodek, stopnja izobrazbe in drugo). Poleg tega pa lahko podatke tudi kupujejo od raziskovalnih agencij (Nickerson in Rogers 2014).

Do leta 2012 so imele ameriške politične kampanje velik problem z združevanjem podatkov v digitalni obliki in anketnih podatkov, zbranih na terenu. Kampanja Baracka Obame je to spremenila z razvojem programa Narwhal, ki je združil vse podatke v eno podatkovno bazo – ta je na začetku kampanje vsebovala deset terabajtov podatkov, na koncu pa kar 50 terabajtov, s čimer je postavila nov mejnik v načrtovanju političnih kampanj (Nickerson in Rogers 2014).

V Sloveniji potekajo vzporedne napovedi volilnih rezultatov z anketami, nismo pa še zasledili resnejše uporabe masovnih podatkov v kontekstu, kot smo ga opisali.

6 EMPIRIČNI DEL: EKSPERTNI INTERVJUJI

V tem poglavju smo za mnenje o masovnih podatkih in prihodnosti vprašali tri strokovnjake, ki se dnevno ukvarjajo s podatki. Njihova ocena je zelo pomembna, saj imajo vpogled v strokovno dogajanje na področju družboslovnega raziskovanja in so seznanjeni tako s kadrovskimi izzivi kot z zakonskimi okviri.

6.1 Intervju s slovenskimi raziskovalci

Da bi lažje razumeli in razložili, kam pelje prihodnost masovnih podatkov in anketnega raziskovanja, smo za mnenje (gre za leto 2017) povprašali tri strokovnjake, ki se poklicno ukvarjajo s podatki. Dr. Ana Slavec je zaposlena kot svetovalka za statistiko v Centru odličnosti InnoRenew CoE, Boro Nikić je zaposlen na Statističnem uradu Republike Slovenije, dr. Blaž Zupan pa se na Fakulteti za računalništvo Univerze v Ljubljani ukvarja z odkrivanjem znanj iz podatkov.

1. Kako po vašem mnenju masovni podatki spreminjajo klasično anketno raziskovanje?

Ana Slavec (2017): »Big data oziroma velepodatki samega anketnega raziskovanja bistveno ne spreminjajo, ampak ga dopolnjujejo. So še en podatkovni vir, ki se lahko kombinira z drugimi in omogoča primerjavo rezultatov z različnih perspektiv. Velepodatki pa lahko

nastajajo tudi v anketnem raziskovanju – pri tem imam v mislih parapodatke, tj. podatke o procesu anketiranja, ki so lahko osnova za izboljšave anketne metodologije.«

Boro Nikić (2017)*:* »Klasično anketno raziskovanje spreminjajo v smislu zbiranja podatkov (nekaj podobnega kot uporaba administrativnih virov), neposredna posledica tega so zmanjševanje obremenitev poročevalskih enot, hitrejša (tudi pogostejša) objava statistik in možnost izračuna novih statistik. Izzivi pa so potrebno znanje "data mininga" (strojno učenje), modeliranja ...«

Blaž Zupan (2017): »Gre za dve različni metodologiji, ki se lahko dopolnjujeta.«

2. Je vzajemno sodelovanje mogoče in ali bodo masovni podatki počasi izpodrinili uporabo anket? V kolikšni meri pravzaprav uporabljate masovne podatke pri svojem delu?

Ana Slavec (2017):

> *Menim, da je sodelovanje mogoče in da velepodatki ne bodo izpodrinili anket. Vsaka metoda raziskovanja ima svoje prednosti in slabosti. Pri velepodatkih smo omejeni le na pojave, za katere obstajajo tovrstni podatki, medtem ko gre pri anketah in intervjujih za spraševanje, s čimer lahko pridobimo tudi podatke o obnašanju in stališčih, ki jih z velepodatki ne bi pridobili. Sama pri svojem delu (še) ne uporabljam velepodatkov, a soglašam z novimi možnostmi vpogleda v obnašanje prebivalstva, ki ga omogočajo. Menim, da so lahko koristni v raziskovalne namene, vendar se je treba zavedati, da vsak podatek še ni koristna informacija, ampak ga mora uporabnik osmisliti z raziskovalnim vprašanjem.*

Boro Nikić (2017)*:*

> *Po mojem mnenju uporabe anket ne bodo nikoli v celoti izpodrinili, ker lahko podatke nekaterih enot zberemo samo od enot opazovanja. Če ne drugega, bodo ankete (raziskovanja) ostale zato, da se bodo velepodatkovne statistike lahko "validirale". Veliko bo pa tudi kombiniranja podatkov anket in velepodatkovnih virov. Velepodatkov ne smemo gledati samo v kontekstu zamenjave podatkovnih virov za obstoječe ankete,*

temveč kot pripravo novih statistik in (ali) hitrejše (pogostejše) objavljanje obstoječih statistik.

Blaž Zupan (2017): »Kot rečeno, gre za dopolnjevalne metode. Moj laboratorij razvija metode za analizo podatkov in jih uporablja na področju biomedicine (tam so zlasti standardni eksperimenti, ne ankete).«

3. Kaj pa napake, ki se pojavijo pri uporabi in interpretaciji tako pridobljenih podatkov, kakšne so rešitve?

Ana Slavec (2017): »Kot sem že omenila, je pogosta napaka, da raziskovalci velepodatkov ne znajo pravilno osmisliti in postaviti pravih raziskovalnih vprašanj ter na koncu interpretirati rezultatov. Predvsem je pomembno, da se raziskovalci zavedajo mogoče pristranskosti velepodatkov zaradi nepokritosti in drugih napak. Po mojem mnenju je rešitev predvsem v kombiniranju velepodatkov s tradicionalnimi načini zbiranja podatkov.«

Boro Nikić (2017): »Glede napak je pa tako: eden največjih izzivov je določiti reprezentativnost (angl. selectivity) teh podatkov, potem je tu uporaba "data mining" tehnik in uporaba modeliranj. Kot sem že prej povedal, tu pridejo prav občasne ankete, s katerimi lahko ocenimo omenjene izzive.«

Blaž Zupan (2017): »Najbolj pogoste napake so zaradi nepoznavanja metodologije in pristopov (tako kot pri anketah).«

4. Kako bo nova uredba GDPR vplivala na zbiranje, hrambo in uporabo podatkov pri vašem delu? Mislite, da so v splošnem raziskovalci v Sloveniji sploh že pripravljeni in vedo, kaj ureditev prinaša?

Ana Slavec (2017):

Z uredbo GDPR sem se seznanila pri delu v Arhivu družboslovnih podatkov (ADP), kjer sem bila vključena v projekt, ki se ukvarja z etičnimi in pravnimi vprašanji uporabe novih podatkovnih virov v družboslovnem raziskovanju. Uredba krepi pravice posameznikov in nalaga strožje predpise za zbiralce in obdelovalce osebnih podatkov. Slovenska zakonodaja je večinoma že skladna z uredbo GDPR, vendar je še nejasno, kakšne bodo administrativne kazni, ki jih naš pravni sistem še ne pozna. ADP je januarja

letos organiziral okroglo mizo na temo etike v znanosti s poudarkom na varnosti osebnih podatkov, s katero je v raziskovalni skupnosti sprožil razpravo o priložnostih in tveganjih, ki jih prinaša nova uredba. Govorilo se je tudi o oblikovanju delovne skupine, ki bi spremljala področje in predlagala spremembe. Sicer pa se mi zdi, da so v splošnem raziskovalci z uredbo še premalo seznanjeni.

Od septembra 2017 sem zaposlena kot svetovalka za statistiko v Centru odličnosti InnoRenew CoE. Moja glavna naloga je svetovanje drugim raziskovalcem pri zbiranju, analizi in upravljanju podatkov. Poleg tega me je InnoRenew imenoval za pooblaščeno osebo za varstvo osebnih podatkov (Data Protection Officer – DPO). Uredba GDPR bo vplivala na raziskave na ljudeh, pri katerih bomo morali biti pozorni, da imamo informirano privolitev sodelujočih za zbiranje, procesiranje in hrambo podatkov, poskrbeti pa bomo morali za ustrezno zaščito osebnih podatkov in drugih informacij, na podlagi katerih bi se lahko določen posameznik identificiral.

Boro Nikić (2017):

GDPR je širši pojem, saj se nanaša na osebne podatke, ki večinoma nastopajo v tradicionalnih virih podatkov. Ta uredba prinaša jasna pravila, ki jih je treba vpeljati pri delu z osebnimi podatki ("auditing" ...). Poleg tega je treba imeti osebo, ki bo spremljala izvajanje uredbe GDPR ("data officer DM"). Na SURS-u je trenutno skupina, ki se s tem aktivno ukvarja. Če bodo podatki anonimizirani, ne vidim razloga, da raziskovalci teh podatkov ne bi dobili. Ali so seznanjeni z uredbo GDPR, pa ne vem.

Blaž Zupan (2017): »Z zbiranjem podatkov se ne ukvarjam.«

5. Imamo dovolj kadra, ki zna delati z masovnimi podatki – jih pridobiti in oplemenititi? Kaj lahko še storimo na tem področju, kar se tiče izobraževanja?

Ana Slavec (2017): »Tovrstnih kadrov je verjetno premalo. S formalnim izobraževanjem se o tem skoraj ničesar ne naučiš, razen verjetno na Fakulteti za računalništvo in informatiko. Večina izobraževanja na tem področju pa je verjetno neformalna oziroma gre za učenje z izkušnjami. Za vpeljavo teh vsebin v učni proces bi potrebovali več strokovnjakov s tega področja.«

Boro Nikić (2017):

> *Kadra seveda ni dovolj. Tu so potrebe po specifičnih znanjih, ki jih trenutno nimamo (no, tudi drugje v Evropi ne). Tisti, ki pa ta znanja imajo, so pa na trgu zelo iskani. Možnost, ki jo izkoriščamo, je ta, da zaposlujemo diplomante in jim nudimo možnost izobraževanja doma in v tujini, sodelovanje pri mednarodnih projektih in drugo. Povezani smo tudi z nekaterimi fakultetami in inštituti (FRI, Jožef Štefan in drugi), kamor zaposlene pošiljamo na izobraževanja. Seveda imamo tudi veliko mednarodnih izobraževanj. Osebno bi si želel še več povezanosti s fakultetami.*

Blaž Zupan (2017): »Kolikor je meni znano, na humanistiki in družboslovju ne poučujejo predmetov s področja znanosti o podatkih. Tudi v srednjih šolah vsebine s tega področja ni. Možnosti za napredek je torej ogromno.«

6.2 Povzetek ekspertnih intervjujev

Če povzamemo, se sogovorniki strinjajo, da je v Sloveniji premalo kadra, ki bi znal delati z masovnimi podatki, in predvsem premalo izobraževalnih programov, ki bi tak kader lahko sproducirali. V prihodnosti še vedno vidijo kombinacijo obojih – anket in masovnih podatkov – in se strinjajo, da se metodi dopolnjujeta, saj brez anket ne moremo doseči vseh želenih podatkov, kot so stališča. Boro Nikić iz SURS-a zelo dobro primerja ankete in masovne podatke: »*Če ne drugega, bodo ankete (raziskovanja) ostale zato, da se bodo velepodatkovne statistike lahko "validirale".«* Kar se tiče napak, se strinjajo, da je možnost zanje velika, če raziskovalci niso dobro podkovani in slabo zastavijo raziskovalni okvir, ter osmišljanje zbranih podatkov. Ana Slavec pravi: *»Pogosta napaka je, da raziskovalci velepodatkov ne znajo pravilno osmisliti in postaviti pravih raziskovalnih vprašanj ter na koncu interpretirati rezultatov.«*

7 ZAKLJUČEK

V diplomskem delu smo pregledali kaj so masovni podatki, kakšna je njihova vloga v družboslovnem raziskovanju in kako jih lahko povežemo z anketnim raziskovanjem. Pojasnili smo pojem »big data« oziroma masovne podatke ter navedli, na kakšen način so proizvedeni in kako dostopamo do njih. Prav tako smo obravnavali vprašanje kakovosti, napak in zanesljivosti ter zakonskih okvirov in predstavili nekaj primerov praktične uporabe.

Zaključujemo z ugotovitvijo, da masovni podatki ne bodo izpodrinili anket. Njihova prednost v primerjavi z anketami je v tem, da so podatki že zbrani, so poceni in zajemajo velik spekter informacij. Pa vendar ugotavljamo, da prihodnost prinaša tesno sodelovanje anket in masovnih podatkov in le na ta način lahko raziskovalci kakovostno podprejo analize, raziskave in sprejemanje odločitev.

Paziti je potrebno na zakonske okvire, ki so še vedno precej v povojih, predvidevamo pa, da se bo to nadalje uredilo s posodobljeno Splošno uredbo EU o varstvu podatkov (General Data Protection Regulation – GDPR), ki preide v veljavo meseca maja v letu 2018.

Največji oviri pri delu z masovnimi podatki sta pomanjkanje kadra in neznanje. Ker se tehnologija izredno hitro spreminja, zastarajo tudi raziskovalne metode. Japec in drugi (2015) predlagajo vzpostavitev multidisciplinarnih ekip, ki bi z združevanjem masovnih podatkov in anket dosegale najboljše rezultate. Trenutno v Sloveniji, kar se tiče družboslovnega raziskovanja, nimamo študijskega programa, ki bi izobraževal kadre posebej za delo z masovnimi podatki. Na Fakulteti za družbene vede bo v prihodnjih letih sicer stekel nov podiplomski študijski program, ki se bo podrobneje ukvarjal s podatkovno analitiko. Trenutno pa manjkajo vsaj študijski predmeti, ki bi študente seznanili s tehnologijami in uporabo, implementacijo in interpretacijo masovnih podatkov.

Tudi strokovnjaki, s katerimi smo govorili, se strinjajo, da je v Sloveniji premalo kadra, ki bi znal delati z masovnimi podatki, in predvsem premalo izobraževalnih programov, ki bi tak kader lahko usposobil. V prihodnosti še vedno vidijo kombinacijo obojih – anket in masovnih podatkov – in se strinjajo, da se metodi dopolnjujeta, saj brez anket ne moremo dobiti vseh želenih podatkov, kot so npr. stališča.

K vsemu temu lahko dodamo, da na drugi strani nekateri raziskovalci, ki se ukvarjajo s trženjem in oglaševanjem v »data driven« podjetjih, ne verjamejo v daljšo prihodnost anket - kar nekaj

jih je prepričanih, da bodo raziskovalci čez približno deset let za spoznavanje potencialnih kupcev uporabljali zgolj še masovne podatke.

Kljub veliki količini masovnih podatkov so torej ankete potrebne za dodatno osmišljanje podatkov, globlje razumevanje problemov in preverjanje veljavnosti masovnih podatkov – slednje velja tudi obratno. Callegaro in Yang (2016) sta predlagala nov termin za masovne podatke, in sicer bogati podatki oziroma »rich data«. S predlogom se lahko v celoti strinjamo in dodajamo mnenje Japec in drugih (2015), ki težijo k multidisciplinarnosti med anketnim raziskovanjem in masovnimi podatki. Delo torej zaključujemo z ugotovitvijo, da masovni podatki še zdaleč niso zgolj modna muha, temveč hitrorastoča metoda zbiranja podatkov, ki lahko v kombinaciji z anketami nudi poglobljene odgovore na pravilno zastavljena raziskovalna vprašanja.

8 LITERATURA

1. 1KA. 2017. *Splošen opis*. Dostopno prek: https://www.1ka.si/c/694/Orodje_1KA/?preid=695 (26. november 2017)
2. Baker, Reginald P. 2016. Big Data: A Survey Research Perspective. V *Total Survey Error: Improving Quality in the Era of Big data*, ur. Paul P. Biemer, Edith De Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker in Brady West. 47–69. Hoboken (NJ): Wiley.
3. Beyer, Mark A. in Douglas Laney. 2012. *The importance of »Big Data«: A Definition*. Stamford (CT): Gartner.
4. Biemer, P. Paul. 2010. Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly* 74 (5): 817–48.
5. Callegaro, M., Katja Lozar Manfreda in Vasja Vehovar. 2015. *Web Survey Methodology*. London: Sage.
6. Callegaro, Mario in Yongwei Yang. 2017. The role of surveys in the era of »Big Data« V *The Palgrave Handbook Of Survey Research*, ur. D.L., Vannette in Krosnick, J.A. 23. New York: Palgrave.
7. Google. 2017. *Google Trends*. Dostopno prek: https://trends.google.com/trends/explore?cat=174&date=all&q=big%20data,Surveys (19. december 2017).
8. Groves, Robert M. 2011. Three Eras of Survey Research. *Public Opinion Quarterly* 75 (5): 861–71.
9. ---, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer in Roger Tourangeau. 2004. *Survey Methodology*. Hoboken (NJ): Wiley.
10. *Hadoop Apache*. Dostopno prek: http://hadoop.apache.org/ (5. avgust 2017).
11. Halford, Susan in Mike Savage. 2017. Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research. *Sociology* 51 (6): 1132–1148.
12. Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul P. Biemere, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neill in Abe Usher. 2015. Big Data in Survey Research. AAPOR Task Force Report. *Public Opinion Quarterly* 79 (4): 839–80.
13. McFarland, Daniel A, in Richard H McFarland. Big Data and the Danger of Being Precisely Inaccurate. *Big Data & Society* 2 (2):1–4.
14. Nickerson, David W. in Todd Rogers. 2012. Political Campaigns and Big Data. *The Journal of Economic Perspectives* 28 (2): 51–73.

15. Nikić, Boro. 2017. Intervju z avtorico. Elektronska pošta, 29. september.
16. R-project. 2017. *What is R?*. Dostopno prek: https://www.r-project.org/about.html (19. december 2017).
17. Slavec, Ana. 2017. Intervju z avtorico. Elektronska pošta, 1. oktober.
18. Informacijski pooblaščenec Republike Slovenije. 2017. *Splošna uredba EU o varstvu podatkov (GDPR - General Data Protection Regulation).* Dostopno prek: https://www.ip-rs.si/zakonodaja/reforma-evropskega-zakonodajnega-okvira-za-varstvo-osebnih-podatkov/ (26. november 2017).
19. Struijs, Peter, Barteld Braaksma in Piet JH Daas. 2014. Official Statistics and Big Data. *Big Data & Society* 1 (1): 1–6.
20. Waldherr, A., D. Maier, P. Miltner in E. Gunther. 2016. Big Data, Big Noise: The Challenge of Finding Issue Networks on the Web. *Social Science Computer Review* 1 (17): 1–17.
21. Zook, Matthew, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson in Frank Pasquale. 2017. Ten Simple Rules for Responsible Big Data Research. Uredil Fran Lewitter. *PLOS Computational Biology* 13 (3). Dostopno prek: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399 (26. november 2017).
22. Zupan, Blaž. 2017. Intervju z avtorico. Elektronska pošta, 5. september.