

STROJNA PREVEDBA INTERNETNIH NASLOVOV NOVIČARSKIH PRISPEVKOV V NASLOV NA WAYBACK ARCHIVE

Janez ŠTEBE¹

¹ Univerza v Ljubljani, Arhiv družboslovnih podatkov

V članku predstavimo dozdajšnje aktivnosti sodelovanja podatkovnih repozitorijev in znanstvenih revij v Sloveniji pri uvajanju citiranja in dostopa do raziskovalnih podatkov, ki spremljajo objave v znanstveni literaturi. Kot nadaljevanje aktivnosti je ADP vodil pilot sodelovanja z nekaterimi družboslovnimi revijami. Prikazan je primer sodelovanja z revijo Javnost, ki je prestavljal poseben izziv za arhiviranje, saj so bili podatki novičarski prispevki, obravnavani v enem izmed člankov v reviji. Problem za redistribucijo so avtorske in sorodne pravice, obenem pa ni zagotovila, da bodo novičarski portali ohranjali objavljene vsebine. Preprosta koda v programskem jeziku Python je avtomatizirala prenos internetnih naslovov na *Wayback Machine*, v področnem podatkovnem arhivu ADP pa je bil shranjen vzorčni seznam vseh v raziskavi obravnavanih internetnih naslovov, katerih povezani novičarski prispevki predstavljajo raziskovalne podatke danega članka. V razpravi na obravnavanem primeru ponovno izpostavimo prednosti shranjevanja in dostopa do podatkov, ki spremljajo znanstveno literaturo.

Ključne besede: reproduktibilnost, raziskovalni podatki, znanstvena računalniška koda, znanstveno založništvo

1 UVOD

Predstavljamo primer računalniške kode, ki smo jo razvili *ad hoc* za trajno shranjevanje in dostop ene izmed prevzetih enot raziskovalnih podatkov v Arhivu družboslovnih podatkov (ADP). Gre za podatke avtorjev Kuhar in sod. (2021), ki so bili podlaga za članek avtorjev Smrdelj in sod. (2021). Podatki vsebujejo seznam referenc na originalne medijske objave. V nadaljevanju opišemo, v čem je bil problem pri arhivski obravnavi omenjenih podatkov. Za razumevanje širše slike podamo izhodišča in možne načine reševanja problema, kot se pojavljajo v literaturi, utemeljimo izbrano rešitev in

predstavimo podrobnosti izsekov kode ter načina njene uporabe. Rešitev bi bila lahko z malo prilagoditve uporabljena tudi v drugih podobnih primerih.

V razpravi potegnemo vzporednice s podanim problemom ohranjanja referenc na novičarske članke in pomenom shranjevanja računalniške kode za namen ponovne uporabe ali pa za zagotovitev reproduktibilnosti v znanstveni literaturi objavljenih rezultatov analiz.

1.1 Izhodišča

Znanstvene revije in znanstvena skupnost se zavedajo pomena dostopnosti podatkov in tudi računalniške kode v povezavi z znanstveno literaturo. V ospredju so možnosti ponovne rabe podatkov za druge namene kot prvotno, lahko samostojno ali pa v povezavi z drugimi podatki. Pomen dvojice podatkov in računalniške kode, namenjene analizi in izdelavi rezultatov, je tudi v zagotavljanju računske replikabilnosti (ang. computational replicability) (Peer, 2024). Dostopni paket podatkov, računalniške kode in drugih elementov se ugotavljanju računske replikabilnosti podvrže preizkusu, če deluje in daje enake rezultate kot so ti, objavljeni v članku v reviji.

Na različnih področjih znanosti, pa tudi znotraj družboslovja, je zavzetost za reproduktibilnost ponekod večja kot drugod. Ekonomisti so zavzetejši, saj je pogostejša uporaba kvantitativnih podatkov in kompleksnih računalniško podprtih analiz (Hoynes, 2023). Sociologija je glede metodoloških pristopov bolj multiparadigmatska. Vprašanje zagotavljanja reproduktibilnosti še ni v večji meri vstopilo v navodila ter priporočila revij in združenj. Zlasti pri kvalitativnem raziskovanju, ki predstavlja velik del socioloških pristopov, je pomen doseganja transparentnosti raziskovanja drugačen, temu primerno pa so drugačni tudi postopki za njeno preverjanje (Weeden, 2023).

V Sloveniji je s pobudo za ozaveščanje znanstvene skupnosti in revij o pomenu deljenja raziskovalnih podatkov v povezavi s članki začela delovna skupina znotraj RDA-vozlišča Slovenije. V letih 2019 in 2020, ko je bilo delovanje vozlišča financirano prek evropskega projekta, je potekal tudi pilot dela z znanstvenimi revijami. Skupina je na podlagi usmeritev (Hrynaszkiewicz in sod., 2020) *Interesne skupine za standardizacijo in uvajanje podatkovnih politik* (RDA Alliance, 2024) pripravila nacionalnemu okolju prilagojene

Smernice (Štebe in sod., 2020a) in z njimi povezana Pojasnila (Štebe in sod., 2020a). Oba dokumenta sta bila namenjena podpori slovenskim znanstvenim revijam pri uvajanju uredniških pravil glede obveznosti in načina citiranja ter dostopnosti raziskovalnih podatkov. Pri uvajanju Smernic so v okviru pilotnega preizkusa sodelovale posamezne revije, pretežno s področij družboslovja in humanistike. Pokazalo se je, da so glavne ovire pomanjkanje strokovne in tehnične podpore na strani revij ter negativna pričakovanja glede prisotnosti kulture deljenja podatkov področnih znanstvenih skupnosti, ki predstavljajo bazen potencialnih avtorjev teh revij (Štebe in sod., 2020c).

Delo z revijami se je nadaljevalo tudi v naslednjem obdobju. ADP je bil vključen v delovno skupino za sodelovanje z revijami pri evropski družboslovni infrastrukturi Consortium of European Social Science Data Archives (CESSDA). Eno izmed poročil delovne skupine je na podlagi pregleda literature in analize intervjujev z izbranimi predstavniki revij po sodelujočih državah ugotavljalo, da imajo globalni založniki večinoma elaborirane osrednje podatkovne politike, da pa se pozna odsotnost naslavljanja področnih specifik družboslovja, kot je vprašanje, kako ravnati z občutljivimi osebnimi podatki in različnimi novimi vrstami podatkov, ki jih uporabljajo družboslovci. Intervjuji s predstavniki nacionalnih uredništev manjših revij pa so pokazali, da se zavedajo pomena in tovrstnih izzivov deljenja podatkov, da pa se sami ne čutijo dovolj usposobljeni za svetovanje avtorjem glede pristopov priprave podatkov za dostop. Tovrstno svetovanje in podporo avtorjem bi bili pripravljene prepustiti nacionalnim organizacijam izvajalcev podatkovnih storitev znotraj združenja CESSDA (Alvanides in sod., 2020). Glavnina aktivnosti v koordinaciji delovne skupine je bila namenjena nadaljevanju preizkušanja modelov sodelovanja podatkovnih arhivov z revijami v nacionalnih okoljih (Štebe in sod., 2022).

V sklopu teh aktivnosti se je ADP povezal z uredništvom nekaterih družboslovnih revij, med njimi z revijo Javnost/The Public. Vabilo k razmisleku o spodbujanju deljenja podatkov v povezavi s članki so pri tej reviji vzeli resno. Boris Mance, pomočnik glavnega urednika, je predlagal, da se kar takoj lotimo dela. Preden bi pripravili svoja navodila glede dostopa do podatkov na ravni revije, so želeli sami model preizkusiti v omejenem obsegu. Ravno so pripravljali posebno številko revije, ki jo je urejal kot gostujoči urednik (Suplement, Javnost, 2021). Avtorje v posebni številki so pozvali, da naj svoje članke obogatijo s povezavo

na uporabljene podatke, in sicer tako, da podatke objavijo v področnem repozitoriju in jih ustrezno citirajo v seznamu literature. ADP je avtorjem ponudil podporo pri pripravi podatkov za objavo v katalogu in zagotovil prednostno obravnavo predanih podatkov. Avtorji so k predaji podatkov pristopili zavzeto, kar kaže, da so strahovi pred nesodelovanjem avtorjev člankov ob ustreznem pristopu odveč (Bezjak, 2022). Pripomoglo je gotovo tudi dejstvo, da so bili članki v sklepni redakciji in da so že prestali recenzijo, tako da sama objava ni bila vprašljiva ne glede na sodelovanje pri pilotu.

Revija Javnost nastaja pod okriljem mednarodne založbe Routledge, ki je del konglomerata Taylor & Francis (T & F). Založnik T & F sicer ima krovna navodila glede deljenja podatkov v povezavi s članki (T & F, 2024a). Opredeljujejo, kaj so raziskovalni podatki ter kateri so razlogi in koristi za deljenje podatkov. V nadaljevanju pa podajajo zgledno izčrpna in razdelana navodila, ki podrobno opredeljujejo potrebne korake. Opišejo pomen priprave Načrta ravnanja z raziskovalnimi podatki in pridobivanja soglasij za deljenje podatkov, kje podatke objaviti, in podajo pojasnila glede načel FAIR. Poudarijo pomembnost deljenja raziskovalne kode. Predstavijo, zakaj je pomembno povezovanje literature in podatkov z vključevanjem permanentnih identifikatorjev (PID) v metapodatke ene in druge entitete. Podani so napotki o tem, da je treba podatke v članku citirati, in sicer svoje pa tudi če uporabljaš podatke od drugod, ter da mora struktura navajanja podatkov vsebovati PID in mesto, na katerem so podatki in spremljajoči metapodatki dostopni. Navodila so skladna z vse bolj uveljavljenimi priporočili za različna znanstvena področja (za jezikovne podatke primerjajte Lenardič in sod., 2020).

1.2 V čem je problem?

Čeprav revija Javnost/The Public spada pod T & F, ima – kot smo že ugotavljali drugje – svoja posebna Navodila za avtorje (Javnost, 2024), ki sama ne omenjajo pomembnosti ali celo obveznosti citiranja in deljenja podatkov. Edino, kar je na strani Navodil, je povezava na krovno založniško politiko (T & F, 2024b), ki sicer vsebuje razdelek o dostopnosti podatkov in njihovi predaji, vendar je politika nezavezujoča za uredniško avtonomne revije, kakršna je tudi naša.

Ko smo Mancetu in kolegom v uredništvu posebne številke predstavili pobudo

(tudi glede na njihovo pripravljenost za sodelovanje), smo se zavedali, da lahko glede na posebnosti primerov člankov, ki so jih pripravljali za objavo, pričakujemo določene izzive. Iz interne korespondence uredniške skupine, ki smo jo lahko spremljali, je razvidno tehtanje vprašanja, kako s podatki medijskih objav:

Bolj pa imava probleme glede člankov, pri katerih gre za vsebinske analize medijskih objav (kar smo se pogovarjali zadnjič pred faksom). Kaj storiti v takem primeru (X. opozarja, da najbrž nimamo pravice arhivirati medijskih objav in da tu ne gre za podatke, ki jih mi generiramo (kot pri intervjujih))? Obenem gre tudi za javno dostopne podatke, torej v takem primeru arhiviranje ni tako nujno? Ali pač? (Štebe, 2021a; osebna imena zakrita).

Izpostavljeni sta dve dilemi. Prva je povezana z vprašanjem pravic (avtorskih in sorodnih) glede arhiviranja in ponovnega objavljanja podatkov člankov, ki so sicer že bili objavljeni drugje, vendar pod različnimi licencami, lahko tudi z omejitvijo redistribucije ali pa brez kakršnih koli oznak glede razpolaganja. Če bi hoteli to razčistiti, bi morali, tako kot pri objavah na socialnih medijih, preveriti, kakšne so licence, in druge morebitne omejitve pravice uporabe (Breuer in sod., 2021; Hagen in sod., 2019; ToS, Terms of Service, Thompson, 2016). Ob omejitvi razpolaganja bi morali, če bi hoteli zagotoviti pravno nedvoumno situacijo in brez tveganja za kršenje pravic, z lastniki medijev doseči dogovor, ki bi dovolil tovrstno uporabo. Ali pa tvegati in računati na to, da zaradi premajhnega interesa nihče izmed lastnikov vsebin ne bi oporekal arhiviranju in dajanju podatkov v dostop za raziskovalne namene, v smislu uporabe v javnem interesu (SSRC, 2019). Lahko pa bi vsebine arhivirali in določili omejen dostop pod posebnimi pogoji, s katerimi bi moral uporabnik soglašati.

Druga dilema v navedku zgoraj je, ali vsebine glede na to, da so že bile javno objavljene, sploh predstavljajo podatke, ki bi jih bilo treba arhivirati in ponovno objaviti pod titulo raziskovalnih podatkov, saj so vsebine že tako ali tako dostopne vsakomur.

1.3 Predlagana rešitev problema

V nadaljevanju bomo prikazali in utemeljili rešitev, ki se je zdela

najprimernejša glede na naravo podatkov in z vidika učinkovite obravnave v arhivu podatkov ADP.¹

Predlagali smo, da se arhivira samo navedke medijskih objav skupaj s povezavami na originalne podatke. Avtorji so namreč že razpolagali z naborom vseh navedkov, ki so ustrezale iskalnim pogojem v določenem obdobju. Dodatno je bilo potrebno upoštevati trajne formate datotek in pripraviti metapodatke.

Argument, da so vsebine medijskih objav že objavljene in se ni potrebno ukvarjati z njihovim dostopom, ima tudi svoje slabost. Pristop z arhiviranjem povezav na objavljene vsebine se zgleduje po ponujenih rešitvah za arhiviranje objav na *Twitterju* (zdaj *X*). Shrani se množica enkratnih identifikacijskih števil vsake izmed objav, ki nato prek API-ja omogoča rekonstrukcijo vsebine in z njo povezanih podatkov (Kaczmirek in sod., 2014; Kinder - Kurlanda in sod., 2017). Tako kot pri vsebinah socialnih medijev, ki se jih *rehidrirajo* na podlagi identifikatorja, je tudi pri povezavah na originalne objave težava, da lahko vsebina ni več dostopna pod določenim naslovom. Objavo na socialnem mediju je lahko uporabnik izbrisal in ni več dostopna. Portal medijskih objav lahko zamenja naslov, stare objave lahko briše, sam medij lahko propade in vsebine niso več dostopne. Da bi rešili še to težavo, saj je cilj nekoliko trajnejši dostop, smo vse vsebine, ki so bile v aktualnem času dostopne, označili za avtomatizirano preselitev na spletni arhiv *Wayback Machine* (2024).

Wayback Machine (WM) je projekt neprofitne digitalne knjižnice *The Internet Archive*. Med prednostmi storitve je, da ima poslanstvo trajnega shranjevanja in da deluje v javnem interesu, kar je tudi pravna podlaga upravičevanja morebitnih izjem glede GDPR in vprašanj pravic intelektualne zaščite. Prav tako pa ponuja različne možnosti strojnega dostopa do zbirk, ki poleg internetnih strani vključujejo tudi zbirke TV-vsebin, e-knjig, programske kode idr.

¹ Glede na jezikovno naravo podatkov bi prišlo v poštev tudi, da bi dajalce podatkov usmerili na repozitorij CLARIN.SI za jezikovne vire in orodja, ki deluje znotraj evropske raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN ERIC, vendar bi s tem samo preložili ukvarjanje s problemom na nekoga drugega, ob tem, da smo avtorjem pri reviji Javnost zagotovili podporo pri pripravi podatkov.

Rezultat izbora medijskih objav predstavlja enkraten izsek določene tematike v določenem časovnem obdobju, tudi zemljepisno zamejeno. Edinstvenost množice objav in njihova raba v kontekstu raziskave pa jim pridajata status raziskovalnih podatkov. Podatki so lahko namenjeni preverjanju in reproduktibilnosti objavljenih ugotovitev ali pa se jih uporabi v nadaljnjih raziskavah z drugačnimi izhodišči in analitičnim pristopom kot izvirno, lahko pa se jih uporabi za primerjavo z drugimi podatki.

Preiskali smo tudi možnost uporabe slovenskega Spletnega arhiva NUK. Kolega Janko Klasinc, spletni arhivar na NUK, s katerim smo se pogovarjali, je kot možnost ponudil prenos vsebin na pobudo, kar bi pomenilo predajo urejenega seznama internetnih naslovov sodelavcem NUK za prenos na njihov Spletni arhiv z njihovim dodatnim naporom indeksacije seznama. Prednost NUK je, da deluje na zakonski podlagi, saj leta 2009 dopolnjeni Zakon o obveznem izvodu publikacij dovoljuje zajem in dostop do arhiviranih javno objavljenih spletnih vsebin.² Omejitev je, da je zajem selektiven, saj temelji na presoji pomena za slovensko kulturno dediščino. Glavna omejitev je, da ne omogoča avtomatizacije zajema na zahtevo prek API. Kolega Klasinc je prijazno za preizkus dejansko poindeksiral nabor enega dela referenc na spletni strani za NUK-ov Spletni arhiv. V naši medsebojni korespondenci je zapisal:

Najprej sem nameraval predlagati, da počakamo na zaključek naše indeksacije in potem primerjamo naše zajeme z njihovimi obstoječimi [mišljeni zajemi na WM, op. a.]. Ampak če pri njih omogočajo brezplačen in enostaven zajem posameznih strani, ki še niso zajete, bi bila to morda boljša rešitev za vas (Štebe, 2021b).

2 OPIS PODATKOV IN METODE SKUPAJ Z RAČUNALNIŠKO KODO

Osnova za prikaz priprave podatkov za podatkovni arhiv so podatki, zbrani za članek o objavah o istospolnih osebah in tujcih med obravnavo zakonov, ki zadevajo obe marginalizirani skupini, v letih 2015 (Zakon o zakonski zvezi in družinskih razmerjih) in 2017 (Zakon o tujcih) (Smerdelj in sod., 2021). Po

² <https://web.archive.org/web/20230531195146/https://arhiv.nuk.uni-lj.si/zakonska-podlaga>

določenih kriterijih in s ključnimi besedami so generirali vzorec objav na štirih novičarskih spletnih portalih: *RTVSLO.si*, *24ur.com*, *SiOL.net* in *Nova24TV.si*. Vzorec je skupaj obsegal 393 enkratnih enot.

Količina enot je bila prevelika za ročno iskanje, če so objave že prenešene v spletni arhiv WM. Če še niso, bi bilo treba ročno vnesti predlog na WM za prenos objave. Odločili smo se, da pripravimo preprosto kodo v programu Python, ki bo postopek avtomatizirala.

V arhivski obravnavi smo dve prejeti besedilni datoteki z navedki zadetkov iz Worda najprej pretvorili v enostavno tekstovno datoteko s končnico *.txt ('interd17_p1_sl_v1_r1.txt' in 'interd17_p2_sl_v1_r1.txt' (Štebe, 2024)), ki se lahko bere prek računalniške kode (tabela 1).

Tabela 1: Izsek treh enot iz dokumenta vzorca navedkov v enostavnem tekstovnem formatu (datoteka 'interd17_p1_sl_v1_r1.txt' (Štebe, 2024)).

1) MMC RTV SLO

1) Je napočil čas, ko bodo istospolni pari poročeni in ne registrirani? (19. 1. 2015) <https://www.rtv slo.si/slovenija/je-napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-registrirani/356207>

2) "Če hočeš ekonomski razvoj, je enakost ne glede na spolno usmerjenost prava pot" (9. 2. 2015) <https://www.rtv slo.si/slovenija/ce-hoces-ekonomski-razvoj-je-enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921>

3) Janša: Nismo brez moči, zato tudi ne brez odgovornosti (14. 2. 2015) <https://www.rtv slo.si/slovenija/jansa-nismo-brez-moci-zato-tudi-ne-brez-odgovornosti/358382>

Programska koda v Python (datoteka 'read_ln_txt.py' v Štebe, 2024) bere zgornjo datoteko vrstico po vrstico in ignorira vrstice, ki ne vsebujejo internetnega naslova. Te vrstice samo po vrsti prepíše v izhodno datoteko.

V vrstici, ki vsebuje internetni naslov, najprej poišče prosti zapis datuma objave in ga pretvori v računalniški datumski format. Koda izlušči internetni naslov originalne objave. Nato z internetnim naslovom naredi poizvedbo na API WM. Pri tem uporabi funkcije iz paketa v Pythonu 'waybackpy' (Mahanty, 2022). Zapisi na *The Internet Archive* API-ju se vrnejo v lokalno opredeljenem formatu CDX. Na voljo je vrsta drugih možnosti za dostopanje do vsebin enot

in metapodatkov prek API-jev, vključno s pretvorbo v format JSON (IA, 2024), ki pa nas tokrat niso zanimala, saj smo želeli čim bolj preprosto in hitro rešitev za naš konkretni namen. Če naslov že prenešen v spletni arhiv, ga program samo zabeleži v izhodnem dokumentu v obliki za citiranje, kot je najbližja datumu objave na WM. Če še ni prenešen, pa da koda ukaz za prenos naslova na WM in po uspešnem prenosu vsebine strani prav tako poleg izvirne zabeleži novo obliko naslova. Pred vsako novo poizvedbo program določi odmor, saj je API na WM občutljiv, če so poizvedbe iz istega naslova prepegoste, in se lahko preneha odzivati.

3 REZULTAT OBRAVNAVE

Vseh 393 izvirnih internetnih naslovov je bilo med preizkusom še aktivnih. Tisti, ki še niso bili do takrat, so bili na novo uspešno prenešeni v WM. Večina naslovov je bila prenešana že predhodno, med manjkajočimi pa so bili večinoma iz *Siol.net*. Manjkajoči naslovi bi bili potencialno izpostavljeni izgubi, če ne bi bilo aktivno spodbujenega prenosa. Obdelava ene in druge datoteke z naslovi je bila popolnoma avtomatizirana, zahvaljujoč tudi doslednosti avtorjev pri navajanju URL-naslovov.

Z naslovi na WB obogateni podatki navedkov skupaj z metapodatki, ki vsebujejo opis raziskave in metode po strukturi Data Documentation Initiative (DDI), so bili po končani arhivski obravnavi dostopni na ADP (Kuhar in sod., 2021). Sami podatki v navedkih naslovov poleg izvirnega naslova vsebujejo še naslov na WM (tabela 2).

Tabela 2: Izsek končne izhodne datoteke vzorca navedkov z dodanim naslovom na spletnem arhivu *Wayback Machine* (krepko označeno) (Kuhar in sod., 2021).

1) MMC RTV SLO

1) Je napočil čas, ko bodo istospolni pari poročeni in ne registrirani? (19. 1. 2015) <https://www.rtvsllo.si/slovenija/je-napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-registrirani/356207>
<<https://web.archive.org/web/20150122200530/http://www.rtvsllo.si:80/slovenija/je-napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-registrirani/356207>>
2) "Če hočeš ekonomski razvoj, je enakost ne glede na spolno usmerjenost prava pot" (9. 2. 2015)
<https://www.rtvsllo.si/slovenija/ce-hoces-ekonomski-razvoj-je->

enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921
<<https://web.archive.org/web/20150710101105/http://www.rtv slo.si/slovenija/ce-hoces-ekonomski-razvoj-je-enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921>>
3) Janša: Nismo brez moči, zato tudi ne brez odgovornosti (14. 2. 2015) <https://www.rtv slo.si/slovenija/jansa-nismo-brez-moci-zato-tudi-ne-brez-odgovornosti/358382>
<<https://web.archive.org/web/20150217101345/http://www.rtv slo.si:80/slovenija/jansa-nismo-brez-moci-zato-tudi-ne-brez-odgovornosti/358382>>

Sama storitev spletnega arhiva WB ima z vidika trajnega shranjevanja tudi svoje pomanjkljivosti. Besedilo se v spletnem arhivu prenese in prikaže neokrnjeno. Videovsebine pa se ne prenesejo, kot kaže podrobna analiza enega izmed primerov³ zajema spletnih strani medijske objave na *24ur.com*. V tem primeru se moramo vprašati, kaj so tiste 'pomembne lastnosti' (*significant properties*), ki si jih prizadevamo trajno ohraniti. V našem primeru sta ohranjena samo besedilni povzetek posnetka oddaje in spremljajoča slika.

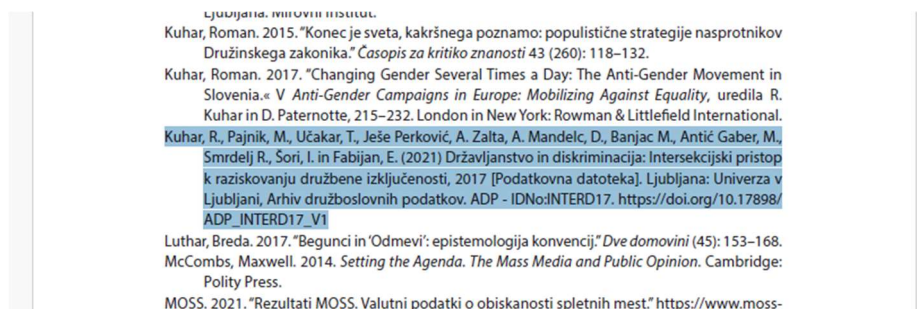
4 RAZPRAVA IN ZAKLJUČEK

Med prednostmi arhiviranja podatkov medijskih objav je v povezavi s člankom v reviji tudi možnost zabeležke veliko več referenc, kot pa jih je uporabljenih v članku. Ta namreč vsebuje samo reference na neposredne navedbe dela novičarskega besedila v samem članku, ki so namenjene ilustraciji ugotovitev, dobljenih z metodo kritičnih študijev diskurza (KŠD) (Smrdelj in sod., 2021). Vsakdo na s člankom povezanih arhiviranih podatkih preveri ustreznost navedkov v smislu replikacije. Razširjen nabor referenc, ki je zajet v arhiviranih podatkih, omogoča nadaljnje analize pa tudi oceno nepristranosti originalne analize.

Podatki so kot vzporedna objava o članku tudi nagrada sodelujočim avtorjem, ki so brez dvoma v pripravo podatkov za predajo v arhiv morali vložiti dodane napore. Podatke je po objavi v arhivu mogoče navajati v literaturi, kar lahko pomeni, da so avtorji lahko deležni novih citatov. Po načelih citiranja podatkov so tudi avtorji zadevnega članka (Smrdelj in sod., 2021) v literaturi navedli lastne uporabljene podatke (slika 1).

³ Primerjajte <https://www.24ur.com/novice/slovenija/med-begunci-tudi-lastniki-tovarn-organizatorji-porok-knjizni-zalozniki-stand-up-komiki.html> in <<https://web.archive.org/web/20170104043031/https://www.24ur.com/novice/slovenija/med-begunci-tudi-lastniki-tovarn-organizatorji-porok-knjizni-zalozniki-stand-up-komiki.html>>

Slika 2: Navajanje v članku uporabljenih podatkov v seznamu literature.



Ta članek je hkrati tudi prikaz, kako lahko tudi programsko kodo arhiviramo in navajamo v literaturi. Skripta v Python je dostopna na GitHub in arhivirana na Zenodo. Alternativna lokacija za arhiviranje kode je po analogiji z internetnim arhivom Software Heritage Archive (2024). Preprosta skripta v Pythonu se lahko dopolni in ponovno uporabi na novih gradivih drugje.

Druga analogija z našim pristopom je pobuda za urejanje kompleksnih navedb, ki so jo predstavili na zadnji konferenci RDA (Stall, 2024). Na veliko znanstvenih področjih se v članku uporablja veliko količino referenc na druge podatke. Pobuda na podoben način, kot smo pokazali pri tem članku, zbere vse reference skupaj s stalnimi identifikatorji v enem samem dokumentu, ki tako postane nov predmet shranjevanja in citiranja.

Preizkus sodelovanja avtorjev posebne številke revije Javnost z arhivom podatkov je dal pozitivne rezultate obogatenih vsebin člankov. Še pomembnejša pa je popotnica pri nadaljnji širitvi sodelovanja arhivov in revij pri vzpostavitvi podatkovnih politik ter njihovi izvedbi. Področni podatkovni arhiv, kot je ADP, je pripravljen svetovati avtorjem in ob posebnih izzivih, kot je bil ta, poiskati priročne rešitve, ki so skladne z zakoni ter učinkovito zagotovijo dostop do podatkov in virov za namene replikacije ali nove analize.

5 LITERATURA

- Alvanides, S., Štebe, J., Morgan de Paula, E., Kleiner, B., Hegedus, P., Glavica, M. in Kranjec, I. (2021). Assessment of Journal Requirements and Needs. CESSDA Work Plan 2020 (3.0). Zenodo. <https://doi.org/10.5281/zenodo.5554436>
- Breuer, J., Borschewski, K., Bishop, L., Vávra, M., Štebe, J., Strapcova, K. in Hegedus, P. (2021). Archiving Social Media Data: A guide for archivists and researchers (1.2). Zenodo. <https://doi.org/10.5281/zenodo.6517880>
- Hagen, S., Bishop, E., Koščik, M., Vavra, M., Štebe, J., Ryan, L., Payne, E., Løvlie, A.,

- Straume, Ø., Rød, L. M. in Høgetveit Myhren, M. (2019). Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data. Deliverable 6.3 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Pridobljeno 31. maja 2024. https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies...__FINAL.pdf
- Hoynes, H. (2023). Reproducibility in Economics: Status and Update. *Harvard Data Science Review*, 5(3). <https://doi.org/10.1162/99608f92.80a1b88b>
- Hrynaskiewicz, I., Simons, N., Hussain, A., Grant, R. in Goudie, S. (2020). Developing a Research Data Policy Framework for All Journals and Publishers. *Data Science Journal*, 19(1). <http://doi.org/10.5334/dsj-2020-005>
- IA, The Internet Archive (2024). Internet Archive Developer Portal. Pridobljeno 31. julija 2024. <https://archive.org/developers/>
- Javnost – The Public (2024). Instructions for authors. Updated 28th February 2024. <https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=rjav20>
- Kaczmarek, L., Mayr, P., Vatrapu, R., Bleier, A., Blumenberg, M., Gummer, T., Hussain, A., Kinder-Kurlanda, K., Manshaei, K., Thamm, M., Weller, K., Wenz, A. in Wolf, C. (2014). Social media monitoring of the campaigns for the 2013 German Bundestag Elections on Facebook and Twitter (GESIS – Working Papers No. 31). <https://www.gesis.org/en/services/sharing-knowledge/publications/archive/gesis-working-papers>
- Kinder - Kurlanda, K., Weller, K., Zenk - Möltgen, W., Pfeffer, J. in Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, 4(2), 1–14. <https://doi.org/10.1177/2053951717736336>
- Kuhar, R., Pajnik, M., Učakar, T., Ješe Perković, A., Zalta, A., Mandelc, D. in Šori, I. (2021). Državljanstvo in diskriminacija: Intersekcijski pristop k raziskovanju družbene izključenosti, 2017 [Podatkovna datoteka]. Ljubljana: Univerza v Ljubljani, Arhiv družboslovnih podatkov. https://doi.org/10.17898/ADP_INTERD17_V1
- Lenardič, J., Erjavec, T. in Fišer, D. (2020). Citiranje jezikovnih podatkov v slovenskih znanstvenih objavah v obdobju 2013–2019. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 8(1), 1–34. <https://doi.org/10.4312/slo2.0.2020.1.1-34>

- Mahanty, A. (2022). Waybackpy 3.0.6 [paket Python kode].
<https://pypi.org/project/waybackpy/>
- Peer, L. (2024). Why and How We Share Reproducible Research at Yale University's Institution for Social and Policy Studies. *Harvard Data Science Review*, 6(1).
<https://doi.org/10.1162/99608f92.dca148ba>
- RDA Alliance. (2024). Data policy standardisation and implementation IG. Pridobljeno 31. maja 2024, <https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig/>
- Smrdelj, R., Učakar, T. in Kuhar, R. (2021). Marginalizacija intersekcijske perspektive: istospolna partnerstva in tujci na slovenskih novičarskih spletnih portalih. *Javnost – The Public*, 28(sup1), S122–S140.
<https://doi.org/10.1080/13183222.2021.2012945>
- Software Heritage archive. Pridobljeno 31. maja 2024, <https://archive.softwareheritage.org/>
- Bezjak, S. (2022). Izzivi publiciranja in citiranja podatkov v družboslovju. Pridobljeno 31. maja 2022. <https://www.adp.fdv.uni-lj.si/blog/2022/05/31/25-let-adp/>
- SSRC and QDR (2019). Addressing Copyright (managing-qualitative-data.org). Pridobljeno 31. maja 2024, <https://managing-qualitative-data.org/modules/3/c/>
- Stall, S. (2024). Complex Citations: Working Towards Recommendations. Pridobljeno 31. maja 2024, <https://www.rd-alliance.org/plenaries/complex-citations-working-towards-recommendations/>
- Suplement, Javnost (2021). Suplement, Javnost – The Public, Vol. 28 – 2021, Suplement, Pridobljeno 31. maja 2024, <https://javnost-thepublic.org/issue/2021/5/>
- Štebe, J. (2021a). Zadeva: Arhiviranje podatkov v ADP – Suplement Javnosti. Osebna korespondenca avtorja s člani uredniškega odbora.
- Štebe, J. (2021b). Zadeva: Arhiviranje spletnih člankov, ki so podlaga znanstvenemu članku. Osebna korespondenca avtorja z Jankom Klasincem.
- Štebe, J. (2024). Read web address from the text file and write add Wayback Archive address (v1.0.2). Zenodo. <https://doi.org/10.5281/zenodo.10992412>
- Štebe, J., Bezjak, S., Alvanides, S., Recker, J., Glavica, M., Kranjec, I., Laaksonen, H., Kondyli, D., Klironomos, N., Linardis, A. in Kleiner, B. (2022). The

- CESSDA Data Archives joint efforts to support journals in data sharing and reproducibility [Panel]. Zenodo. <https://doi.org/10.5281/zenodo.6683796>
- Štebe, J., Bezjak, S. in Dolinar, M. (2020a). Smernice za oblikovanje politik znanstvenih založb glede navajanja raziskovalnih podatkov v znanstvenih publikacijah in zagotavljanja dostopa do primarnih podatkov, uporabljenih v člankih (2.4). Zenodo. <https://doi.org/10.5281/zenodo.3757288>
- Štebe, J., Bezjak, S. in Dolinar, M. (2020b). Pojasnila za razumevanje politik znanstvenih založb glede zagotavljanja dostopa do primarnih podatkov, uporabljenih v člankih (1.0). Zenodo. <https://doi.org/10.5281/zenodo.3754783>
- Štebe, J., Bezjak, S., Alvanides, S., Recker, J., Glavica, M., Kranjec, I., Laaksonen, H., Kondyli, D., Klironomos, N., Linardis, A. in Kleiner, B. (2022). The CESSDA Data Archives joint efforts to support journals in data sharing and reproducibility [Panel]. Zenodo. <https://doi.org/10.5281/zenodo.6683796>
- Štebe, J., Dolinar, M., Bezjak, S. in Inkret, A. (2020c). Implementing the RDA Research Data Policy Framework in Slovenian Scientific Journals. *Data Science Journal*, 19(1), 49. <https://doi.org/10.5334/dsj-2020-049>
- T & F (2024a). Taylor & Francis Editorial Policies. Pridobljeno 31. maja 2024, <https://authorservices.taylorandfrancis.com/editorial-policies/>
- T & F (2024b). Sharing and citing data. Pridobljeno 31. maja 2024, <https://authorservices.taylorandfrancis.com/data-sharing/>
- Thomson, S. D. (2016). Preserving Social Media. Digital Preservation Coalition. DOI: <http://dx.doi.org/10.7207/twr16-01>
- Wayback Machine (2024). Pridobljeno 31. maja 2024, <https://archive.org/web/>
- Weeden, K. A. (2023). Crisis? What Crisis? Sociology's Slow Progress Toward Scientific Transparency. *Harvard Data Science Review*, 5(4). <https://doi.org/10.1162/99608f92.151c41e3>

MACHINE TRANSFER OF INTERNET ADDRESSES OF THE ONLINE NEWS ON THE WAYBACK ARCHIVE ADDRESSES

Past activities of collaboration of Slovenian research data repositories and scientific journals regarding data citation access to data used in the article is described. Recent continuation of those collaboration lead to some interesting use cases. We describe the pilot activity with the journal Javnost – The Public. A challenge for archival were online news content used in the analysis of one of the articles. Copyright and special terms of media are a limitation regarding content redistribution, on the other hand there is no guarantee that the portal will preserve the content for future. Simple Python code made the transfer of internet addresses of the online news on the Internet Archive. The disciplinary data archive preserves the list of sample addresses of news used in the article as research data, described its origin and accessible for reuse. Discussion put emphasis on importance of preserving and access to data related to scientific literature.

Keywords: reproducibility, research data, scientific software code, scientific publishing

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva –
Deljenje pod enakimi pogoji 4.0 Mednarodna.

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0
International.

<https://creativecommons.org/licenses/by-sa/4.0/>

