Procedura pohrane i objave podataka u arhivu CROSSDA

2. prosinca 2022.
Radionica za zaposlenike Instituta za razvoj i međunarodne odnose

Vedran Halamić

“AS OPEN AS POSSIBLE, AS CLOSED AS NECESSARY”

_svi podaci u društvenim znanostima ne mogu biti potpuno otvoreni

- Podaci sadrže informacije pomoću kojih je moguće identificirati pojedinca
- Istraživači su sudionicima u istraživanju obećali da će se podaci koristiti samo u određene svrhe

_ipak, podatke je moguće pohraniti i objaviti pomoću 3 osnovna mehanizma:

- Informirani pristanak (etički i pravni)
- Anonimizacija
- Regulaciju pristupa

_o regulaciji pristupa brinu arhivi podataka

- CROSSDA trenutno pohranjuje i objavljuje samo podatke koji ne sadrže osobne podatke
- Sljedeći korak u razvoju arhiva CROSSDA je implementacija mehanizama kontrole pristupa podacima koji sadrže osobne podatke (udaljeni pristup ili sigurna soba)

_moguće je objaviti dvije verzije podataka - jednu koja je “maksimalno” anonimizirana, drugu koja sadrži potencijalne identifikatore i dostupna je uz registraciju i kontrolu

REGULIRANJE PRISTUPA PODACIMA U ARHIVU CROSSDA

A. Slobodno dostupni podaci (otvoreni pristup)

- Dostupni pod uvjetima licence Creative Commons Autorstvo 4.0 Međunarodna (CC BY 4.0)

B. Podaci dostupni uz registraciju

- Podaci dostupni samo za znanstvene svrhe
  - Priprema i provedba znanstvenih istraživanja (projekti, radovi, doktorske disertacije)
  - Provjera rezultata iznesenih u znanstvenom radu (recenzenti ili drugi istraživači)
  - Izrada diplomskih i poslijediplomskih specijalističkih radova

- Podaci dostupni za znanstvene svrhe i za korištenje u nastavi
  - Osim u znanstvene svrhe, dozvoljeno je i korištenje u nastavi (npr. vježbe iz statistike, studentski seminarski radovi, završni radovi na razini preddiplomskih studija)

PROCEDURA POHRANE

_istraživač šalje upit o pohrani na arhiv.podataka@ffzg.hr

_dogovara se sastanak s osobljem arhiva, razgovor o:

- Pripremi podataka
- Razinama pristupa podacima
- Uvjetima korištenja podataka
- Sljedećim koracima

_potpisuje se izjava o pohrani

- Istraživač daje dozvole za korištenje podataka i arhivu i krajnjim korisnicima
- Voditelj arhiva potvrđuje da su podaci zaprimljeni i da će s podacima postupati kako je dogovoreno u izjavi

- Predložak izjave: za otvorene podatke / samo u znanstvene svrhe / u znanstvene svrhe te u svrhu podučavanja i učenja

PROCEDURA POHRANE

_istraživač šalje zahtjev za otvaranjem korisničkog računa u Dataverse-u

- Potrebno je prijaviti se u sustav korištenjem AAI@EduHr korisničkog računa

_osoblje arhiva dodjeljuje istraživaču potrebne ovlasti u Dataverse-u

_istraživač kreira skup podataka i prenosi datoteke s podacima i popratnom dokumentacijom u Dataverse

- Podaci u SPSS-u ili STATA-i se automatski konvertiraju u otvoreni format (tekstualna datoteka s podacima + metapodaci u DDI formatu)
- Za skup podataka rezervira se DOI broj

_istraživač unosi podatke o skupu podataka u Dataverse

- Upute za izradu metapodataka

PROCEDURA POHRANE

_osoblje arhiva pregledava sadržaj predanih datoteka

- Jesu li sve varijable i kodovi za vrijednosti varijabli jasno dokumentirani?
- Jesu li u podacima sadržane vrijednosti koje nisu navedene u kodovima?
- Je li dostavljena sva potrebna dokumentacija za razumijevanje skupa podataka?
- Je li postignuta odgovarajuća razina anonimizacije?

_osoblje arhiva šalje istraživaču izvještaj o skupu podataka sa savjetima za doradom

_istraživač ispravlja podatke i nadopunjuje dokumentaciju prema savjetima i dostavlja novu verziju datoteka

PROCEDURA POHRANE

_osoblje arhiva pregledava sadržaj predanih datoteka

- Jesu li sve varijable i kodovi za vrijednosti varijabli jasno dokumentirani?
- Jesu li u podacima sadržane vrijednosti koje nisu navedene u kodovima?
- Je li dostavljena sva potrebna dokumentacija za razumijevanje skupa podataka?
- Je li postignuta odgovarajuća razina anonimizacije?

_osoblje arhiva šalje istraživaču izvještaj o skupu podataka sa savjetima za doradom

_istraživač ispravlja podatke i nadopunjuje dokumentaciju prema savjetima i dostavlja novu verziju datoteka

PROCEDURA POHRANE

_osoblje arhiva preimenovat će datoteke u skladu s internim pravilima

- Dodaje se prefiks datotekama (arhivski broj + oznaka za vrstu datoteke)
  npr. cda1010_dat_SOCRES_RapidCATI_SUF11032021_hr.tab

_ukoliko je potrebno, datoteke se konvertiraju u formate prikladnije za korištenje i/ili dugoročnu pohranu

_osoblje arhiva kontrolira i nadopunjuje metapodatke (ključne riječi iz tezaurusa, predmetna klasifikacija i drugo)

_prije objave, zapis se šalje na provjeru

_nakon što istraživač odobri promjene, osoblje arhiva objavljuje skup podataka

Preduvjeti za pohranu podataka: čišćenje, dokumentacija, anonimizacija.

PREDUVJETI ZA POHRANU PODATAKA

01 Organizacija podataka

02 Dokumentacija

03 Čišćenje podataka

04 Anonimizacija

PREDUVJETI ZA POHRANU PODATAKA

_dobra organizacija datoteka s podacima i popratnom dokumentacijom u kojima se (vi sami) možete snaći

_dovoljno dokumentacije za razumijevanje sadržaja datoteka s podacima

_podaci očišćeni i pripremljeni za analizu

_anonimizirani podaci

_istraživanje provedeno u skladu s etičkim načelima i pravnim propisima (GDPR)

_razjašnjena prava vlasništva nad podacima i popratnom dokumentacijom

_CESSDA Data Management Expert Guide https://dmeg.cessda.eu/

ORGANIZACIJA PODATAKA

_organizacija varijabli

_struktura foldera

_imenovanje datoteka

_verzioniranje

_formati podataka

Održani CROSSDA webinari:

Kako organizirati podatke u istraživačkom projektu?
Rad s tabličnim podacima

_trivijalan zadatak ili ozbiljan organizacijski problem?

_savjet: jedna osoba u projektu neka bude zadužena za organizaciju podataka

DOKUMENTACIJA

_informacije o projektu/istraživanju

_metodološki izvještaj: opis uzorka i uzorkovanje, korišteni instrumenti, korišteni softver za prikupljanje podataka, opis postupka prikupljanja podataka, upute za anketare i kontrola anketara, odaziv...

_kodne knjige (značenje varijabli i kodova)

_upitnici, pokazne kartice

_bibliografski zapisi o publikacijama u kojima su korišteni podaci

_dokumentacija o transformacijama datoteke s podacima

ČIŠĆENJE I PRIPREMA ZA ANALIZU

_pod očišćenim podacima podrazumijevamo ispravljene podatke, očišćene od neispravnih, nepotpunih nepravilno oblikovanih ili duplih zapisa podataka

_sve promjene u podacima u odnosu na sirovu verziju treble bi biti evidentirane, najbolje uz pomoć sintakse statističkog program koji je za to korišten

Održana radionica: Per Rspera ad astra: Priprema podataka pomoću programskog jezika R

QAMyData

- “Data health check” alat razvijen unutar UK Data Service-a
- Omogućava provjeru kvalitete podataka/identificira česte probleme unutar skupa podataka
- Bez grafičkog sučelja (u razvoju), parametri se unose uređivanjem config file-a
- Nudi nekoliko vrsta automatske “provjere” stanja podataka

```yaml
---
##################################################
## QAMYDATA: Health Checks for Your Data Files ##
##################################################

# Welcome to the default configuration (config) file for QAMYDATA.
# The file is written in YAML (YAML Ain't Markup Language), which is a human-readable language commonly used for configuration files.
# The config is divided into 4 types of tests: Basic File Checks, Metadata Checks, Data Integrity Checks and Disclosure Control Checks.
# Lines starting with '#' are comments so they are ignored.


########################
## Basic File Checks ##
########################

basic_file_checks:
  # Checks whether the file name contains illegal/odd/non-compliant characters
  bad_filename:
    setting: "^([a-zA-Z0-9]+)\\.([a-zA-Z0-9]+)$"
    desc: "File name should match the user specified pattern"

######################
## Metadata Checks ##
######################

metadata:
  # Checks high-level grouping (for example, useful if dataset can be grouped by household)
  primary_variable:
    setting: HouseholdID
    desc: "Counts the unique occurrences for the grouping variable specified"

  # Checks whether any variables do not have labels
  missing_variable_labels:
    setting: true
    desc: "Variables should have a label"

  # Checks whether any user-defined missing values do not have labels (sysmis) - SPSS only
  value_defined_missing_no_label:
    setting: true
    desc: "User-defined missing values should have a label (SPSS only)"
```

QAMyData

teaching-data%set.sav

Raw Case Count: 10210
Aggregated Case Count: 0
Total Variables: 188
Data Type Occurrences: Numeric: 186, String: 2
Created At: 2019-02-18 13:37:39
Last modified at: 2019-02-18 13:37:39
File Label:
File Format Version: 2
File Encoding: WINDOWS-1252
Compression type: Rows

Basic File Checks

Name	Status (N)	Description
Bad file name	failed (1)	File name should match the user specified pattern

Metadata Checks

Name	Status (N)	Description
Missing variable labels	failed (8)	Variables should have a label
Variable odd characters	failed (2)	Variable names and labels should not contain the specified characters ["&", "#", " ", "@", "*", "ç", "ð", "ü"]
Variable label max length	failed (6)	Variable labels should not exceed the defined number of characters (79 characters)

ANONIMIZACIJA

_anonimizacija je rezultat obrade osobnih podataka kako bi se nepovratno spriječilo utvrđivanje identiteta sudionika istraživanja

_podrazumijeva uklanjanje direktnih i/ili indirektnih identifikatora iz skupa podataka pri čemu se podaci mogu izmijeniti ili drugačije organizirati kako ne bi bilo moguće otkrivanje identiteta sudionika istraživanja

_onemogućavanje izdvajanja sudionika u svrhu reidentifikacije (k-anonimnost)

_jedan od ciljeva anonimizacije podataka jest omogućavanje dijeljenja podataka u svrhu istraživanja uz minimalan rizik od reidentifikacije

_proces anonimizacije uvelike ovisi o kontekstu te je ispravnu odluku moguće donijeti razmatrajući ne samo podatke, već i okolnosti njihovog prikupljanja i pohrane

ANONIMIZACIJA

_sve više podataka se prikuplja, pohranjuje i analizira što otvara nove mogućnosti povezivanja podataka

_briga oko privatnosti podataka

_etička odgovornost istraživača -> anonimizacija predstavlja etički problem o kojemu treba razmišljati kroz cijeli istraživački proces

_zakonski okvir (GDPR)

_važnost drugih “sigurnosnih” mjera (informirani pristanak i reguliranje pristupa podacima)

ANONIMIZACIJA

“The anonymisation Decision-Making Framework: European Practitioners’ Guide” (Elliot, Mackey & O’Hara, 2020)

Data situation audit -> Risk analysis and control -> Impact management

- “Potpuna” anonimizacija nije moguća/korisna

_apsolutna, formalna, statistička ili **funkcionalna** anonimizacija?

_potreba za holističkim pristupom anononimizaciji

https://dmeg.cessda.eu/Data-Management-Expert-Guide/5.-Protect/Anonymisation

ANONIMIZACIJA - ALATI

sdcMicro

- Open source softverski alat (R paket) za anonimizaciju
- Koristi jednostavno grafičko sučelje (nije potrebno predznanje kodiranja)
- Uključuje više metoda/algoritama i mjera rizika
- Korisna dokumentacija (https://sdcpractice.readthedocs.io/en/latest/sdcMicro.html)

Amnesia

- Softverski alat za anonimizaciju razvijen u sklopu OpenAire-a
- Koristi grafičko sučelje, iznimno jednostavan za korištenje
- Rad u oblaku, podatke je potrebno uploadati

ARX

- Sveobuhvatan softver za anonimizaciju

sdcMicro GUI

About/Help | Microdata | Anonymize | Risk/Utility | Export Data | Reproducibility | Undo

What do you want to do?

- Display microdata
- Explore variables
- Reset variables
- Use subset of microdata
- Convert numeric to factor
- Convert variables to numeric
- Modify factor variable
- Create a stratification variable
- Set specific values to NA
- Hierarchical data

Reset inputdata

Loaded microdata

The loaded dataset is `testdata` and consists of `4580` observations and `15` variables. No variables were dropped because of all missing values.

Show 20 entries

Search:

urbrur	roof	walls	water	electcon	relat	sex	age	hhcivil	expend	income	savings	o
2	4	3	3	1	1	1	46	2	90929693	57800000	116258.5	
2	4	3	3	1	2	2	41	2	27338058	25300000	279345	
2	4	3	3	1	3	1	9	1	26524717	69200000	5495381	
2	4	3	3	1	3	1	6	1	18073948	79600000	8695862	
2	4	2	3	1	1	1	52	2	6713247	90300000	203620.2	
2	4	2	3	1	2	2	47	2	49057636	32900000	1021268	
2	4	2	3	1	3	2	13	1	63386309	22700000	8119166	
2	4	2	3	1	3	2	19	1	1106874	89100000	9881406	

Showing 1 to 20 of 4,580 entries

Previous 1 2 3 4 5 … 229 Next

View/Analyze existing sdcProblem

Show summary

Explore variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PRAM (simple)

PRAM (expert)

Supress values with high risks

Anonymize numerical variables

Top/bottom coding

Microaggregation

Adding noise

Rank swapping

Reset SDC problem

Summary of dataset and variable selection

The loaded dataset consists of `4580` records and `15` variables.

Categorical key variable(s): `urbrur`

Numerical key variable(s): `income savings`

Computation time

The current computation time was ~ `0.06 seconds` .

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
urbrur	2 (2)	2290.000 (2290.000)	646 (646)

Risk measures for categorical key variables

We expect `2.00` (`0.04%`) re-identifications in the population, as compared to `2.00` (`0.04%`) re-identifications in the original data.

`0` observations have a higher risk than the risk in the main part of the data, as compared to `0` observations in the original data.

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	0 (0.000%)
3-anonymity	0 (0.000%)	0 (0.000%)
5-anonymity	0 (0.000%)	0 (0.000%)

REFERENCE/KORISNI LINKOVI

- Benschop, T. (2021) sdcMicro GUI manual documentation.
- Benschop, T., Machingauta, C., & Welch, M. (2019). Statistical disclosure control: A practice guide.
- CESSDA Training Team (2017 - 2022). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Retrieved from https://dmeg.cessda.eu/
- Elliot, M., Mackey, E., & O'Hara, K. (2020). The anonymisation decision-making framework 2nd Edition: European practitioners' guide.
- Elliot, M., Mackey, E., O'Hara, K & Tudor, C. (2016). The anonymisation decision-making framework. UKAN Publications.
- Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. Computer Law & Security Review, 34(2), 204-221.
- EU Data Protection Working Party (2022, October 3) European Commission. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Hvala

arhiv.podataka@ffzg.hr
crossda.hr