

# Viri podatkov

**1. srečanje: Uvodne teme  
(predstavitev načina dela in obveznosti  
študentov pri predmetu; razdelitev tem;  
osnovni pojmi)**

*Mojca Bavdaž ([mojca.bavdaz@ef.uni-lj.si](mailto:mojca.bavdaz@ef.uni-lj.si))*



Petek, 19.2.2021

# Izvajalci

## Mojca Bavdaž

E-mail: [mojca.bavdaz@ef.uni-lj.si](mailto:mojca.bavdaz@ef.uni-lj.si)

Skype: mojcab

Tel.: 01 5892-630

Lokacija: EF, R-412

Govorilne ure: po dogovoru

## Janez Štebe

E-mail: [janez.stebe@fdv.uni-lj.si](mailto:janez.stebe@fdv.uni-lj.si)

Tel.: 01 5805-292

Lokacija: FDV, C 233

Govorilne ure: torek 10:00 – 12:00

IŠ 1. torek v mesecu 15:00 – 16:00

## Gostje iz prakse





**Kakovost  
podatkov**

**Interpretacija  
Komunikacija**

# Teme

## Kakovost podatkov

- Uradna statistika (tudi vir ekonomskih podatkov)
- Viri podatkov s področja javnega zdravja in medicine
- Masovni podatki

Iskanje in ponovna uporaba podatkov

Pravni in etični vidiki uporabe podatkov

## Prikazovanje podatkov



# Izvedba predmeta

- Kontaktne ure kot kombinacija:
  - predavanj z aktivno udeležbo študentov (diskusije, predstavitve)
  - seminarjev z gosti
  - individualnega in skupinskega reševanja nalog/problemov
- Sprotno delo študentov:
  - Priprave na kontaktne ure
  - Predstavitve
  - Seminaraska naloga
- Izpit



# Struktura ocene

---

Izpit	50 točk
-------	---------

Sprotno delo	50 točk
--------------	---------

---

*Iz vsakega dela je potrebno doseči vsaj 25 točk*

---

---

Ocena:

10	nad 90 točk
----	-------------

9	nad 80 točk
---	-------------

8	nad 70 točk
---	-------------

7	nad 60 točk
---	-------------

6	nad 50 točk
---	-------------

---

# Sprotno delo

Max 50 točk.

- A. Priprava na uro s področja uradne statistike (do 5)
- B. Predstavitev primera »big data« (do 10)
- C. Priprava na uro s področja družboslovnih podatkov (do 5)
- D. Seminarska naloga (do 25)
  - Osnutek grafične predstavitve (do 5)
  - Končna predstavitev z zagovorom (do 20)
- E. Končni trije aha momenti (do 5)

---

+ Izpitni bonus (do 10 izpitnih točk)

# Pregled pomembnih datumov

Datum dogodka ali rok oddaje	Oblika dela	Točke
Rok oddaje: sreda, 24.2.	A. Priprava na uro s področja uradne statistike	5
Rok oddaje do predavanj petek, 5.3.	Bonus	10 (izpitne)
Predstavitev in rok oddaje: petek, 19.3.	B. Predstavitev primera big data	10
Rok oddaje: sreda, 31.3.	C. Priprava na uro s področja družboslovnih podatkov	5
Konzultacije: petek, 23.4.	D. Osnutek grafične predstavitve	5
Predstavitev in rok oddaje: petek, 7.5.	D. Končna predstavitev z zagovorom	20
Rok oddaje: petek, 7.5.	E. Končni aha moment	5



# A. Priprava na uro s področja uradne statistike

## Viri:

The Economist (26. maj 2018). Plunging response rates to household surveys worry policymakers / Don't even ask! Dostopno na

<https://www.economist.com/international/2018/05/24/plunging-response-rates-to-household-surveys-worry-policymakers>

Evropski statistični sistem (2017). Kodeks ravnanja evropske statistike. Brošura dostopna na

<https://www.stat.si/statweb/FundamentalPrinciples/CodeOfPractice> ALI <http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice>

Seljak, R. (6. junij 2020). O neodvisnosti in ugledu uradne statistike. *Delo. Sobotna priloga*.

# A. Priprava na uro s področja uradne statistike

## Naloga:

Napišite eno stran vsebinskega poročila, povzetka (ne copy-paste), ki bo zajel bistvo vseh štirih virov.

Navedite tri »aha momente«.

Navedite eno vprašanje, ki bi ga želeli prediskutirati.

## Oddaja:

Format: pdf

Oddajte na spletni strani predmeta v poglavju Sprotno delo

Rok: **sreda, 24.2.**

# Bonus

**Petek, 5.3., 12:00:** Metka Zaletel (*Nacionalni inštitut za javno zdravje, Zdravstveno podatkovni center*)

Naloga (rok oddaje pred predavanjem na e-mail [mojca.bavdaz@ef.uni-lj.si](mailto:mojca.bavdaz@ef.uni-lj.si)):

- Od treh oseb zbrati podatke za vprašalnik NIJZ.
- Pripraviti poročilo (max 2 strani) o izkušnji z zbiranjem podatkov (izbor oseb, razmerje do osebe, zadržki glede podajanja podatkov, trajanje, občutki anketarja in anketirancev po izvedbi, ocena kakovosti - popolnosti in resničnosti – podatkov) ter razmislekom, kako na osnovi tovrstnih podatkov oblikovati ukrepe.
- Udeležba na predavanju gosta in sodelovanje v diskusiji.
- Do 10 izpitnih točk (šteje v kvoto 25 minimalno zahtevanih izpitnih točk; nadomešča izpitna vprašanja na to temo).

## B. Predstavitev primerov »big data«

### Naloga:

Ustna predstavitev: max 3 min (pitch)

- vsebinski opis (vira) podatkov s poudarkom na opisu načina zbiranja oz. mehanizma, kako podatki nastanejo;
- dobre in slabe plati vira oz. njegove uporabe;
- ključna spoznanja.

Zastavljeno vsaj eno vprašanje na predstavitvah.

Pisni izdelek (2 – 3 strani).

**Ustna predstavitev:** v okviru srečanja 19.3.

**Oddaja pisnega izdelka:** 19.3.

Format: pdf

Oddajte na spletni strani predmeta v poglavju Sprotno delo.

## B. Predstavitev primerov »big data«

1. Cene
2. Facebook in družba
3. Google Flu Trends
4. Humanizacija masovnih podatkov
5. Netflix
6. Pametne stavbe
7. Pristranskost zaradi kvantifikacije
8. Promet
9. Target

# C. Priprava na uro s področja družboslovnih podatkov (1)

CESSDA Data Management Expert Guide (izbrana poglavja)

1. Plan

(<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan>)

2. Organise & Document

(<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document>)

5. Protect

(<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/5.-Protect>)

6. Archive & Publish

(<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish>)

7. Discover

(<https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover>)

# C. Priprava na uro s področja družboslovnih podatkov (2)

## RDA Recommendations

1. An open, universal literature-data cross-linking service - RDA/WDS Publishing Data Services WG Recommendations, DOI: <http://dx.doi.org/10.15497/RDA00002>
2. Developing a Research Data Policy Framework for All Journals and Publishers, DOI: [10.5334/dsj-2020-005](https://doi.org/10.5334/dsj-2020-005)
3. Eleven Quick Tips for Finding Research Data, DOI: [DOI: 10.1371/journal.pcbi.1006038](https://doi.org/10.1371/journal.pcbi.1006038)
4. Engaging Researchers with Data Management: The Cookbook, DOI: [DOI: 10.11647/OBP.0185](https://doi.org/10.11647/OBP.0185) (izbrano poglavje)
5. FAIR Data Maturity Model: specification and guidelines, DOI: [10.15497/rda00050](https://doi.org/10.15497/rda00050)
6. Metadata Standards Directory Working Group Recommendations
7. Repository Audit and Certification DSA–WDS Partnership WG Recommendations, DOI: [Requirements: https://doi.org/10.17026/dans-22n-gk35](https://doi.org/10.17026/dans-22n-gk35)
8. Sharing COVID-19 Epidemiology Data, DOI: [10.15497/rda00049](https://doi.org/10.15497/rda00049)
9. The final version of the RDA COVID-19 Recommendations and Guidelines for Data Sharing, published 30 June 2020 (drugo poglavje)
10. Wheat Data Interoperability Recommendations, DOI: [DOI: http://dx.doi.org/10.15497/RDA00018](http://dx.doi.org/10.15497/RDA00018)

# C. Priprava na uro s področja družboslovnih podatkov (3)

## Naloga:

Vsak študent povzame eno poglavje/priporočilo na eni strani (max dva študenta na temo).

Navedite tri »aha momente«.

Navedite eno vprašanje, ki bi ga želeli prediskutirati.

## Oddaja:

Format: pdf

Oddajte na spletni strani predmeta v poglavju Sprotno delo

Rok: **sreda, 31.3.**



## D. Seminarska naloga

**Vir podatkov:** SLORA: Slovenija in rak. Epidemiologija in register raka. Onkološki inštitut Ljubljana. [www.slora.si](http://www.slora.si)

**Namen:** Razvoj veščin za grafično in vsebinsko predstavitev podatkov

**Cilj:** Priprava grafične predstavitve za splošno in strokovno javnost vključno s kratkimi vsebinskimi pojasnili in z metodološkimi opombami glede značilnosti in omejitvami podatkov.

**Koraki priprave:**

1. Oblikovanje skupin.
2. Izbira vsebinskega vidika in določitev obsega naloge ter pridobitev potrditve na [mojca.bavdaz@ef.uni-lj.si](mailto:mojca.bavdaz@ef.uni-lj.si).
3. Diskusije na predavanjih in individualne konzultacije.
4. Predstavitev.
5. Oddaja končnega dokumenta.

## D. Seminarska naloga (alternativa)

- **Replikacijska raziskava:** ponovitev analiz iz enega članka
- **Združevanje, harmonizacija in urejanje podatkov iz več virov, ali več vrst podatkov, ali podatkov na različnih populacijah na teme v povezavi s COVID-19 in cepljenjem.**

### Koraki priprave:

1. Izbira vsebinskega vidika in določitev obsega naloge ter pridobitev potrditve na [janez.stebe@fdv.uni-lj.si](mailto:janez.stebe@fdv.uni-lj.si).
2. Diskusije na predavanjih dr. Štebeta in individualne konzultacije.
3. Predstavitev.

## E. Končni aha momenti

Kaj se vam bo najbolj vtisnilo v spomin pri tem predmetu?

Navedba treh a-ha momentov in argumentacija.

# Izpitni roki

Razpisani izpitni roki:

- 1.6. (tor)
- 22.6. (tor)
- 25.8. (sre)

*Dodatni predrok v sredini maja?*



# Sledilnik COVID-19

**Cilj:** povečati ozaveščenost o pomenu kakovosti podatkov in potrebni infrastrukturi

**Izvedba:** Študenti se razdelijo v tri skupine

## Naloga 1:

Viharjenje možganov: pripravite čimveč vprašanj na temo zbiranja podatkov o COVID-19.

Ko boste imeli vsaj 25 vprašanj, jih razvrstite v zaokrožene skupine.

# Kakovost podatkov

Relativna: odvisna od namena

*Primer: merjenje temperature*

- Vrsta spremenljivke
  - Opisna
  - Številska
- Merska lestvica
  - Imenska (nominalna)
  - Urejenostna (ordinalna)
  - Razmična (intervalna)
  - Razmernostna (proporcionalna)

# Kakovost podatkov

Večdimenzionalna:

- Točnost (angl. accuracy)
- Ustreznost statističnih konceptov (angl. relevance)
- Veljavnost merjenja (angl. validity)
- Zanesljivost (angl. reliability)
- Pravočasnost in točnost objave (angl. timeliness and punctuality)
- Skladnost med podatki (angl. coherence)
- Primerljivost v času in prostoru (angl. comparability)
- Dostopnost in jasnost (angl. accessibility and clarity)

# I. Način pridobivanja

- Primarni podatki  
zbrani za znan namen  
*field research*: (neomejeno) ustvarjaš vprašanja
- Sekundarni podatki  
iz obstoječih virov  
ponovno uporabljeni  
*desk research*: izbiraš med razpoložljivimi  
vprašanji/opazovanji



# Ponudniki sekundarnih podatkov

ponujajo primarno zbrane podatke

ponujajo podatke iz sekundarnih virov



# Prednosti sekundarnih podatkov

Ker so že zbrani in (če) jih zberejo kompetentni ponudniki:

- prihranki
- ne obremenjujejo poročevalskih enot
- namen raziskave ne vpliva na odgovore poročevalskih enot, torej potencialno boljše razkritje informacij
- večji doseženi vzorci
- dodatne spremenljivke/informacije za triangulacijo
- pogosto edini vir za longitudinalno in mednarodno analizo
- stalnost in periodičnost zbiranja podatkov, predvidljivost objave

# Slabosti sekundarnih podatkov

Izvirni »greh«: nezmožnost vplivanja na metodologijo in odsotnost nadzora nad kakovostjo:

- neskladnost s potrebami tokratnega raziskovanja
- neustreznost osnovnih opredelitev koncepta, enote, ciljne populacije itd.
- nezadostna razčlenjenost podatkov
- prepozne objave
- problemi časovnih serij (dolžina, prelomi)
- neprimerljivost merskih enot
- dostop

## II. Namen uporabe

Kvantitativni podatki (izraženi številsko, numerično)

Kvalitativni podatki (izraženi opisno, nenumerično)

Narava podatka načeloma izhaja iz obravnave in namena uporabe:

Ali želimo pojav opisati ali ugotoviti pogostost?

Vprašanje:

Ali kvantitativni podatki temeljijo na kvalitativni presoji?

Ali lahko kvalitativne podatke spremenimo v kvantitativne?

## III. Stopnja razkritja

### Mikro podatki

- neanonimizirani mikro podatki (data enclaves),
- anonimizirani mikro podatki,
- statistično zaščiteni mikro podatki,
- public use microdata/files

### Agregirani podatki

- v tabelah/bazah
- v grafičnih prikazih

## IV. Obseg podatkov

Small data

Big data = 3V

(Gartner. IT Glossary: Big data. Najdeno na <http://www.gartner.com/it-glossary/big-data/>)

high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation.

+ Veracity etc.



# Druge možne tipologije

- geografska pokritost1:
  - lokalni (občinski, regionalni)
  - državni
  - mednarodni
  - globalni, svetovni
- geografska pokritost2:
  - slovenski (SURS)
  - tuji (Statistics Netherlands)
- geografska pokritost3:
  - evropski (SURS, Eurostat)
  - neevropski (US Census)
- število vključenih držav:
  - nacionalni (SURS, ONS = Britanski statistični urad, INSEE = francoski statistični urad)
  - mednarodni (Eurostat; World Bank)
- jezik
  - enojezični (angleški, kitajski, španski, ruski, portugalski, francoski...)
  - dvo- in večjezični

- področje
  - tehnični (ELES)
  - ekonomski (DURS)
  - naravoslovni (ARSO)
  - družboslovni (Facebook)
- raven obravnave:
  - mikro podatki (gospodinjstvo, podjetje)
  - mezo podatki (občina, dejavnost)
  - makro podatki (država)
- zavezanost Zakonu o varstvu osebnih podatkov
  - osebni podatki (katerikoli podatki, ki se nanašajo na posameznika, ne glede na obliko, v kateri so izraženi)
  - neosebni, drugi podatki
- preverjenost, zanesljivost:
  - preverjeni, recenzirani podatki (učbenik)
  - nepreverjeni podatki, podatki z neznano zanesljivostjo (Wikipedia)



- (zaveстно) sodelovanje udeležениh oseb
  - s privolitvijo in sprotim nadzorom nad posredovanimi podatki (anketiranje, intervju)
  - s privolitvijo in brez sprotnega nadzora nad posredovanimi podatki (kartice zvestobe, piškotki)
  - brez privolitve (štetje prometa)
- vrsta zapisa
  - digitalni
  - analogni (zdravniške kartoteke)
- oblika komuniciranja oz. izražanja:
  - številski
  - tekstovni
  - slikovni
  - zvočni
  - multimedia (kot kombinacija različnih oblik podatkov)
- vključenost enot:
  - populacijski (Poslovni register Slovenije)
  - vzorčni (Raba Interneta v Sloveniji)
- sprotnost zajema:
  - registrski in registrirani podatki (CRP, skenirano trgovsko blago)
  - z zamikom (Anketa o potrošnji gospodinjstev)

- časovno sledenje:
  - presečni
  - longitudinalni
- periodika zbiranja:
  - enkratni podatki
  - večkrat zbrani podatki (različne občasne ankete)
  - periodični podatki (RIC)
- aktualnost:
  - arhivski podatki (zgodovinski zapisi)
  - sodobni podatki (socialna omrežja)
- formalne implikacije:
  - statistični podatki (Anketa o delovni sili, Statistični del Poslovnega registra Slovenije)
  - administrativni podatki (CRP)
- stroški dostopa:
  - plačljivi podatki (GURS)
  - brezplačni podatki (ESS)
- oviranost dostopa:
  - prost dostop
  - dostop z registracijo, identifikacijo

- vrsta ponudnika:
  - uradni (AJPES)
  - komercialni/profitni (gvin.com)
  - nekomercialni (Oqali (transformed food products))
- sektor lastnika:
  - javni (ARSO)
  - zasebni (bančni izpiski)
- dostopnost:
  - zunanji/javno dostopni (Wikipedia, izmerjena temperatura zraka)
  - notranji/zasebni (klicane mobilne številke, izmerjena telesna temperatura)
- število uporabljenih virov
  - posamezni
  - združeni
    - avtomatsko združeni (prepletene storitve=mashup)
    - preverjeno združeni (OECD)

# Varstvo osebnih podatkov

**Cilj:** povečati ozaveščenost o problematiki varstva osebnih podatkov

**Izvedba:** Študenti se razdelijo v tri skupine

## Naloga 2:

Viharjenje možganov: identificirajte čim več situacij, v katerih se postavi vprašanje varstva osebnih podatkov.

Ko boste imeli vsaj 25 situacij, določite (etične, pravne, praktične) dileme, ki se pri teh situacijah pojavljajo.

# Gosti – obvezna udeležba!

- 5.3. Metka Zaletel (*Nacionalni inštitut za javno zdravje, Zdravstveno podatkovni center*)
- 12.3. Luka Renko in Maja Založnik (*COVID-19 Sledilnik*)
- 16.4. Urban Brulc (*samostojni svetovalec Informacijske pooblaščenke RS*)
- 23.4. Vesna Zadnik (*Onkološki inštitut Ljubljana, Epidemiologija in register raka*)

# Citiranje podatkovnih nizov (APA stil)

Skupno ne glede na vrsto vira:

Author/Rightsholder. (Year). *Title of data set* (Version number) [Description of form].

Dodatek glede na vrsto vira:

Fizični vir: Location: Name of producer.

Elektronski vir – naslov vira: Retrieved from <http://>

Elektronski vir – naslov strani: Available from <http://>

Za neobjavljene podatke:

Author, F. N. (Year). [Description of study topic].

Unpublished raw data.

# Citiranje podatkovnih nizov (APA stil)

## Primeri:

United States Department of Housing and Urban Development. (2008). *Indiana income limits* [Data file]. Retrieved from [http://www.huduser.org/Datasets/IL/IL08/in\\_fy2008.pdf](http://www.huduser.org/Datasets/IL/IL08/in_fy2008.pdf)

Pew Hispanic Center. (2008). *2007 Hispanic Healthcare Survey* [Data file and code book]. Available from <http://pewhispanic.org/datasets/>

Smith, J.A. (2006). [Personnel survey]. Unpublished raw data.

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies. (2013). *Treatment episode data set -- discharges (TEDS-D) -- concatenated, 2006 to 2009* [Data set]. doi:10.3886/ICPSR30122.v2

# Citiranje grafičnih podatkov (APA stil)

## Primeri:

Centers for Disease Control and Prevention. (2005). [Interactive map showing percentage of respondents reporting "no" to, During the past month, did you participate in any physical activities?]. Behavioral Risk Factor Surveillance System. Retrieved from <http://apps.nccd.cdc.gov/gisbrfss/default.aspx>

Solar Radiation and Climate Experiment. (2007). [Graph illustration the SORCE Spectral Plot May 8, 2008]. Solar Spectral Data Access from the SIM, SOLSTICE, and XPS Instruments. Retrieved from [http://lasp.colorado.edu/cgi-bin/ion-p?page=input\\_data\\_for\\_spectra.ion](http://lasp.colorado.edu/cgi-bin/ion-p?page=input_data_for_spectra.ion)