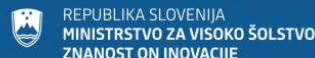


Načrtovanje in pridobivanje raziskovalnih podatkov

Doc. dr. Ana SLAVEC, Fakulteta za matematiko, naravoslovje in
informacijske tehnologije Univerze na Primorskem in InnoRenew CoE

Ljubljana in online, Centralna tehniška knjižnica Univerze v Ljubljani,
Usposabljanje podatkovnih strokovnjakov, 14. – 17. 10. 2024





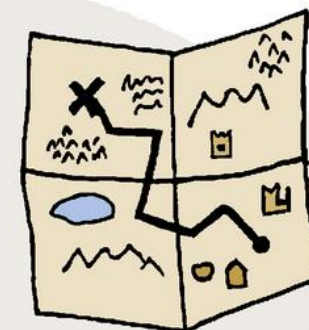
Vsebina

- Pregled obstoječih podatkov in ocena možne uporabe sekundarnih podatkov
- Izбира metod, instrumentov, predvideni dobljeni podatki, parametri meritev in ocena kakovosti podatkov
- Ustvarjanje metapodatkov
- Provenienca in opis podatkov
- Načrtovanje obdelave podatkov
- Predelava surovih podatkov

DATA



METADATA



 Dataedo /cartoon

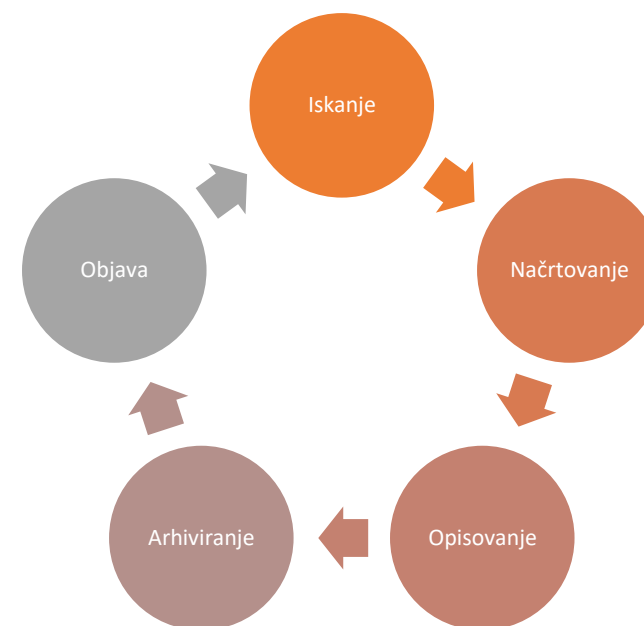
Piotr@Dataedo

Vir: [Dataedo.com](https://dataedo.com)



Življenjski krog raziskovalnih podatkov

- **Iskanje in ponovna raba podatkov**
- **Načrt ravnanja s podatki**
- **Opis podatkov**
- Arhiviranje podatkov
- Objava podatkov





Viri sekundarnih podatkov – nekaj primerov

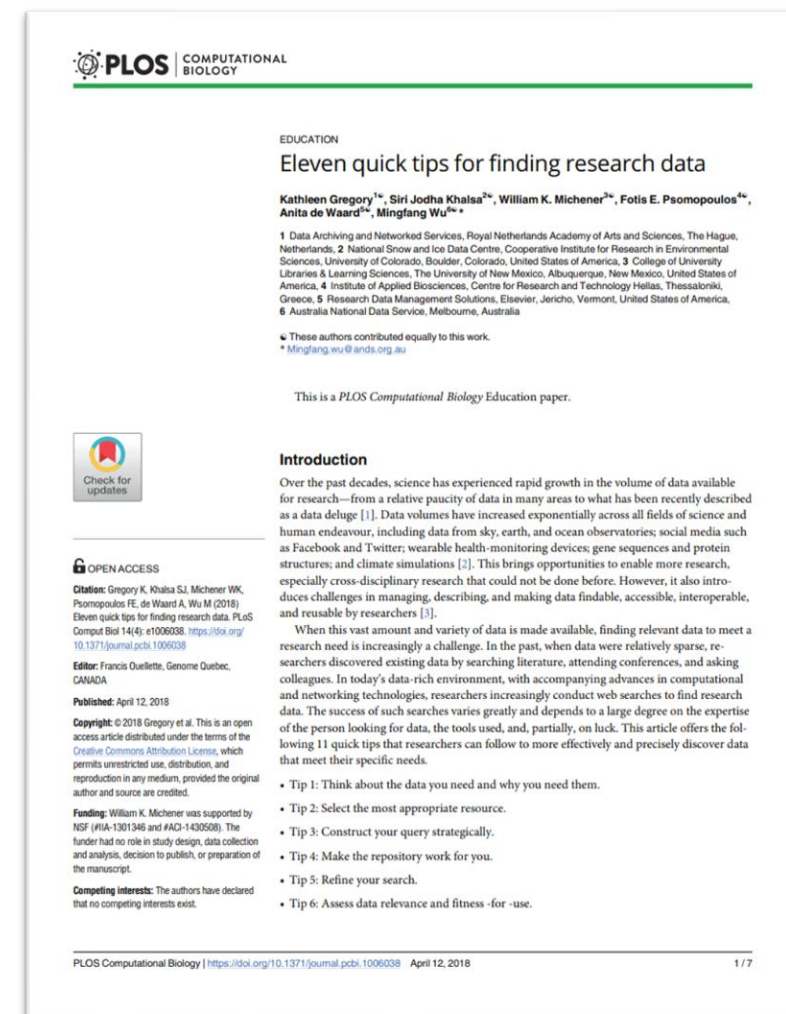
- **Uradni registri** (registri prebivalstva, popisi, matične knjige, poslovni registri, zemljiška knjiga, itd.) **in drugi uradni podatki** (nacionalni statistični uradi, Eurostat, itd.)
- **Mednarodne organizacije** (Združeni narodi, Organizacija za gospodarsko sodelovanje in razvoj, Svetovna banka, itd.)
- **Javne organizacije** (arhivi, agencije, izobraževalne ustanove, knjižnice, muzeji, zdravstvene ustanove, itd.)
- **Podatki podjetij** (letna poročila, podatki iz analitskih orodij)
- **Spletni iskalniki** (npr. [Google DataSet Search](#))
- **Arhivi in repozitoriji raziskovalnih podatkov**
- **Podatki v znanstvenih in strokovnih člankih** (vključno s podatkovnimi objavami, podatki v glavnem besedilu, v prilogah, dopolnilnih materialih)
- ...



11 nasvetov glede iskanja raziskovalnih podatkov

- 1: Premislite, katere podatke potrebujete in zakaj.
- 2: Izberite najprimernejše vire.
- 3: Strategično oblikujte svoje iskalno povpraševanje.
- 4: Naj repozitorij deluje vam v prid.
- 5: Izboljšajte svoje iskanje.
- 6: Ocenite njihovo ustreznost in primernost za uporabo.
- 7: Shranite podrobnosti iskanja in vira podatkov.
- 8: Išcite podatkovne storitve, ne le podatkov.
- 9: Spremljajte najnovejše podatke.
- 10: Občutljive podatke obravnavajte odgovorno.
- 11: Vračajte skupnosti (navedite vir in delite podatke).

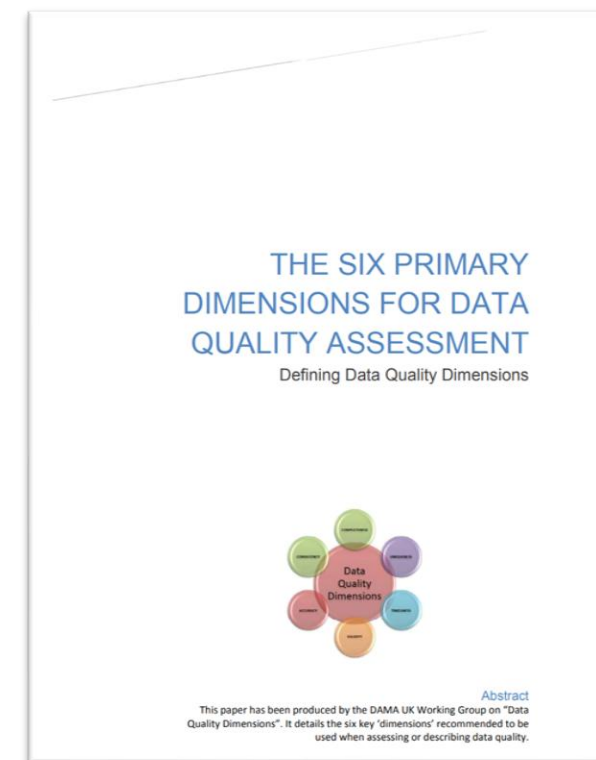
Vir: [Eleven quick tips for finding research data](https://doi.org/10.1371/journal.pcbi.1006038) (Gregory idr. 2018)





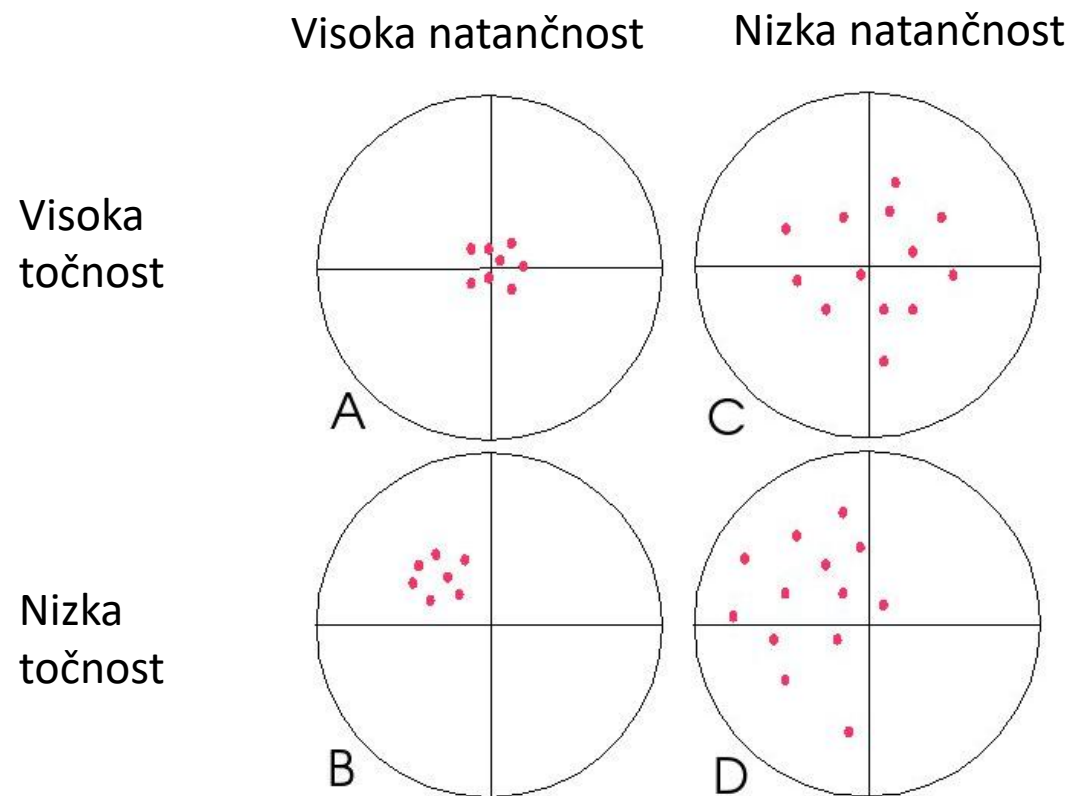
Kakovost podatkov

- **Kakovost podatkov** je opredeljena kot primernost podatkov za določen namen uporabnika. Koncept je relativno odvisen od standardov, ki jih opredeljujejo pričakovanja uporabnikov.
- [6 dimenzij, kot jih opredeljuje Data Management Association UK:](#)
 1. **Popolnost** (prisotnost podatkov v ustreznih poljih)
 2. **Edinstvenost** (brez podvajanja zapisov)
 3. **Pravočasnost** (so na voljo, ko jih uporabniki potrebujejo)
 4. **Veljavnost** (ustrezanje postavljenim pravilom)
 5. **Natančnost** (odražanje resničnosti)
 6. **Doslednost** (skladnost med različnimi viri)





Natančnost in točnost podatkov





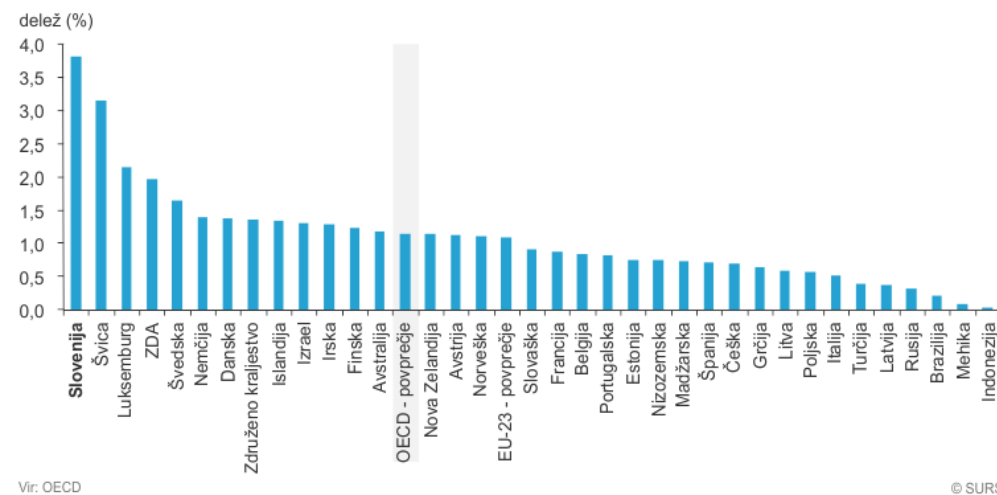
Različne vrste napak

- **Slučajne** napake so nepredvidljive in naključne, npr. rahel tresljaj, ki povzroči nepravilnost pri merjenju.
- **Sistematične** napake se zgodijo zaradi napake v postopku ali sistemu, npr. napačna kalibracija instrumentov ali napačna konfiguracija sistema.
- **Nenamerne napake** zaradi nepozornosti, pomanjkanja informacij ali razumevanja, ali napačnih predpostavk
- **Namerne napake** zaradi [različnih tipov goljufij z raziskovalnimi podatki \(izmišljanje, potvarjanje, zavajanje, plagiarizem, sabotaža\)](#) (Vir: DiRROS Data)



Primer napake zaradi napačnih predpostavk

Delež odraslih, starih 25–64 let, z doktorsko ali primerljivo izobrazbo, države OECD, 2018



METODOLOŠKO OPOZORILO

Za statistično-analične namene se izidi izobraževanja v Sloveniji razvrščajo po Klasifikacijskem sistemu izobraževanja in usposabljanja (KLASIUS), natančneje s prvim podsistemom KLASIUS-SRV glede na segment, raven ter vrsto. KLASIUS-SRV vsebuje osem ravni, dve od teh tvorita po dve podravni. Po klasifikaciji KLASIUS-SRV se izidi terciarnega izobraževanja v Sloveniji razvrščajo na tri ravni: šesto (Visokošolska izobrazba prve stopnje in podobna izobrazba), sedmo (Visokošolska izobrazba druge stopnje in podobna izobrazba) in osmo (Visokošolska izobrazba tretje stopnje in podobna izobrazba). Doktorat znanosti, prejšnji in tretje bolonjske stopnje, se razvršča v najvišjo, osmo raven, natančneje v podraven 8/2 (Doktorat znanosti in podobna izobrazba). Za mednarodno poročanje pa izide izobraževanja razvrščamo po Mednarodni standardni klasifikaciji izobraževanja (ISCED). Po zdaj veljavni verziji ISCED 2011 se izidi terciarnega izobraževanja razvrščajo na štiri ravni: peto (Short-cycle tertiary education), šesto (Bachelor's or equivalent level), sedmo (Master's or equivalent level) in osmo (Doctoral or equivalent level). Doktorat znanosti se razvršča v najvišjo, osmo raven.

Ker klasifikaciji KLASIUS-SRV in ISCED nista neposredno primerljivi, se za mednarodno sporočanje podatkov s področja izobraževanja uporablja povezovalna tabela, katere podlaga je mednarodno validiran "ISCED mapping 2013". Po tej shemi se v najvišjo, osmo raven (Doctoral or equivalent level) klasifikacije ISCED 2011 razvrščajo trije izidi javnoveljnega izobraževanja v Sloveniji: doktorat znanosti (prejšnji), doktorat znanosti (tretje bolonjske stopnje) in magisterij znanosti (prejšnji). Magisterij znanosti (prejšnji) je po značilnostih predbolonjskega magistrskega izobraževanja namreč bolj podoben doktoratu znanosti kot magistrski izobrazbi druge bolonjske stopnje, zato se razvršča v osmo raven izobraževanja klasifikacije ISCED.

Med prebivalci Slovenije 13.182 doktorjev znanosti

V Sloveniji je na začetku leta 2019 živelo 13.182 prebivalcev z doktoratom znanosti, kar je 0,6 % celotnega prebivalstva.



Kako odkriti napake?

- **Križna preverjanja z referenčnimi podatki** iz zanesljivih virov
- **Iskanje manjkajočih vrednosti**, kjer bi morale biti prisotne
- **Statistična analiza** za identifikacijo izstopajočih vrednosti
- **Grafični prikazi podatkov**, ki omogočajo prepoznavo nenavadnih vrednosti
- **Logične preveritve** za ugotavljanje, ali so podatki med seboj skladni
- **Preverjanje uporabe predpisanih pravil** za določanje veljavnih vrednosti
- **Identifikacija dvojnikov** za preverjanje podvojenih zapisov
- **Analiza revizijskih sledi** za sledenje spremembam, kdo in kdaj je spreminjal podatke
- **Avtomatski testi podatkovne kakovosti** s programskimi orodji za preverjanje podatkov



Načrti ravnanja z raziskovalnimi podatki (NRRP)

NRRP je formalen dokument, ki opisuje ravnanje s podatki med projektom in po zaključku projekta. Običajno vključuje naslednje elemente:

1. Opis podatkov, ki jih je treba zbrati oziroma ustvariti
2. Dokumentacija in metapodatki
3. Etika in skladnost s pravnimi predpisi
4. Shranjevanje in varnostno kopiranje tekom projekta
5. Izbor in ohranjanje podatkov
6. Deljenje podatkov z drugimi
7. Odgovornost in sredstva za ravnanje s podatki



1. Opisovanje podatkov

- Vir oz. izvor podatkov
- Tip, format in obseg podatkov
- Standardi in metodologije
- Struktura in imena datotek
- Upravljanje različic
- Procesi zagotavljanja kakovosti



Izvor podatkov in provenienca

- **Izvor** se nanaša na informacije o okoliščinah nastanka podatkov (npr. avtorji, čas nastanka, raziskovalna oprema, ipd.), ki so pomembne za zagotavljanje interoperabilnosti in ponovne uporabljivosti podatkov.
- **Provenienca** gre korak dlje in se osredotoča na celoten process nastanka, vključno s spremembami, ki so se zgodile skozi čas, ter morebitnih obdelavah, ki so bile opravljene na podatkih, npr. čiščenje, transformacije, uteževanje, ipd.



Priporočeni formati za različne vrste podatkov

Vrsta podatkov	Priporočeni formati	Sprejemljivi formati
Tabelarni podatki	csv, .tab, .por, .xml	.txt, xls, .dbf, .ods, .sav, .dta, .mdb
Geoprostorski podatki	.shp, .shx, .dbf, .prj, .sbx, .sbn, .tif, .tfw, .dwg, .gml	.mdb, .mif, .kml, .ai, dxf, .svg
Besedilni podatki	.rtf, .txt, .xml	.html, .doc
Slikovni podatki	.tif	.jpg, .gif, .tif, .tiff, .raw, .psd, .bmp, .png, .pdf
Zvočni podatki	.flac	.mp3, .aif, .wav
Video podatki	.mp4, .ogv, .ogg, .mj2	.avchd
Dokumentacija in skripte	.rtf, .pdf, .xhtml, .htm, .odt	.txt, .doc, .xls, .xml

Vir: [UK Data Service](#)



Tabelarni podatki – relacijski podatkovni model

Relacijski podatkovni model je najbolj razširjen in priljubljen način organiziranja podatkov, vendar obstajajo tudi drugi (hierarhični, omrežni, objektni, dokumentni...).

Enote

Atributi (lastnosti)

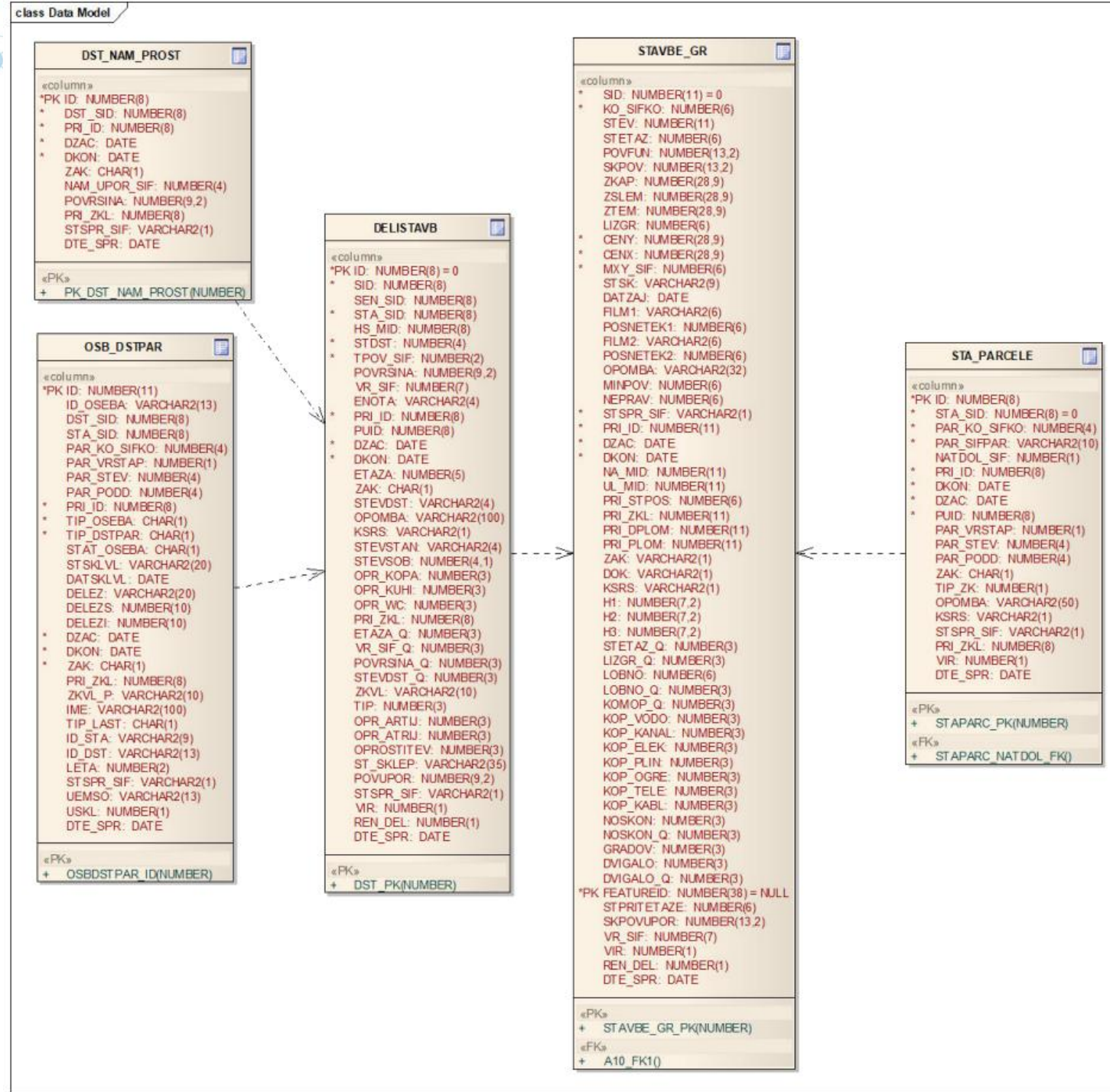
Ime države	Populacija	Površina v kvadratnih metrih	Leto vstopa v EU	...
Avstrija	8,857,960	301,318	1995	...
Italija	60,404,355	83,858	1957	...
Slovenija	2,070,050	20,273	2004	...
...	...	...	...	...

Vrednosti



Relacije med tabelami

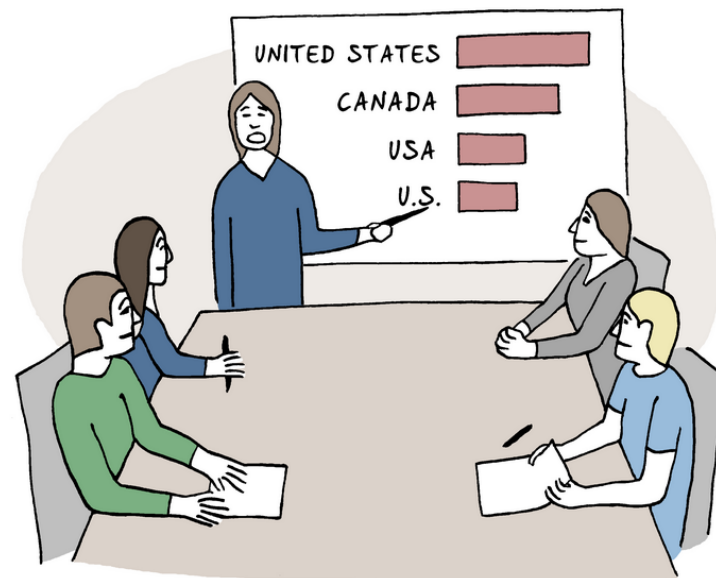
Primer zemljiškega katastra
(Vir: [Geodetski inštitut Slovenije](#))





Geslovniki, ontologije in tezavri

- Priporoča se uporaba splošno priznanih in uporabljenih geslovnikov, ontologij ali tezavrov
- **Geslovniki** so zbirke besed in njihovih definicij, ki se uporabljajo za razumevanje njihovega pomena.
- **Ontologije** so formalne predstavitve znanja v neki domeni, ki opisujejo entitete, njihove lastnosti in odnose med njimi.
- **Tezavri** so sistematične zbirke besed, ki so organizirani glede na pomene in odnose, kot so sinonimi, antonimi in hierarhične povezave (npr. [ELLST](#))



AS YOU CAN SEE, OUR TOP MARKETS ARE
UNITED STATES, CANADA, USA AND THE U.S.



Poimenovanje datotek

- Ustvarite smiselna, a kratka imena.
- Razvrščanje datotek glede na ime
- Uporaba datumov v formatu LLLL-MM-DD
- Uporaba pomišljajev (-) in podčrtajev (_)
- Izogibanje zelo dolgim imenom
- 3-črkovne končnice za specifične oznake formata datotek
- Upravljanje različic

Vir: [UK Data Service](#)



CREATE TABLE.
NAME... DOESN'T MATTER.
DONE.

ONE YEAR LATER...



WHAT THE #@%!
IS THIS TABLE?!



Procesi zagotavljanja kakovosti v različnih fazah raziskave

- Tekom **zbiranja podatkov** (kalibracija, število meritev, standardizacija metod, uporaba sistemov za računalniško podprto zbiranje podatkov...)
- Pri **vnašanju podatkov** in [digitizaciji](#) (pravila preverjanja veljavnosti, vmesniki za vnos podatkov, uporaba nadzorovanih besednjakov in kodnih seznamov za zmanjšanje ročnega vnosa podatkov, podrobno označevanje imen spremenljivk, oblikovanje spremna dokumentacija o podatkih...)
- **Preverjanje podatkov** (kodiranje in vrednosti izven razpona, popolnost podatkov, dodajanje oznak spremenljivk in vrednosti, preverjanje naključnih vzorcev, dvojni vnos podatkov, statistične analize, strokovni pregled...)
- Zagotavljanje **avtentičnosti** (hranjenje ene glavne datoteke, dodelitev odgovornosti eni osebi, ureditev dostopa **avtentičnosti** o različic podatkovnih datotek, ohranitev starih datotek, redno arhiviranje, razviti postopki za uničenje datotek)



2. Dokumentacija in metapodatki

- **Dokumentacija raziskave** vključuje podrobnosti o uporabljeni metodologiji, analitičnih in postopkovnih informacijah, definicijah spremenljivk, slovarjih, enotah merjenja, predpostavkah ter formatu in vrsti datotek podatkov.
 - **Metapodatki** lahko nastanejo avtomatsko ali pa se jih pripravi z ročnim beleženjem informacij; za pripravo metapodatkov se uporabljajo tudi obstoječi viri.
 - **Metapodatkovna shema** je niz pravil in smernic, ki določajo strukturo metapodatkov, njihove elemente in povezave med njimi.
 - **Metapodatkovni standard** je metapodatkovna shema, ki je navadno tudi mednarodno preverjena in potrjena s strani pristojne organizacije.
 - Splošne sheme oz. standardi (npr. Dublin Core)
 - Področno specifične sheme oz. standardi (npr. DDI)
- ➔ Imenik metapodatkovnih standardov [RDA Metadata Standards Directory](#)



Splošne sheme in standardi

- CERIF (Common European Research Information Format)
- Data Package
- DataCite Metadata Schema
- Dublin Core
- MODS (Metadata Object Description Schema) _
- OAI-ORE (Open Archives Initiative Object Reuse and Exchange)
- Repository-Developed Metadata Schemas
- ...



Dublin Core

- Splošni standard metapodatkov, ki se lahko uporablja in prilagaja za različne vrste vsebin, vključno z dokumenti, slikami, spletnimi stranmi, idr.
- Ključni elementi: **Naslov, Avtor, Tema, Opis, Založnik, Sodelujoči, Datum, Vrsta, Tip, Format, Identifikator, Vir, Jezik, Razmerja, Pokritost, Pravice**



DataCite

- Standard posebej zasnovan za citiranje raziskovalnih podatkov
- Ponuja način opisa podatkovnih nizov omogoča enostavno odkrivanje, dostopanje in citiranje podatkov.
- Elementi: **Identifikator (DOI), Naslov, Avtor, Založnik, Datum izdaje, Področje, Sodelavci, Datum, Jezik, Tip vira, Povezani identifikatorji, Velikost, Format, Različica, Pravice, Opis, Lokacija, Financiranje**



Področno specifični metapodatkovni standardi

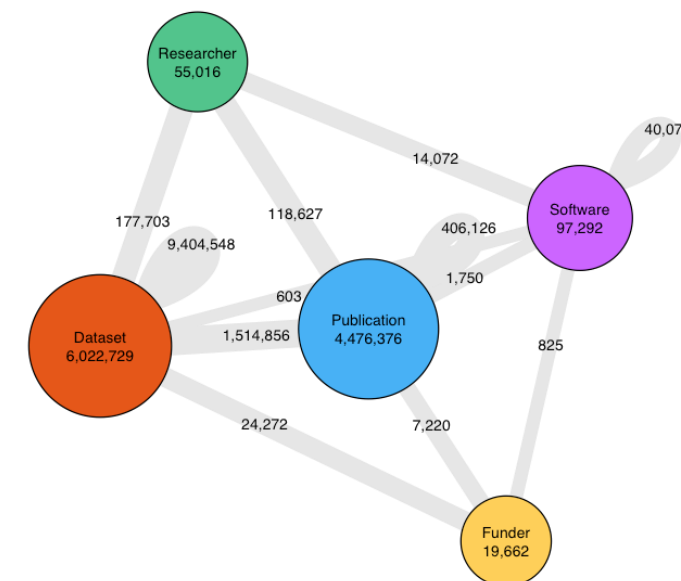
- Humanistika: [EAD](#), [TEI](#), [DDI](#), [MARC](#), [MIDAS](#), [OAI-ORE](#), [RO-Crate](#)
- Tehnika: [CIF](#), [CSMD](#), [ISA-Tab](#), [MIBBI](#), [NeXus](#), [openPMD](#)
- Vede o življenju: [ABCD](#), [Darwin Core](#), [EML](#), [Genome Metadata](#), [ISA-Tab](#), [MIBBI](#), [NeXus](#), [Observ-OM](#), [ODAM](#), [OME-XML](#), [openPMD](#), [PDBx/mmCIF](#), [Protocol Data Element Definitions](#), [Repository-Developed Metadata Schemas](#), [RO-Crate](#), [UKEOF](#)
- Fizikalne vede in matematika: [AgMES](#), [AVM](#), [CF](#), [CIF](#), [CIM](#), [CSMD](#), [DIF](#), [FGDC/CSDGM](#), [FITS](#), [International Virtual Observatory Alliance Technical Specifications](#), [ISO 19115](#), [NeXus](#), [Observations and Measurements](#), [openPMD](#), [PDBx/mmCIF](#), [Repository-Developed Metadata Schemas](#), [SDAC](#), [SPASE Data Model](#), [UKEOF](#)
- Družboslovje: [DDI](#), [MIDAS](#), [OAI-ORE](#), [QuDEx](#), [SDMX](#)



Trajni identifikatorji (PID)

- Trajni identifikatorji (angl. Persistent Identifiers oz. PID) predstavljajo enolične in trajne povezave do različnih objektov. Primeri PID:
 - Za digitalne objekte: [DOI](#), [HNDL](#), [ARK](#), [URN](#), ...
 - Za nedigitalne objekte: [ORCID](#), [ROR](#), ...
- Odnose in povezave med različnimi entitetami metapodatkov (avtorji, podatkovni nizi, ...) lahko prikažemo s PID grafi

PID Graph
Number of nodes and connections (30 June 2019)





3. Etika in zakonska skladnost

- Raziskovalne načrte, ki vključujejo zbiranje občutljivih podatkov, je treba predložiti v presojo etični komisiji
- Pridobivanje informiranega soglasja za zbiranje, obdelavo in dolgoročno hrambo podatkov
- Neposredne in posredne identifikatorje je treba odstrani ali poskrbeti za združevanje ali psevdoanonimizacijo podatkov
- V primerih, ko bi anonimizacija ovirala ponovno uporabnost podatkov, se uporabi omejitev dostopa do podatkov
- Poskrbeti je treba za skladnost s Splošno uredbo o varstvu podatkov
- Obravnava zaupnih podatkov in poslovnih skrivnosti



Soglasje za deljenje podatkov

- Zagotoviti je treba, da obrazci za soglasje, licence in pogodbe ne omejujejo možnosti deljenja podatkov.
- Če nameravamo podatke deliti z drugimi, sodelujočim v vabilu ne smemo zagotavljati, da bodo podatki uporabljeni samo za namen tekočega projekta. Seznaniti jih moramo tudi, da se lahko podatki uporabijo v druge namene.



4. Shranjevanje in varnostno kopiranje

- Avtorske pravice in pravice intelektualne lastnine
- Licence za ponovno uporabo
- Omejitev ponovne uporabe podatkov tretjih oseb
- Omejitve deljenja podatkov (obdobje embargo)
- Varnostni ukrepi in standardi za zaupne podatke



5. Izbor in ohranjanje podatkov

- Katere podatke je treba obdržati/uničiti za pogodbene, pravne ali regulativne namene
- Predvidene raziskovalne uporabe podatkov (validacija raziskovalnih rezultatov, načrtovanje novih raziskav, izobraževanje)
- Dolžina hrambe
- Izbira repozitorija
- Ocena stroškov hrambe



6. Deljenje podatkov

- Kje, kako, s kom?
- Pogoji za deljenje podatkov (pogodbe o izmenjavi podatkov)
- Kdaj bodo podatki postali dostopni
- Trajni identifikatorji
- Oris pričakovanih omejitev zaradi zaupnosti ali pravic intelektualne lastnine
- Sporazumi o nerazkritju z zaščito zaupnih podatkov



7. Odgovornost in sredstva za ravnanje s podatki

- Vloge in odgovornosti za zajem podatkov, izdelavo metapodatkov, zagotavljanje kakovosti podatkov, shranjevanje in varnostno kopiranje, arhiviranje podatkov in skupno rabo podatkov
- Ugotavljanje, ali potrebuje pomoč (podatkovnih) strokovnjakov
- Ugotavljanje, ali je potrebna dodatna strojna ali programska oprema



Organizacije in viri

- [Research Data Alliance](#)
- [CODATA](#)
- [Digital Curation Centre](#)
- [UK Data Service](#)
- [GO FAIR project](#)
- [FOSTER project](#)
- [Open Science MOOC](#)
- ...

Hvala za pozornost!



Univerza v Ljubljani

