

Ponovna uporaba raziskovalnih podatkov



UNIVERZA
V LJUBLJANI

dr. Brane LESKOŠEK, Medicinska fakulteta Univerze v Ljubljani, Inštitut za biostatistiko in medicinsko informatiko, Središče ELIXIR Slovenija

dr. Marko VIDAK, Medicinska fakulteta Univerze v Ljubljani, Inštitut za biostatistiko in medicinsko informatiko, Središče ELIXIR Slovenija

dr. Bojan KVERH, Medicinska fakulteta Univerze v Ljubljani, Inštitut za biostatistiko in medicinsko informatiko, Središče ELIXIR Slovenija

dr. Polonca FERK, Medicinska fakulteta Univerze v Ljubljani, Inštitut za biostatistiko in medicinsko informatiko, Središče ELIXIR Slovenija

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



Vsebina

- Področja ponovne uporabe podatkov
- Ponovna uporaba kot del življenjskega kroga podatkov
- Ponovna uporaba kot del podatkov FAIR
- Definicija, vloga v znanosti, prednosti in primer uporabe
- Področja za zagotavljanje ponovne uporabe
- Izzivi ponovne uporabe

Področja ponovne uporabe podatkov

- **Naravoslovje (Naravoslovne vede)**
- **Družboslovje (Družboslovne vede)**
- **Formalne znanosti** (npr. matematika, informatika in računalništvo)

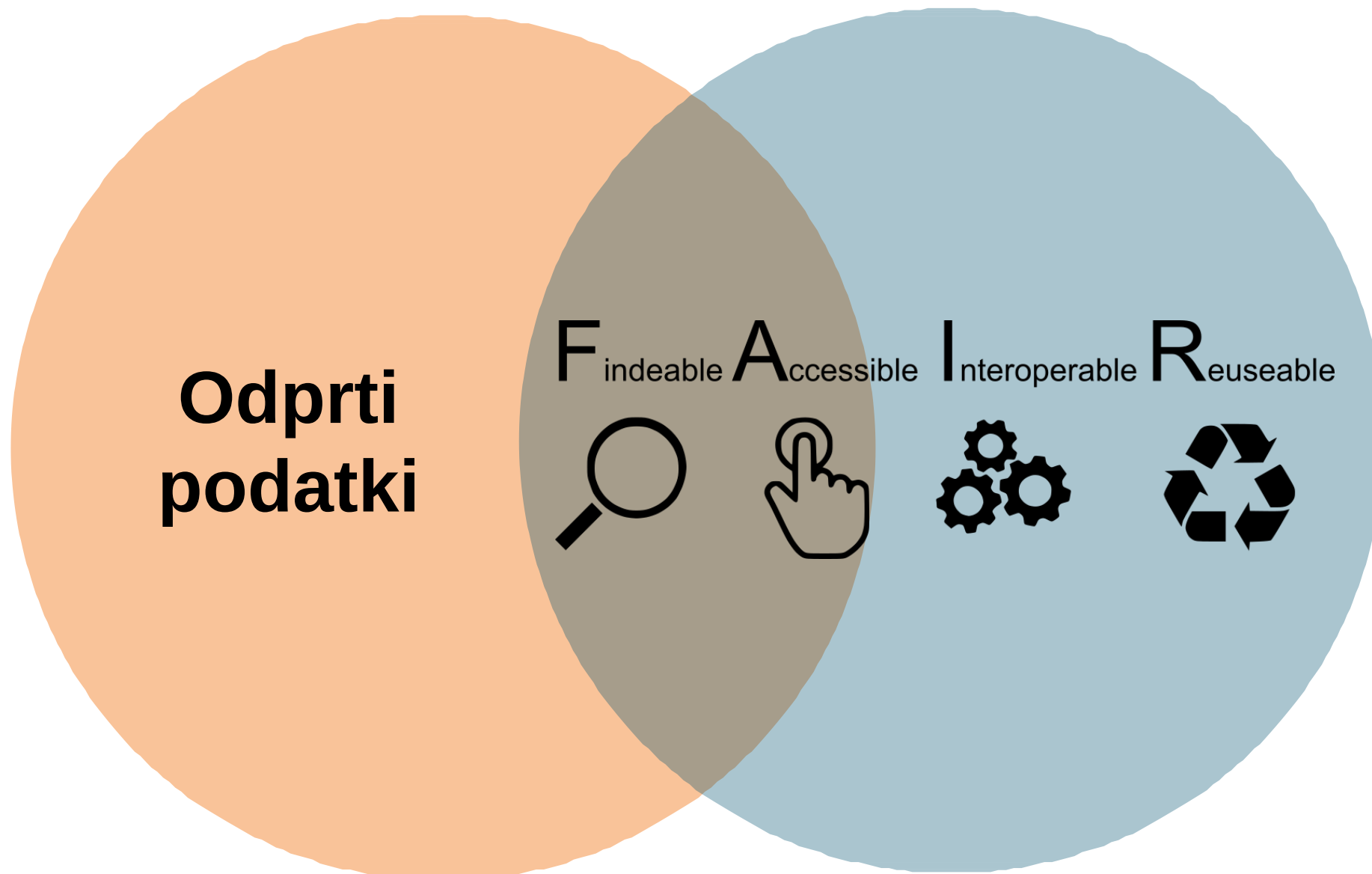
Življenjski krog podatkov

- Načrtovanje
- Zbiranje
- Obdelava
- Analiza
- Shranjevanje / Arhiviranje
- Deljenje
- **Ponovna uporaba**



Odprti podatki in podatki FAIR

- Odprti podatki in odprta znanost (*Open data, open science*)
- Osnovno odprtosti raziskovalnih podatkov: »podatki naj bodo odprti, kolikor je mogoče, zaprti, kolikor je nujno«.
- FAIR (*Findable, Accessible, Interoperable, Reusable*) pomeni najdljivi, dostopni, interoperabilni in ponovno uporabni.
- Ponovna uporabnost je poleg odprtosti podatkov ena od ključnih komponent načela FAIR za ravnanje s podatki.



“Odprti dostop, kolikor je mogoče, ter zaprti dostop, kolikor je nujno.”

“As open as possible, as closed as necessary.”

Definicija ponovne uporabe podatkov

- Ponovna uporaba podatkov je uporaba obstoječega nabora podatkov za namene, ki so drugačni od tistih, za katere so bili ti podatki prvotno zbrani.
- Ponovna uporaba omogoča raziskovalcem, da iz podatkov potegnejo nova spoznanja in tako povečuje (dodano) vrednost podatkov.

Vloga ponovne uporabe podatkov v znanosti

- Ponovna uporaba podatkov omogoča raziskovalcem, da nadgradijo obstoječa spoznanja, in tako pospešuje razvoj znanosti.
- Ponovna uporaba omogoča racionalizacijo stroškov zaradi manj podvajanja raziskav.
- Ponovna uporaba spodbuja interdisciplinarno raziskovanje in inovativne znanstvene pristope.

Prednosti ponovne uporabe podatkov

- Stroškovna učinkovitost
- Spodbujanje raziskav in razvoja
- Transparentnost in ponovljivost
- Spodbujanje dostopnosti in delitve podatkov (FAIR-ifikacija)

Stroškovna učinkovitost

- Ponovna uporaba omogoča raziskovalcem, da se osredotočijo na analizo podatkov namesto na njihovo zbiranje.
- Ponovna uporaba podatkov omogoča izvedbo meta-analiz. Tako lahko združimo rezultate več raziskav brez ponavljanja eksperimentalnega dela.

Spodbujanje raziskav in razvoja

- Ponovna uporaba podatkov olajša pot do novih idej in inovacij, saj omogoča kombinacijo spoznanj iz različnih virov.
- S pomočjo podatkov iz preteklosti lahko identificiramo trende, napovemo dogodke v prihodnosti ter razvijemo inovativne rešitve za kompleksne probleme.

Transparentnost in ponovljivost

- Ponovna uporaba podatkov poveča ponovljivost raziskav in je uporabna za potrjevanje pravilnosti ter zanesljivosti znanstvenih spoznanj.
- Ponovna uporaba podatkov spodbuja sprejemanje odločitev, ki temeljijo na dokazih. Tako se pospešuje družbeni razvoj na splošno.

Spodbujanje dostopnosti in delitve podatkov

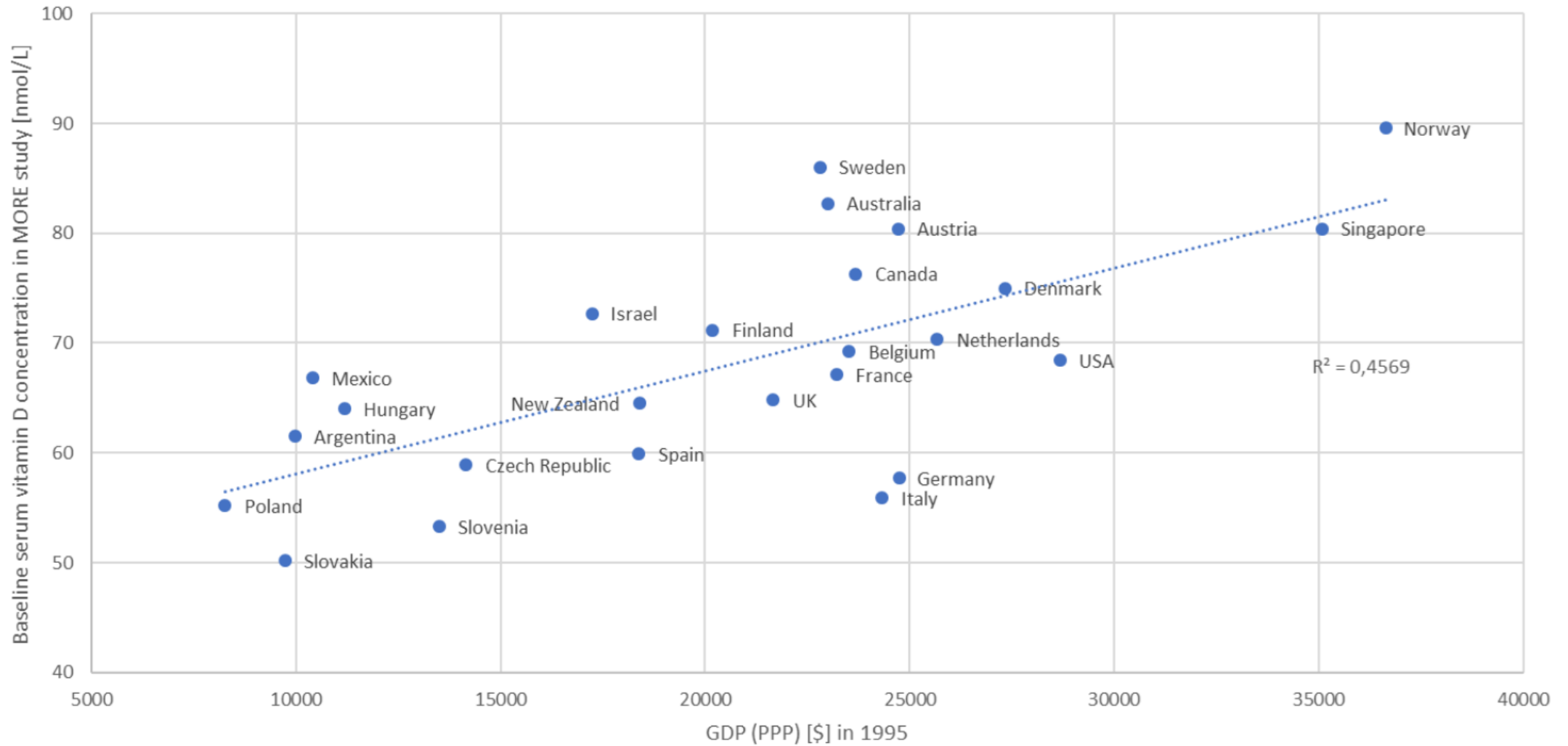
- Ponovna uporaba se navezuje tudi na dostopnost in interoperabilnost podatkov iz načel FAIR - omogoča večjo dostopnost podatkov za širšo uporabo.
- Ponovna uporaba spodbuja kulturo delitve podatkov in s tem večje sodelovanje med raziskovalnimi skupinami.
- Delitev podatkov omogoča inovacije, tudi takšne z velikim vplivom na družbo.

Primeri ponovne uporabe podatkov

- Primer iz mednarodne klinične raziskave MORE, v kateri so preizkušali zdravilo raloksifen za osteoporozo*.
- Najdemo korelacijo med serumsko koncentracijo vitamina D pri pacientkah z osteoporozo in gospodarsko razvitostjo države, merjeno z bruto družbenim proizvodom po kupni moči.

* Lips P et al. A global study of vitamin D status and parathyroid function in postmenopausal women with osteoporosis: baseline data from the multiple outcomes of raloxifene evaluation clinical trial. J Clin Endocrinol Metab. 2001 Mar;86(3):1212-21. J Clin Endocrinol Metab 2001 Jul;86(7):3008.

GDP (PPP) vs Baseline serum vitamin D concentration (MORE study, 1995)



Področja za zagotavljanje ponovne uporabe podatkov

- **Dostopnost in najdljivost podatkov in postopkov**
 - Za ponovno uporabo podatkov je nujno, da so podatki in postopki dostopni in da jih lahko najdemo / odkrijemo
- **Metapodatki in dokumentacija (provenienca)**
 - Podrobni opisni podatki so ključni za razumevanje konteksta in uporabnosti podatkov

FAIRifikacija!

Področja za zagotavljanje ponovne uporabe podatkov

- **Standardni podatkovni formati** FAIRifikacija!
 - Uporaba standardiziranih podatkovnih formatov izboljšuje združljivost in interoperabilnost podatkov med različnimi uporabniki in informacijski sistemi
- **Podatkovni repozitoriji in ostale spletne podatkovne zbirke/storitve**
 - Platformi kot sta **GitHub** ali **Zenodo** podpirata izmenjavo in ponovno uporabo podatkov
 - **Beacon** na področju biomedicine omogoča lažje odkrivanje genomskih podatkov

Upravljanje z dostopom do podatkov

(Data Governance and Access Control)

- **Varnostne politike in politike dostopa do podatkov**
urejajo načine dostopa do podatkov in tako omogočajo odgovorno (ponovno) uporabo podatkov.
 - Za dostop do genomskih podatkov preko repozitorijev ENA ali FEGA se uporablja način Odborov za dostop do podatkov (DACs = *Data Access Committees*)
 - Nadzori dostopa so ključni za vzdrževanje integritete in varnosti podatkov. Ne pozabimo osnovnega načela odprtosti raziskovalnih podatkov »odprti, kolikor je mogoče, zaprti, kolikor je nujno«.

Izzivi pri ponovni uporabi podatkov

- Analiza podatkov
- Kakovost in format podatkov
- Varnost podatkov in občutljivi podatki
- Licence
- Repozitoriji in arhivi podatkov
 - Prenos in iskanje podatkov
- Metapodatki
- Trajni identifikatorji
- Ontologije

Analiza podatkov

- Kako narediti analizo podatkov FAIR tj. da je ponovljiva?
 - damo na voljo programsko kodo
 - damo na voljo okolje za izvajanje programov
 - damo na voljo delotoke
- Kje dobiti oz. kam dati podatke? Repozitorij GitHub, GitLab
 - Okolje za programsko opremo: Conda ali Bioconda
 - Vsebniki (*containers*): Docker ali Singularity
 - Delotoki: Galaxy, Nextflow, CWL, Snakemake, ...

Kakovost in format podatkov

- Težave lahko nastopijo zaradi manjkajočih vrednosti v podatkih, nezdružljivih formatov ali pomanjkanja standardizacije.
- Za zagotavljanje uporabnosti podatkov je velikokrat potrebno podatke prečistiti in pretvoriti v drug format.

Format podatkov

- Format podatkov je ključna informacija, ki jo moramo navesti v metapodatkih
- Podatki naj bodo v formatih, ki se pogosto uporabljajo in so splošno razširjeni
- Dostopnost formata z odprtokodnimi orodji
 - Nelastniški formati (*Non-proprietary formats*)

Primer formata podatkov: binarni format

- Pri pretvorbi v binarni format lahko pride do izgube podatkov.
- Podatke lahko urejamo le s posebnimi programi, npr. C++ editor.

Microsoft | Learn | Documentation | Training | Certifications | Q&A | Code Samples | Assessments | Shows | Events

... / Work with resource files / Resource editors /

Binary Editor (C++)

Article • 11/11/2021 • 4 minutes to read • 9 contributors [Feedback](#)

Caution

Editing resources such as dialog boxes, images, or menus in the **Binary Editor** is dangerous. Incorrect editing could corrupt the resource, making it unreadable in its native editor.

The **Binary Editor** allows you to edit any resource at the binary level in either hexadecimal or ASCII format. You can also use the [Find command](#) to search for either ASCII strings or hexadecimal bytes. Use the **Binary Editor** only when you need to view or make minor changes to custom resources or resource types not supported by the Visual Studio environment. The **Binary Editor** is not available in Express editions.

Version: Visual Studio 2022

Filter by title

- Accelerator editor
- Binary editor**
- Dialog editor
- Image editor for icons
- Menu editor
- String editor
- Toolbar editor
- Version information editor

> Deploy native desktop applications

Download PDF

Scribble - Microsoft Visual C++ [design] - Scribble.rc (IDD_ABOUTBOX - Data)

File Edit View Project Build Debug Schema Format Tools Component Window Help

VS Home Page | Scribble.rc (IDD_ABOUTBOX - Data)

00000000	C0 00 C8 80 00 00 00 00	04 00 00 00 00 00 EB 00	A.b.o.u.t.
00000010	37 00 00 00 00 00 41 00	62 00 6F 00 75 00 74 00	7....S.c.r.i.b.b.l.
00000020	20 00 53 00 63 00 72 00	69 00 62 00 62 00 6C 00	M.S..S.h.
00000030	65 00 00 00 08 00 4D 00	53 00 20 00 53 00 68 00	e.l.l..D.l.g...
00000040	65 00 6C 00 6C 00 20 00	44 00 6C 00 67 00 00 00	...P.....P
00000050	03 00 00 50 00 00 00 00	0B 00 11 00 14 00 14 00	...(.w.....
00000060	FF FF FF FF 82 00 FF FF	80 00 00 00 80 00 02 50	...S.c.r.i.b.b.l.
00000070	00 00 00 00 28 00 0A 00	77 00 08 00 FF FF FF FF	e...V.e.r.s.i.o.
00000080	82 00 53 00 63 00 72 00	69 00 62 00 62 00 6C 00	n...1...0.....
00000090	65 00 20 00 56 00 65 00	72 00 73 00 69 00 6F 00	...P...(.w.....
000000a0	6E 00 20 00 31 00 2E 00	30 00 00 00 00 00 00 00	C.o.p.y.r.
000000b0	00 00 02 50 00 00 00 00	28 00 19 00 77 00 08 00	i.g.h.t...(.C.)..
000000c0	FF FF FF FF 82 00 43 00	6F 00 70 00 79 00 72 00	..2.0.0.1.....
000000d0	69 00 67 00 68 00 74 00	20 00 28 00 43 00 29 00	...P.....2...
000000e0	20 00 32 00 30 00 30 00	31 00 00 00 00 00 00 00	O.K.....
000000f0	01 00 03 50 00 00 00 00	B2 00 07 00 32 00 0E 00	
00000100	01 00 FF FF 80 00 4F 00	4B 00 00 00 00 00 00	

Address from start of file | Hexadecimal data | ASCII value

Varnost občutljivih podatkov

- Varnost podatkov je bistvenega pomena še posebej v primerih, ko gre za občutljive osebne podatke.
- Potrebno je spoštovati zakonodajo o varstvu podatkov in zasebnosti, npr. v EU velja Splošna uredba o varstvu podatkov (*General Data Protection Regulation* - GDPR), v drugih delih sveta imamo drugo zakonodajo kot npr. HIPPA (*Health Insurance Portability and Accountability Act*) v ZDA ali PIPEDA (*Personal Information Protection and Electronic Documents Act*) v Kanadi.

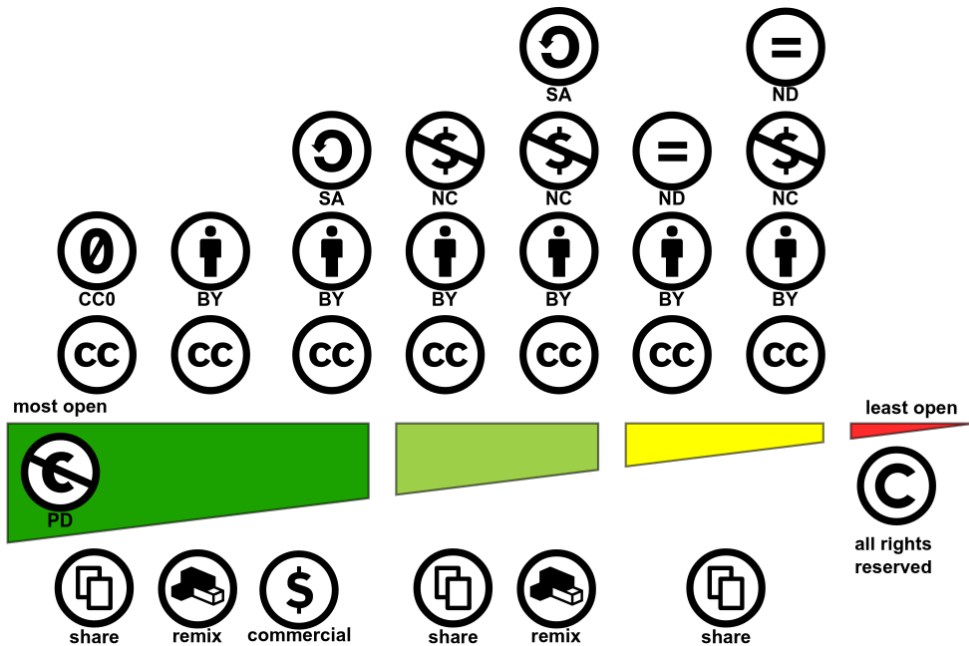
Licence ter pravni in etični pomisleki

- Lastništvo podatkov, pravice intelektualne lastnine in spoštovanje dogovorov o uporabi podatkov.
- Etični pomisleki glede ponovne uporabe podatkov, kar vključuje lastništvo nad podatki in privolitev za uporabo osebnih podatkov (obveščeni pristaneek = *informed consent*)
- Pomen vzpostavitve etičnih smernic za odgovorno (ponovno) uporabo podatkov

Licence za uporabo podatkov

EUDAT izbirnik licenc

<https://ufal.github.io/public-license-selector/>



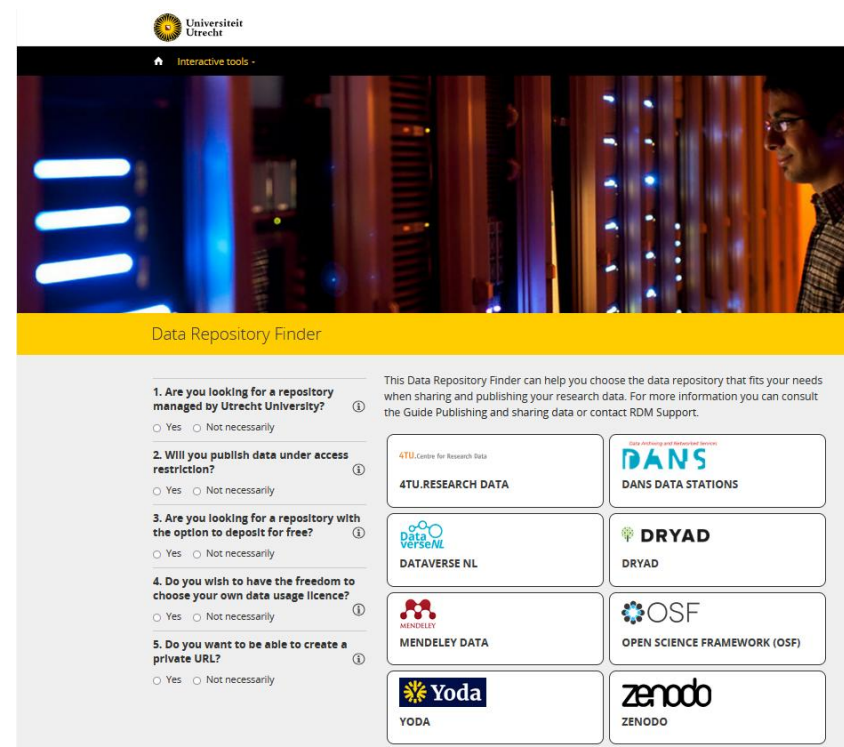
- Licence
 - CC
 - GNU
 - Mozilla
 - BSD
 - MIT
 - ...

Prenos podatkov

- Javni oblaki: GDrive, OneDrive, Amazon, Dropbox
- Interni oblaki: NextCloud
- Objektni arhivi podatkov
- **Repozitoriji**

Iskalniki repozitorijev

- Iskalnik Univerze v Utrechtu, podprt s strani EOSC
- <https://tools.uu.nl/repository-decision-tool/>



The screenshot shows the 'Data Repository Finder' tool from Utrecht University. The header includes the university logo and the text 'Interactive tools'. Below the header is a yellow banner with the title 'Data Repository Finder'. The main content area contains a series of five questions on the left and a grid of repository options on the right. The questions are:

1. Are you looking for a repository managed by Utrecht University?
☐ Yes ☐ Not necessarily
2. Will you publish data under access restriction?
☐ Yes ☐ Not necessarily
3. Are you looking for a repository with the option to deposit for free?
☐ Yes ☐ Not necessarily
4. Do you wish to have the freedom to choose your own data usage licence?
☐ Yes ☐ Not necessarily
5. Do you want to be able to create a private URL?
☐ Yes ☐ Not necessarily

On the right, there is a grid of repository options, each with a logo and name:

- 4TU Centre for Research Data (4TU.RESEARCH DATA)
- DANS (DANS DATA STATIONS)
- DataVerseNL (DATAVERSE NL)
- DRYAD (DRYAD)
- Mendeley (MENDELEY DATA)
- OSF (OPEN SCIENCE FRAMEWORK (OSF))
- Yoda (YODA)
- zenodo (ZENODO)

Below the questions, there is a small text box that reads: 'This Data Repository Finder can help you choose the data repository that fits your needs when sharing and publishing your research data. For more information you can consult the Guide Publishing and sharing data or contact RDM Support.'

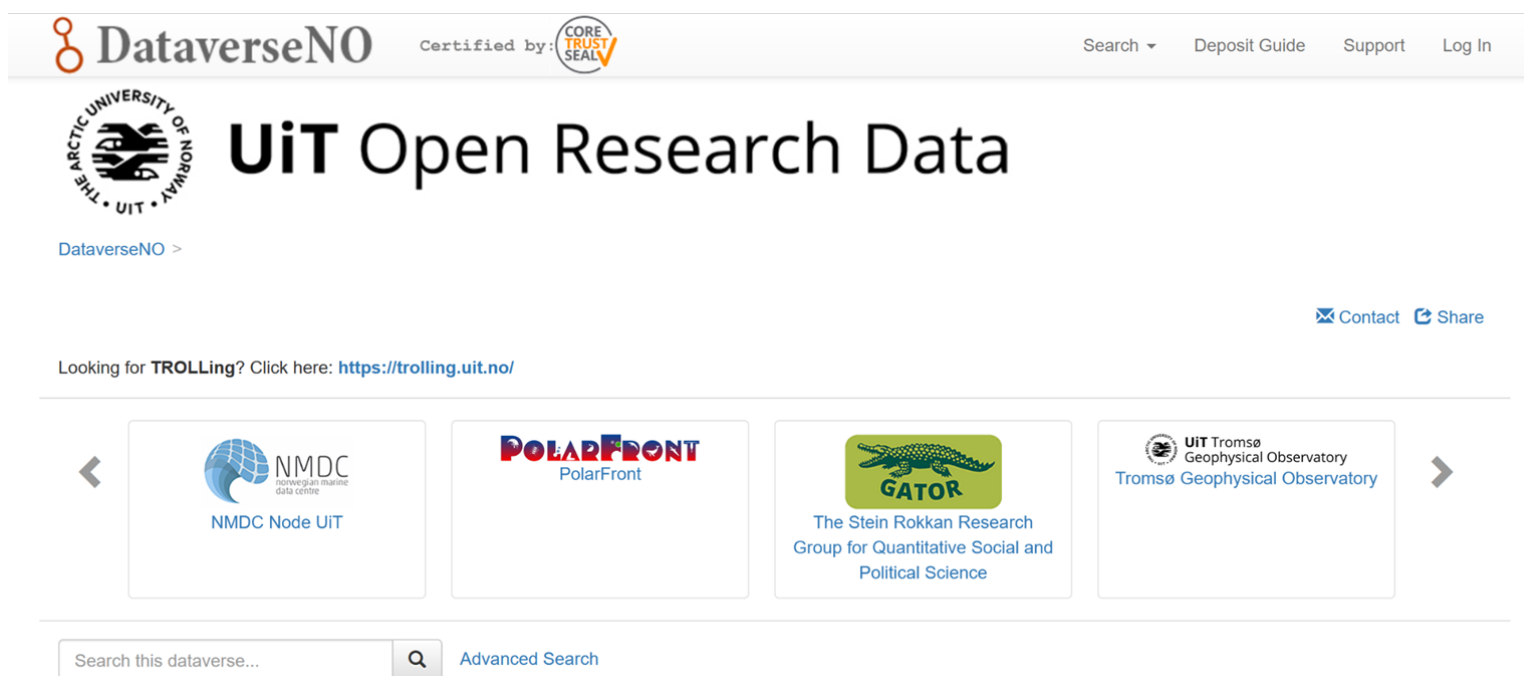
Nacionalni repozitoriji

- Podatki iz določene države, drugih omejitev ni
- Primer: Dataverse.no
 - Norveški nacionalni repozitorij



Nacionalni repozitoriji - Dataverse.no

- Dataverse.no (<https://dataverse.no/dataverse/uit>)
 - Vsebuje tudi podatke z družboslovnega področja



Institucionalni repozitoriji

- Podatki iz določene institucije
- Ni drugih omejitev
- Primer: RUL



Institucionalni repozitoriji - RUL

- Repozitorij Univerze v Ljubljani - RUL
 - Med drugim vsebuje vse doktorske, magistrske in diplomske naloge UL



Splošni repozitoriji

- Ni omejitev
- Primer: GitHub, Zenodo

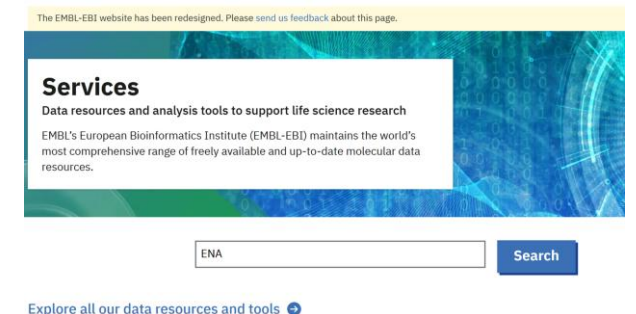


Specifični repozitoriji

- Primer: BioStudies (<https://www.ebi.ac.uk/biostudies/>)



- Primer: ENA (*European Nucleotide Archive*)
- <https://www.ebi.ac.uk/ena/browser/home>



Metapodatki

- Metapodatki = podatki o podatkih (lepše slovensko je tudi opisni podatki)
 - Metapodatki omogočajo lažje iskanje po podatkih
 - Podobno kot je npr. stvarno kazalo v knjigi

Standardi za metapodatke

- Pri opisu podatkov moramo upoštevati standarde za metapodatke
- Standarde lahko najdemo s specializiranimi iskalniki
 - Primer iskalnika: <https://FAIRsharing.org>

Trajni identifikatorji (PID = *persistent identifiers*)

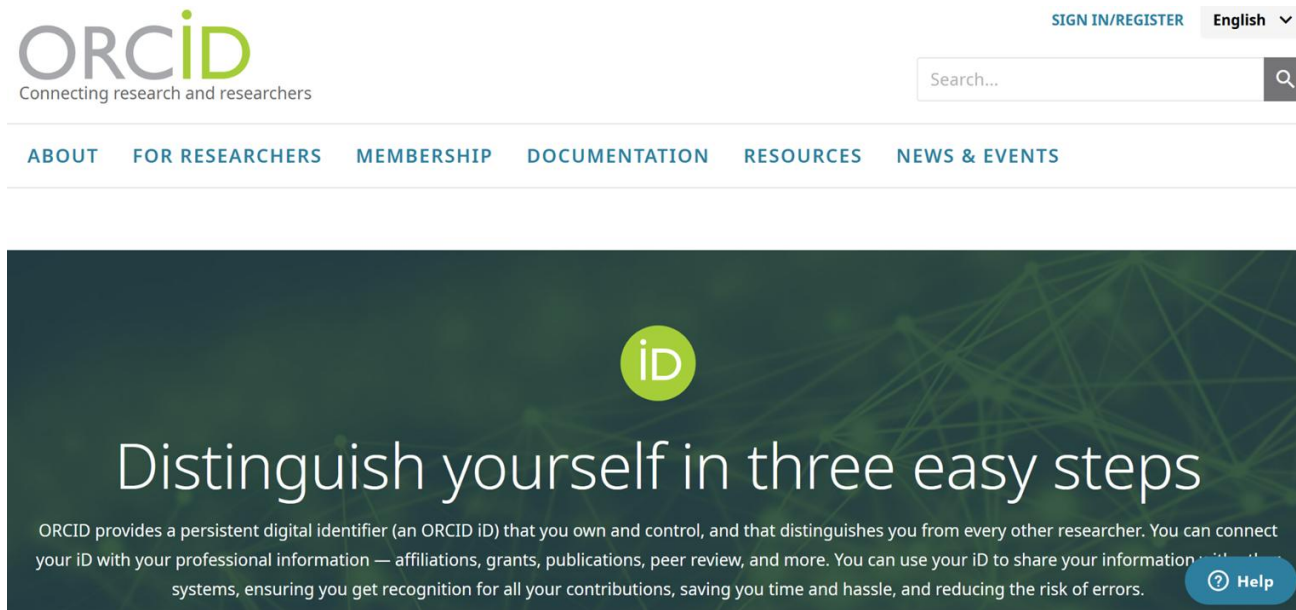
- Trajni identifikatorji so primer obveznih metapodatkov:
 - Identifikator je stalna šifra podatka (podobno kot npr. serijska številka pri avtomobilu)
 - Identifikator nam omogoča, da najdemo podatek tudi v spremenjenih okoliščinah (npr. selitev podatkov)

Primeri trajnih identifikatorjev

- Osebni identifikator za raziskovalce:
 - ORCID
- Identifikator za objave na internetu:
 - DOI (*Digital Object Identifier*)
- Identifikator za spletne naslove:
 - URL (*Uniform Resource Locator*)
- Nacionalni identifikatorji za raziskovalce in ustanove:
 - SICRIS

Trajni identifikatorji - pregled

- Open Researcher and Contributor Identity Document (ORCID)
<https://orcid.org/>





The DOI® System

ISO 26324



This is the web site of the International DOI Foundation (IDF), a not-for-profit [membership organization](#) that is the governance and management body for the [federation of Registration Agencies](#) providing Digital Object Identifier (DOI)

Resolve a DOI Name

Type or paste a DOI name, e.g., 10.1000/xyz123, into the text box below.
(Be sure to enter all of the characters before and after the slash. Do not include extra characters, or sentence punctuation marks.)

Clicking on a DOI link (try this one: <https://doi.org/10.1109/5.771073>) takes you to one or more current URLs or other services related to a single resource. If the URLs or services change over time, e.g., the resource moves, this same DOI will continue to resolve to the correct resources or services at their new locations.

Check the current status of the DOI system at doi.statuspage.io.



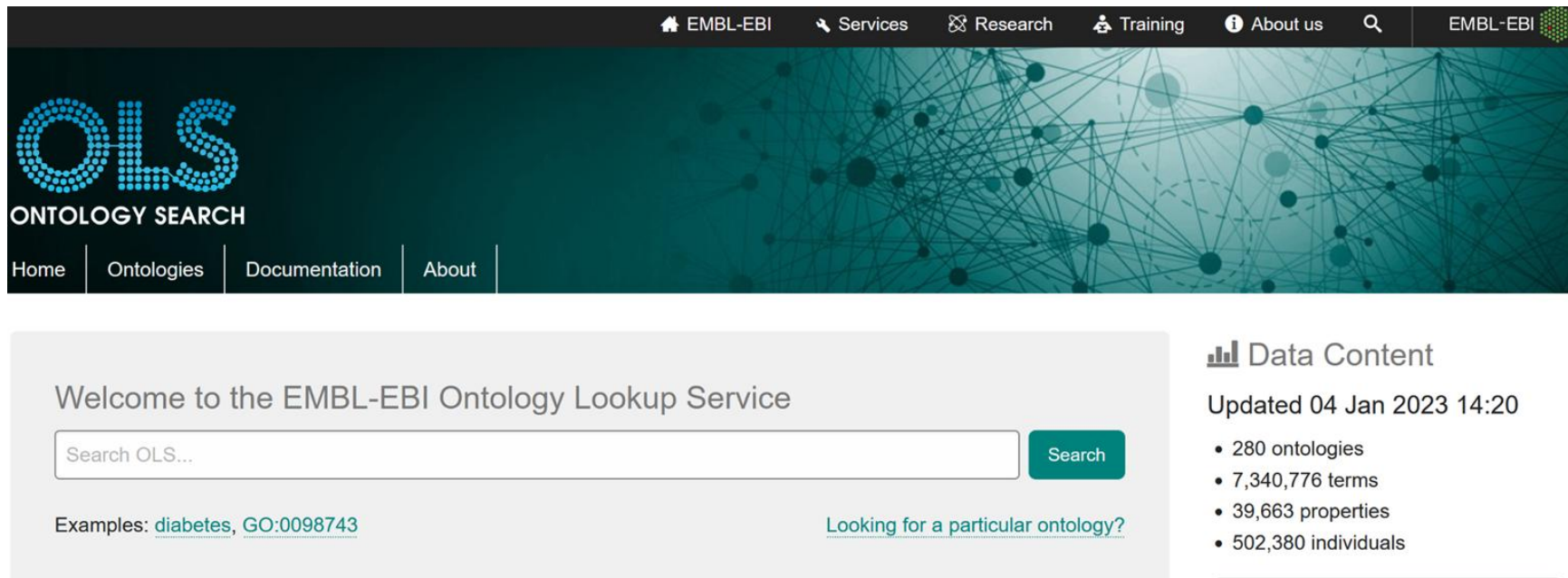
Enhance the value of your content.
Join the DOI Community

Ontologije in urejeni besednjaki

- Ontologije lahko poiščemo z iskalniki ontologij
 - Primer iskalnika za ontologije v biomedicini je OLS (*Ontology Lookup Service*)
- S pomočjo OLS lahko najdemo posamezne ontologije:
 - ENVO - ontologija za pojme, povezane z okoljem
 - GAZ - ontologija za geografske lokacije
- Primeri urejenega besednjaka iz biomedicine
 - MeSH (*Medical subject headings*)
<https://www.ncbi.nlm.nih.gov/mesh/>

Ontology Lookup Service (OLS)

- <https://www.ebi.ac.uk/ols/index>



The screenshot shows the EMBL-EBI Ontology Lookup Service (OLS) homepage. The top navigation bar includes links for EMBL-EBI, Services, Research, Training, and About us. The main header features the OLS logo and the text "ONTOLOGY SEARCH". Below the header is a navigation menu with links for Home, Ontologies, Documentation, and About. The main content area is divided into two sections. The left section, titled "Welcome to the EMBL-EBI Ontology Lookup Service", contains a search bar with the placeholder text "Search OLS..." and a green "Search" button. Below the search bar, it provides examples: "diabetes" and "GO:0098743", and a link "Looking for a particular ontology?". The right section, titled "Data Content", shows the update date "Updated 04 Jan 2023 14:20" and a list of statistics: 280 ontologies, 7,340,776 terms, 39,663 properties, and 502,380 individuals.

EMBL-EBI Services Research Training About us

OLS
ONTOLOGY SEARCH

Home Ontologies Documentation About

Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... Search

Examples: [diabetes](#), [GO:0098743](#) [Looking for a particular ontology?](#)

Data Content
Updated 04 Jan 2023 14:20

- 280 ontologies
- 7,340,776 terms
- 39,663 properties
- 502,380 individuals

The Environment Ontology - ENVO (okoljski podatki)

- <https://sites.google.com/site/environmentontology/>

OLS / The Environment Ontology **ENVO** / **ENVO:02500000** 

environmental system process 

 http://purl.obolibrary.org/obo/ENVO_02500000 

A process in which includes the components of an environmental system as participants.

Tree view

Term mappings

entity

occurent

process

environmental system process

Graph view

Reset tree

Show all siblings

Term information

comment

- This is a convenience class for organisation and should not be used for annotation.

Term relations

Equivalent to:

- process and has participant some (astronomical body part or astronomical object or astronomical body or environmental system or environmental material)

Zaključek

Ponovna uporaba podatkov je temelj odprte znanosti, saj omogoča večjo transparentnost, sodelovanje in inovacije. Sledenje načelom FAIR (najdljivi, dostopni, interoperabilni in ponovno uporabni) povečuje vrednost raziskav, pospešuje znanstveni napredek ter spodbuja sodelovanje in učinkovitost v znanstveni skupnosti.

Odprti podatki in
FAIRifikacija!

If you want to go fast,
go alone.
If you want to go far,
go together.



Vprašanja ?

brane.leskosek@mf.uni-lj.si

UL MF, IBMI, Središče ELIXIR-SI