



ZRC SAZU

O odprtih podatkih, jezikoslovnem raziskovanju in potrebah –

Znanje je edina dobrina, ki se množi, če ga delimo. (B. Stevanovič)

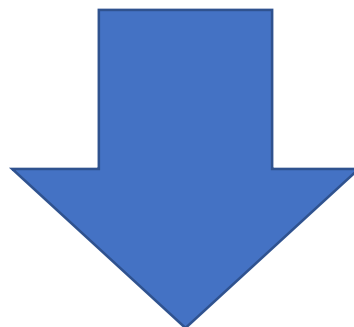
Doc. dr. Mateja Jemec Tomazin, ZRC SAZU, Podiplomska šola

Kdor išče, mora najti ...

(<https://www.aris-rs.si/sl/dostop/predstavitev.asp>)

Kako o svojih odkritjih povedati svetu?

- Galileo Galilei (1564–1642) – heliocentrični sistem
- Johannes Kepler (1571–1630) – zakon o gibanju planetov
- Isaac Newton (1643–1727) – zakoni težnosti, klasična optika



Objava odkritij (z anagrami in šiframi) v časopisu, da je zanje izvedela javnost.

Vir: M. Nielsen: ***Reinventing Discovery: The New Era of Networked Science***, Princeton University Press, 2011, 173.

Projekt Spoznaj in Slovar odprte znanosti

- **Ožja delovna skupina:** Mateja Jemec Tomazin, ZRC SAZU, Sonja Bezjak, Irena Vipavc Brvar, ADP
- **Sodelavke in sodelavci avtoric:** Ana Božič Ranfl, Peter Čerče, Mojca Kotar, Bernarda Korez, Dunja Legat, Branimir Leskošek, Milan Ojsteršek, Miro Pušnik, Mojca Ogrin, Alenka Kavčič Colić, Ana Inkret, Miha Peče, Anže Županič
- **Potrjevanje slovarja:** Terminološka sekcija RDA, delovna skupina 5 projekta SPOZNAJ
- **Slovenski terminološki portal:** https://terminoloski.slovenscina.eu/iskanje?q=*&d=48
- Termini so iz nastajajočega priročnika o odprti znanosti v projektu SPOZNAJ, nabor je pripravila Sonja Bezjak, urednica priročnika, ter ga dala v pregled in komentiranje terminološki skupini, ki jo vodi Ana Inkret, ADP.
- Slabosti: - nima ISBN, problem *trajnega identifikatorja* = **enolični zapis, ki trajno nedvoumno označuje digitalni objekt ter omogoča enoznačno prepoznavo (DOI, HANDLE ... EMŠO, DŠ ☺)**

Temeljna načela odprte znanosti (ARIS)

- **Takojšnja in splošna dostopnost znanstvenih del**, ki so rezultat javnega financiranja znanstvenoraziskovalnega dela (na primer **znanstvene objave** ali **raziskovalni podatki**) na svetovnem spletu brez naročniških omejitev;
 - **Plačevanje odprtih objav**
 - takojšnja shranitev končne, recenzirane različice članka (VoR) ali recenziranega rokopisa (AAM) v zaupanja vreden repozitorij – **STROJNA BERLJIVOST**
 - podatki, nujno potrebni za ponovitev raziskav
 - ponovna uporaba rezultatov raziskav v drugih raziskavah
 - **Uporabniške raziskave** (naglas; informatorji; snemanje; prednostni termin, poznavanje frazemov, pregovorov, prepoznavanje oblik, potrditev rabe)
 - **Ponovitev**: spremembe/potrditve naglasov, podatki o frekvenci, n-grami

Deljenje in skupna uporaba programske, strojne in druge opreme

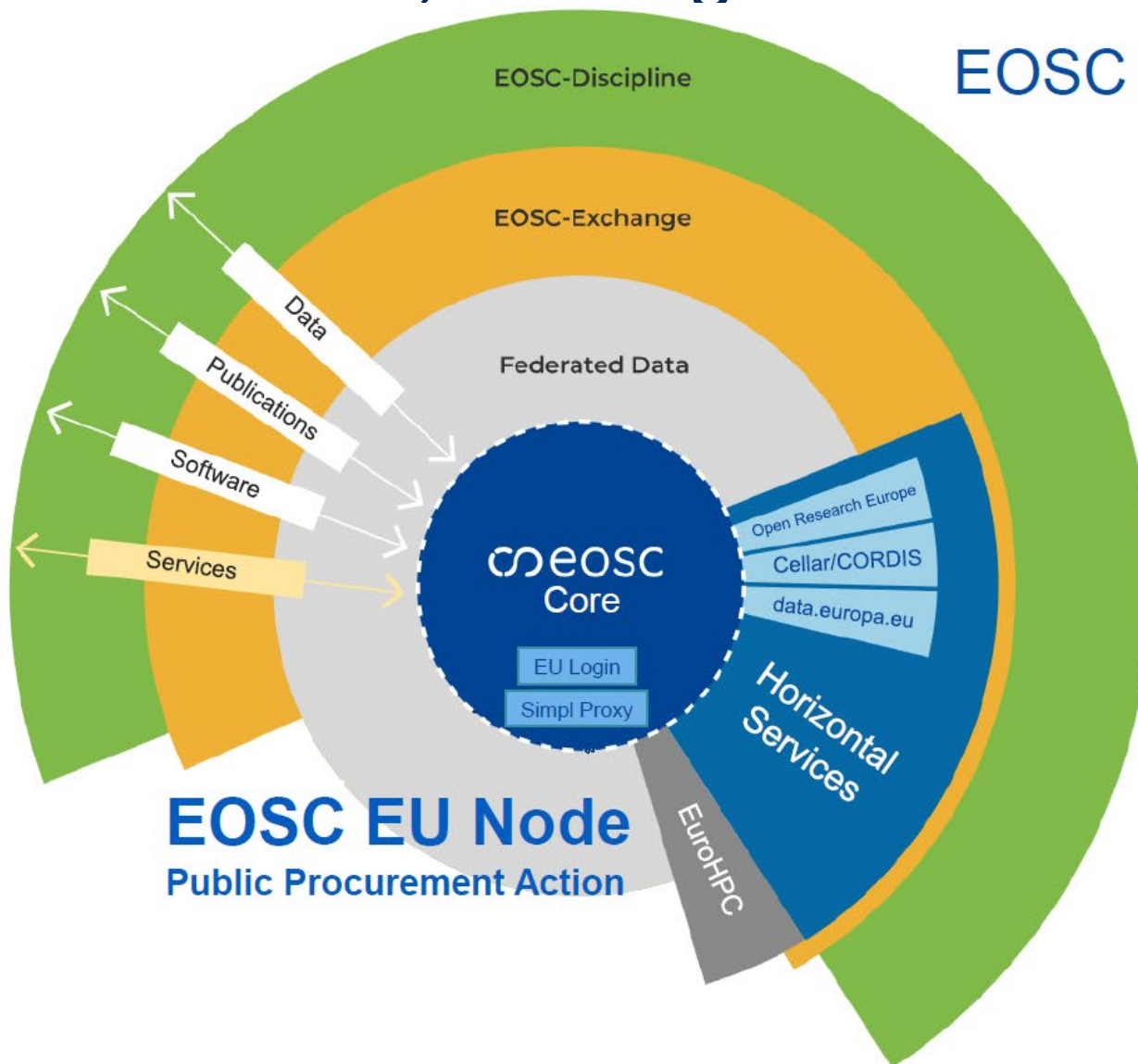
- Skupni nakup strojne opreme – projekt RSDO – FRI UL
- Sredstva Clarin.si – vzhodna – zahodna kohezijska regija – računalniki, strežniki na UM FERI
- Izum.si – Vega – zahtevki + čakalne vrste
- Programski jeziki R; Python
- Repozitoriji: [GitHub](#), GitLab; RUL, Openscience.si
- SLING – Slovensko nacionalno superračunalniško omrežje (<https://www.sling.si/>)



Delavnice, praktični primeri uporabe

Evropski oblak za odprto znanost (EOSC); Berlin 2024, P. Szegedi

EOSC EU Node Value Proposition



- **Facilitate** the creation of the “*Web of FAIR data and interoperable services*” (aka. EOSC Federation) under the Open Science Policy
- **Put** a “*seed in the ground*” by operationalizing the first recognised EOSC Node at the European level for the initial 3 years
- **Offer** “*core services*” for scientific research infrastructures to federate (single-sign-on, catalogues, knowledge graph, application workflow, monitoring, accounting, helpdesk) and common “*horizontal services*” for end-users to benefit from (compute, containers, data transfer, notebooks, file sharing, open research data)
- **Define** the *pathway and blueprint* (EOSC Interoperability Framework) for other potential EOSC Node operators to join the federation

Zaupanja vredni repozitoriji

- **Definicija:** repozitorij, ki je namenjen shranjevanju, ohranjanju in dostopu do digitalnih objektov za daljše časovno obdobje ter je praviloma mednarodno certificiran ali vključen v uradne seznane
- **Področni repozitorij:** repozitorij, ki je specializiran za shranjevanje in dostop do digitalnih objektov posebej opredeljenega znanstvenega področja:
 - ADP
 - CLARIN.SI
- **Institucionalni repozitorij:** repozitorij, ki ga upravlja posamezna ustanova pod svojimi pogoji in na katerem sta omogočena shranjevanje ter dostop do digitalnih objektov raziskovalcev in drugih, ki raziskujejo pod okriljem te ustanove
 - **RUL**
 - **DKUM**
 - **RUP**
 - **RUNG**
 - **DiRROS** - <https://dirros.openscience.si/info/index.php/slo/>

Splošni repozitoriji – namenjeni raziskovalnim podatkom

- Zenodo: <https://zenodo.org/> (projekt OpenAIRE)
- Figshare: <https://figshare.com> (Podpira odprto deljenje raziskovalnih rezultatov (podatkov, grafov, videov, itd.).
- Dryad: <https://datadryad.org> (Osredotočen na podatke, povezane z znanstvenimi publikacijami.
- Dataverse: <https://dataverse.org> (Odprt repozitorij, ki omogoča vzpostavljanje lastnih namestitev (npr. Harvard Dataverse).
- Dlib (?): <https://www.dlib.si/>



NAČELA **FAIR, CARE**

Načela FAIR

Načela, po katerih morajo biti digitalni objekti v odprti znanosti objavljeni, najdljivi, dostopni, interoperabilni in uporabljivi

- **F – najdljivost:** načelo ravnanja z digitalnimi objekti, po katerem morajo biti ti vključeni v repozitorije, kjer jih iskalniki enostavno zaznajo in vključijo med rezultate
- **A – dostopnost:** načelo ravnanja z digitalnimi objekti, po katerem mora biti razvidno, kako je mogoče do njih priti, in vključuje tudi morebitne postopke avtentikacije in avtorizacije
- **I – interoperabilnost:** načelo ravnanja z digitalnimi objekti, po katerem morajo ti znotraj sistema za svoj opis uporabljati skupni jezik, kar omogoča semantično in sintaktično razumevanje z drugimi sistemi, kot je npr. izmenjava
- **R – uporabljivost:** načelo ravnanja z digitalnimi objekti, po katerem morajo biti ti strukturirani in opremljeni s takšnim opisom, da jih je mogoče replicirati ali vključiti v nove ali razširjene raziskave

Načela CARE

Načela, po katerih morajo podatki o staroselskih, manjšinskih in drugih skupnostih podpirati njihovo kolektivno korist, skupnostim zagotavljati pravico nadzora nad podatki ter odgovornost in etičnost raziskovalca pri ravnanju z njimi

- **C - kolektivna korist:** načelo ravnanja s podatki CARE, po katerem morajo ti služiti skupnostim, na katere se nanašajo, vključujoč njihov razvoj in upoštevajoč njihove vrednote
- **A - pravica nadzora:** načelo ravnanja s podatki CARE, po katerem mora imeti pripadnik skupnosti dostop do podatkov o svoji skupnosti in biti vključen v odločitve o ravnanju in skrbništvu nad njimi
- **R - odgovornost:** načelo ravnanja s podatki CARE, po katerem mora obdelovalec podatkov svoje izsledke deliti s skupnostjo, ji z njimi pomagati pri višji stopnji samoodločbe in jo krepiti pri njenem razvoju
- **E - etičnost:** načelo ravnanja s podatki CARE, po katerem so pravice in blaginja skupnosti postavljene v središče življenjskega kroga podatkov, upoštevajoč presojo skupnosti

Načela TRUST

Načela, po katerih se zaupanja vredni digitalni repozitoriji presojuje glede na svojo transparentnost, odgovornost, osredotočenost na uporabnika, trajnost in tehnološkost

- **T - transparentnost:** načelo delovanja TRUST za digitalne repozitorije, ki vključuje najdljive, pregledne, jasne in dostopne informacije o repozitoriju in njegovih storitvah za ciljne uporabnike
- **R - odgovornost:** načelo delovanja TRUST za digitalne repozitorije, po katerem mora repozitorij zagotavljati skrbništvo nad digitalnimi objekti in storitvami za ciljne uporabnike
- **U – osredotočenost na uporabnika:** načelo delovanja TRUST za digitalne repozitorije, po katerem mora repozitorij upoštevati potrebe ciljnih uporabnikov, tako pri shranjevanju kot uporabi digitalnih objektov
- **S – trajnost:** načelo delovanja TRUST za digitalne repozitorije, po katerem mora repozitorij zagotavlja trajno skrbništvo nad digitalnimi objekti in storitvami za ciljne uporabnike
- **T – tehnološkost:** načelo delovanja TRUST za digitalne repozitorije, po katerem repozitorij temelji na medsebojnih vplivih med ljudmi, procesi in tehnologijo, tj. strojno in programsko opremo, tehničnimi storitvami, kar omogoča varne, obstojne in zanesljive storitve ciljnim uporabnikom

Temeljna načela odprte znanosti 2

- **transparentnost** znanstvenih metod, protokolov in eksperimentov – uporaba UI;
- omogočanje **uporabljenosti/ponovne rabe** raziskovalnih podatkov;
- **spodbujanje** poštenih in transparentnih metod **vrednotenja** raziskovalnega dela - merila;
- **spodbujanje in promocija povezovanja znanosti z javnostjo**;
- spodbujanje **praks odprte znanosti na drugih področjih**, kot sta npr. izobraževanje in kultura – **promocija raziskovalnih rezultatov**;
- spodbujanje sprememb v kulturi raziskovalnega dela v smeri **sodelovanja med raziskovalci in omejevanja tekmovalnosti**

Raziskovalni podatki – definirati po področjih

- zapisi o dejstvih, ki predstavljajo osnovno podlago za znanstveno raziskovanje in ki v okviru znanstvene skupnosti veljajo kot ustrezno sredstvo za preverjanje veljavnosti raziskovalnih spoznanj.
 - številčni podatki
 - besedilni, zvočni in slikovni zapisi
- Meritve: jezikoslovje, slovaropisje
- Digitalna slovarska baza (CVJT) – [podatkovni model](#)
- Skupna podatkovna baza ISJFR – priprava modela, [Pregovori](#):
 - Besedna oblika (iztočnica) – besedilna različica – govorjena različica – kako zapisati podatke
 - Pripisovanje povezanih podatkov – leme, razlage, definicije, vezljivost, udeleženosť v terminih, frazemih
 - Individualnost – generičnost rešitev

Licence – kaj smem in česa ne?

Večinoma jih določi avtor sam ali pa založnik v svoji založniški politiki.

Licenca

CC BY

CC BY-SA

CC BY-ND

CC BY-NC

CC BY-NC-ND

Kaj dovoljuje

Uporaba, deljenje, priredba

Kot CC BY, a deliti pod isto licenco

Dovoljeno deljenje, **brez priredb**

Samo nekomercialna uporaba

Kombinacije

Založba ZRC: CC BY-NC-ND

Pogoj

Navedba avtorja

Deljenje pod istimi pogoji

Brez spreminjanja

Ni dovoljenja za komercialno rabo

Brez spreminjanja in prepovedano
za komercialno rabo

Copyleft = Svobodna avtorska pravica (GNU GPL)

Dovoljuje uporabo, spreminjanje in distribucijo, a pod pogojem, da vse izpeljane verzije ostanejo odprte in pod isto licenco.

ZRCola 2 - <https://zrcola.zrc-sazu.si/>

Projekti, kot je [Kiberpipa](#) in [Ljudmila](#) – **laboratorij za znanost in umetnost**, so spodbujali uporabo **copyleft** licenc za umetniška in medijska dela.

Občanska znanost – ja ali ne? Poznavanje metod?

- Zbiranje podatkov (npr. merjenje temperature, opazovanje ptic, novih besed – trenutno funkcija *Predlagajte novo besedo* – ne deluje)
- Verifikacija pomenov in rabe – množičenje – dober opis naloge, kontrola (Besedomat)
- Označevanje, ročna anotacija korpusov – omejitev pri frazeologiji, terminologiji, prepoznavanju stalnosti, žargonih
- Zbiranje in iskanje starih virov - npr. starih tiskov, rokopisov, lokalnih zapisov, ki pomagajo razvozlati stare pomene besed ali njihovo rabo.

Prednosti

- Zajem **živega jezika**, ki ga slovaropisci težko sami beležijo sproti.
- Večji doseg – pokrivanje različnih regij, generacij, strok. – Motivacija (nagrade?)
- Večja **jezikovna ozaveščenost** in sodelovanje skupnosti.
- Povečanje **zaupanja in transparentnosti** pri izdelavi slovarjev.

Slabosti: dosledna metodologija; zanesljivost rezultatov (Slovar sopomenk – množičenje : posameznik)

Podatkovne objave in *surovi* ter *nesurovi* (*kuhani?*) podatki

- **Podatkovna objava** = objava v podatkovnem repozitoriju, ki vsebuje celovito dokumentirane podatke, opremljene z bibliografskimi informacijami in s trajnim identifikatorjem, ter pri kateri je opravljen pregled kakovosti
- **Surovi podatki** = podatki, ki so ustvarjeni ali zbrani iz različnih virov ter niso očiščeni, predelani, organizirani, interpretirani
- **Izvirni podatki** = surovi podatki, ki se zajemajo neposredno z mesta nastanka
- **Obdelani podatki** = podatki, ki so očiščeni, predelani in strukturirani, npr. razvrščeni, tipologizirani, z namenom njihove uporabe

Načrt ravnanja z raziskovalnimi podatki

- Ali obstajajo vede brez raziskovalnih podatkov?

Raziskovalni podatki niso vedno **kvantitativni** (številčni). V širšem smislu so to:

- meritve,
- opazovanja,
- zapisi,
- ankete,
- intervjuji,
- terenski zapiski,
- rokopisi, artefakti,
- introspektivni zapisi (v filozofiji, psihologiji),
- digitalne sledi (v jezikoslovju, antropologiji ipd.)

Digitalne sledi

Aktivne digitalne sledi

- Podatki, ki jih uporabnik namerno ustvari ali deli:
 - objave na družbenih omrežjih,
 - elektronska pošta,
 - izpolnjevanje obrazcev,
 - komentiranje na forumih ali blogih,
 - nalaganje videov, slik, datotek

Pasivne digitalne sledi

Podatki, ki se samodejno beležijo, brez aktivnega sodelovanja uporabnika: IP-naslovi in lokacija, piškotki (cookies), zgodovina brskanja, trajanje ogleda videoposnetka, kliki, pomiki z miško, čas preživet na strani.

Etika in zasebnost, če je vse odprto

- podpisano soglasje (pravica umika)
- informirano soglasje
- varstvo osebnih podatkov (GDPR)
- transparentna raba podatkov
- upoštevanje etičnega kodeksa (primer: jezik v komentarjih na YouTubeu - digitalne sledi uporabnikov – to je realno gradivo, ki ni bilo ustvarjeno namensko za raziskavo, a ponuja vpogled v rabo jezika v naravnem okolju)

Zakonodaja:

Resolucija o znanstvenoraziskovalni in inovacijski strategiji Slovenije 2030 (ReZrIS30) → Akcijski načrt

[Zakon o znanstvenoraziskovalni in inovacijski dejavnosti \(ZZrID\)](#)

[ZASP](#) (izjema za podatkovno rudarjenje: [57a.](#), [57b.](#), [57c.](#), [57d](#))

[Uredba o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odprte znanosti](#)

Primer NRO – terminologija I

1. Opis podatkov

1. Vrsta podatkov: terminološki zapisi, definicije, specializirani korpusi, besedila, metapodatki, ustrezniki, klasifikacije.
2. Format: .xml, .csv, .tbx (TermBase eXchange), .xls(x), .txt, .docx.
3. Izvor podatkov: primarni (luščenje), sekundarni (obstoječi terminološki slovarji, baze)

2. Zbiranje in dokumentiranje podatkov

1. Zajem, ročno beleženje
2. Strojno luščenje
3. Potrditev/validacija strokovnjakov
4. Metapodatki (npr. izvor, avtor, pomen, kontekst, raba, področje)
5. Orodje: Excel, Word, namenski program

3. Shranjevanje in varnost

1. Mesto hrambe (lokalno, strežnik, oblak)
2. Varnostno kopiranje, zaščita z geslom, omejen dostop

Primer NRO – terminologija II

4. Etika in pravni vidiki:

1. Avtorske pravice: ustrezno citiranje; uporaba besedil bo omejena na dele, dovoljenje bo pridobljeno, kjer potrebno.
2. Osebni podatki: če se uporablja terminologija iz intervjujev, bo pridobljeno informirano soglasje, podatki pa anonimizirani.
3. Licenciranje rezultatov: rezultati bodo objavljeni pod odprto licenco (npr. CC BY 4.0 ali ODC-BY za podatkovne baze), uporabnikom bo omogočeno prosto uporabo terminov s sklicem na vir.

5. Raba in deljenje podatkov

1. Dostopnost: podatki bodo deljeni v repozitoriju (npr. RDA-SI, CLARIN.SI, Zenodo), DOI
2. Način deljenja: javno dostopen terminološki slovar prek spleta (npr. Terminologišče)
3. Ciljna skupina: raziskovalci, prevajalci, študenti, stroka.

6. Dolgotrajno hranjenje

1. Repozitorij: dolgoročno shranjevanje v Zenodo, **Clarin.si**?
2. Trajanje hrambe: 10 let

NRO III

7. Odgovornost

1. Glavni odgovorni za podatke: vodja projekta
2. Tehnična podpora: inštitutski IT ali repozitориjski partner (npr. CLARIN.SI).
3. Redno posodabljanje dokumentacije: vsakih 6 mesecev.

Podatki v standardizirani obliki (TBX, RDF, JSON-LD, če je mogoče).

Povezava terminoloških zapisov z ontologijami ali konceptualnimi shemami (če je relevantno).

Oznaka statusa terminoloških kandidatov: delovni, preverjeni/potrjeni, predlagani.

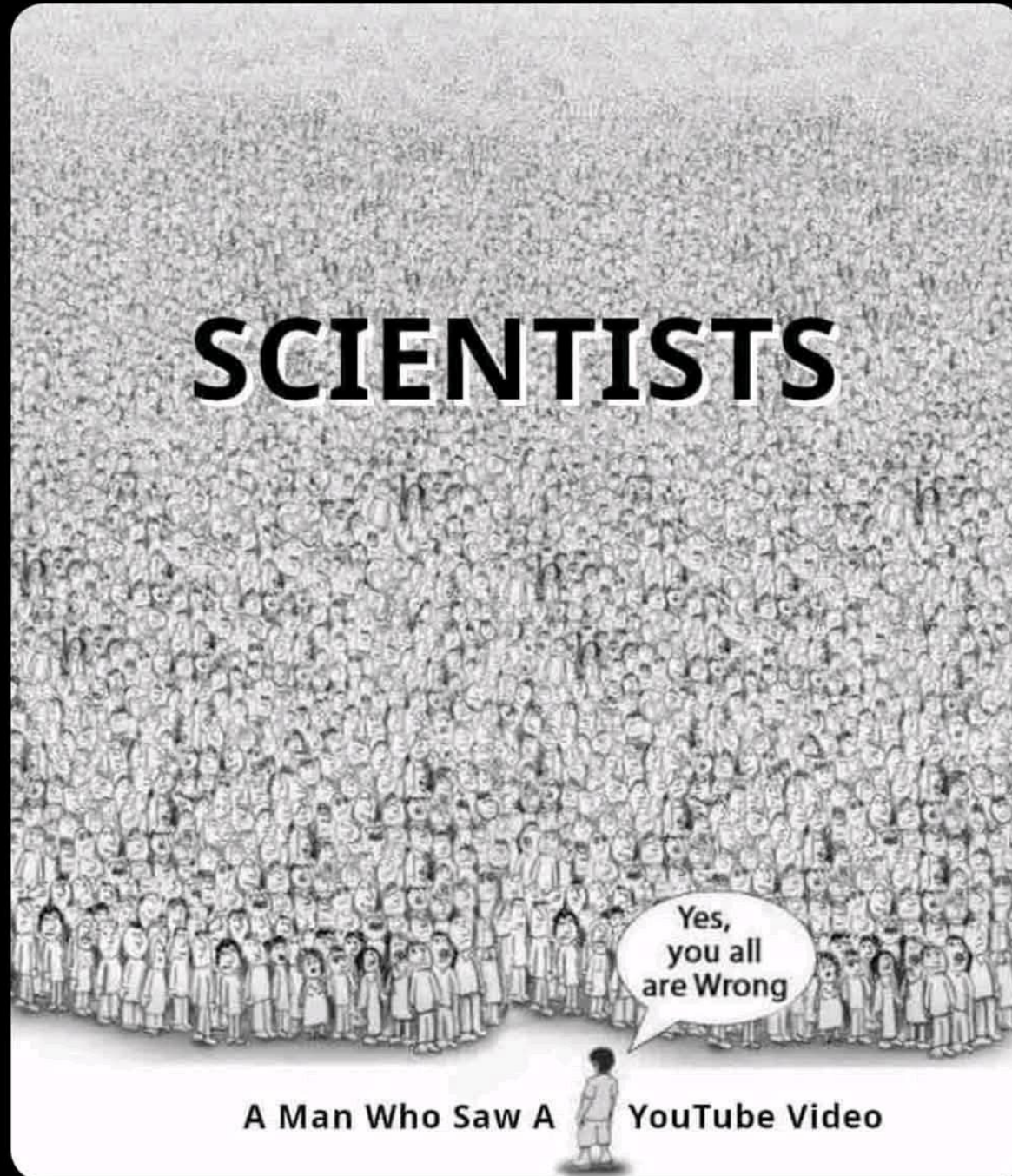
Življenjski krog raziskave

[Življenjski krog podatkov](https://projekt-spoznav.si/video/definicija-in-zivljenjski-krog-raziskovalnih-podatkov/) (<https://projekt-spoznav.si/video/definicija-in-zivljenjski-krog-raziskovalnih-podatkov/>)

Arhiviranje



ZRC SAZU



A Man Who Saw A YouTube Video



Viri

- [Projekt SPOZNAJ](#)
- [Slovenska skupnost odprte znanosti](#)
- [Nacionalni portal odprte znanosti](#)
- P. Kamocki, E. Ketzan, J. Wildgans: [Language Resources and Research under General Data Protection Regulation](#), White paper CLIC, september 2018
- Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavrilidou, Pavel Straňák: [Processing personal data without the consent of the data subject for the development and use of language resources](#), 2019
- Tomaž Erjavec: [Raziskovalna infrastruktura za jezikovne vire in tehnologije CLARIN.SI](#)



ZRC SAZU

• Inštitut za slovenski jezik Frana Ramovša • Inštitut za arheologijo • Zgodovinski inštitut Milka Kosa • Umetnostnozgodovinski inštitut Franceta Steleta • Muzikološki inštitut • Inštitut za slovensko literaturo in literarne vede • Inštitut za slovensko narodopisje • Inštitut za slovensko izseljenstvo in migracije • Filozofski inštitut • Geografski inštitut Antona Melika • Inštitut za raziskovanje krasa • Paleontološki inštitut Ivana Rakovca • Biološki inštitut Jovana Hadžija • Glasbenonarodopisni inštitut • Družbenomedicinski inštitut • Inštitut za kulturno zgodovino • Inštitut za antropološke in prostorske študije • Inštitut za kulturne in spominske študije •