Pregled storitev CLARIN.SI za ravnanje z jezikovnimi raziskovalnimi podatki

Kaja Dobrovoljc
Filozofska fakulteta UL
Institut Jožef Stefan

Letno srečanje Doktorske šole UL 2022
UL FDV, 2. junij 2022

CLARIN

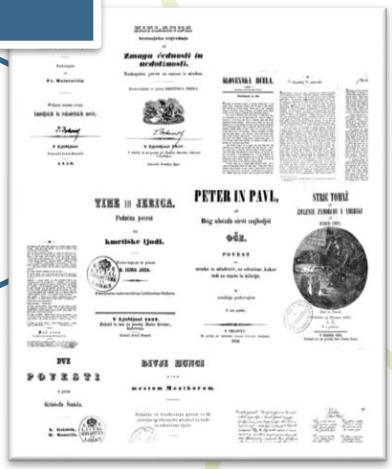
Nekaj primerov jezikovnih podatkov

Korpus pisnih besedil Gigafida

- časopisi, revije, spletne novice, priročniki, leposlovje
- 38.000 besedil
- 1 milijarda besed

Korpus starejših besedil IMP

- besedila od konca 16. stol. do 1918
- 650 enot
- 14 milijonov besed



Korpus spletnih besedil JANES

- tviti, besedila forumov, komentarji na novice, blogi …
- 13 milijonov zapisov, 252 milijonov besed

Korpus parlamentarnih razprav ParlaMint

- parlamentarne razprave v 17 jezikih
- za slovenščino 2014-202
- 23 milijonov besed

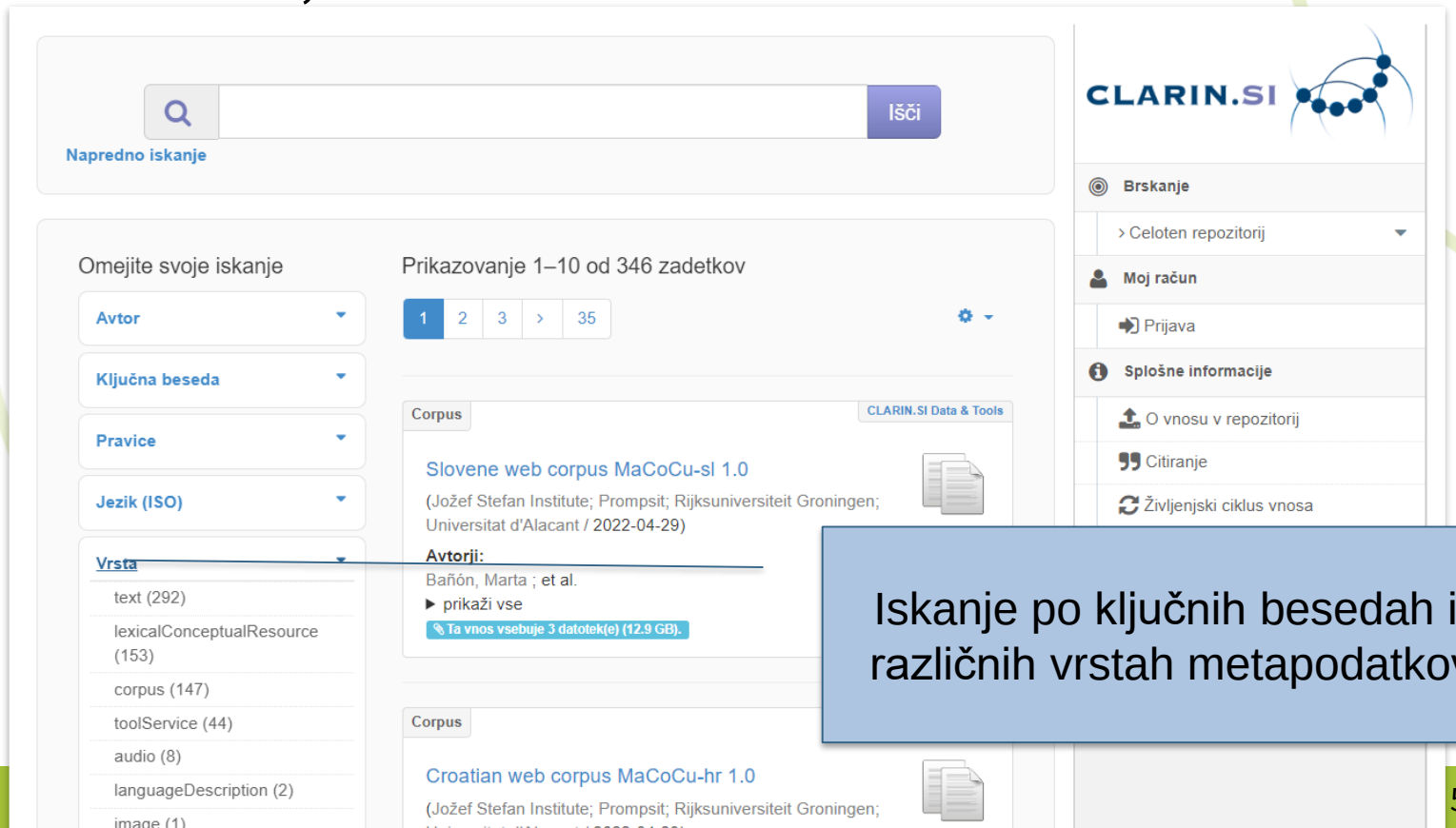
UPORABA OBSTOJEČIH PODATKOV

Pregled obstoječih podatkovnih zbirk

- zbiranje, obdelava in objava novih podatkov vzamejo **veliko časa** in zahtevajo **dodatna znanja**
- uporabni vstopni točki za pregled obstoječih jezikovnih raziskovalnih podatkov:
  - globalno: **Virtual Language Observatory**
  - lokalno: **repozitorij CLARIN.SI**

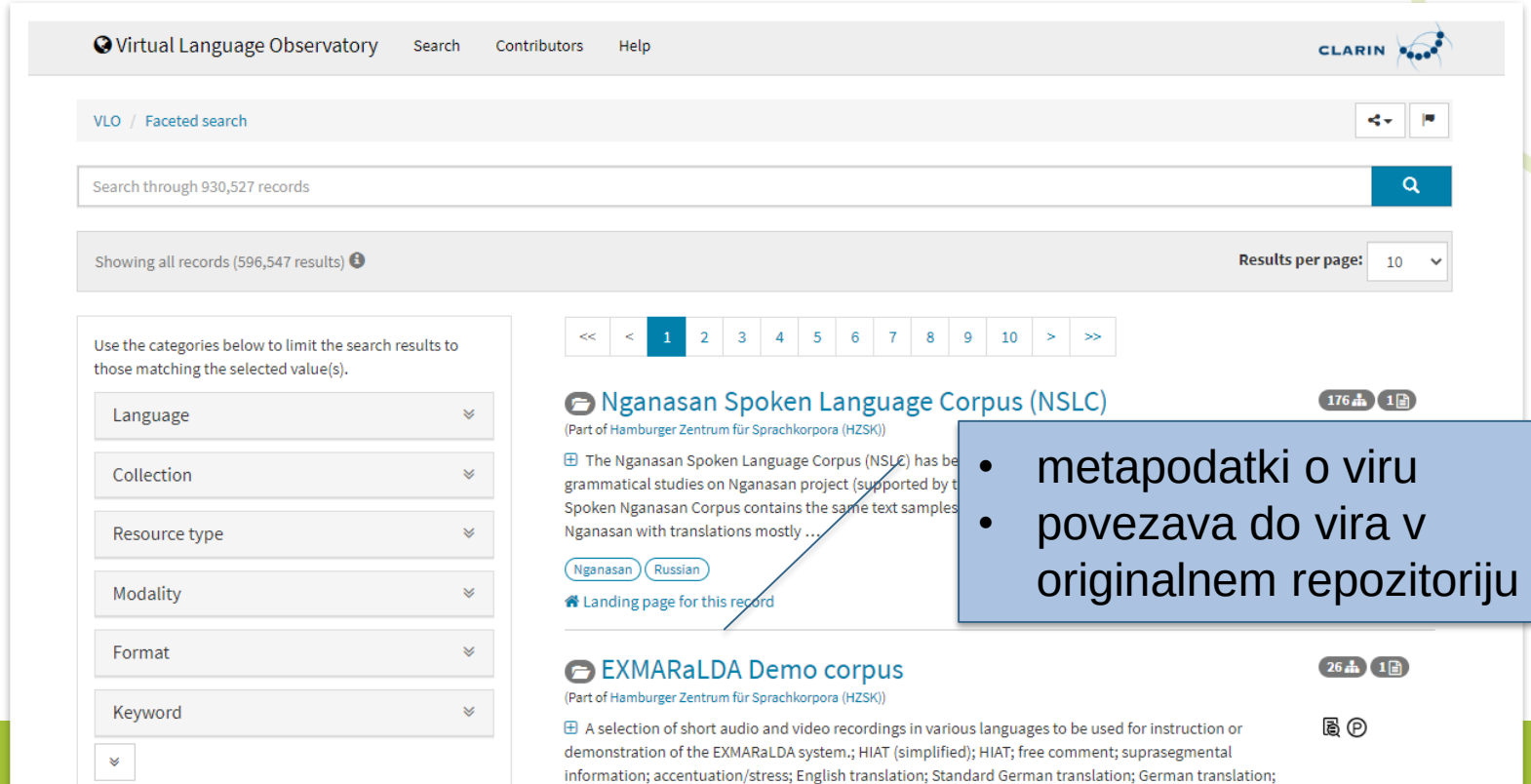
Iskanje po repozitoriju CLARIN.SI

- https://www.clarin.si/repository/xmlui/
- stalna in varna hramba jezikovnih virov (tudi) za slovenščino; več kot 300 vnosov



Iskanje po Virtual Language Observatory

- vlo.clarin.eu
- skupni iskalnik po vseh jezikovnih virih v nacionalnih repozitorijih CLARIN in drugih podatkovnih repozitorijih

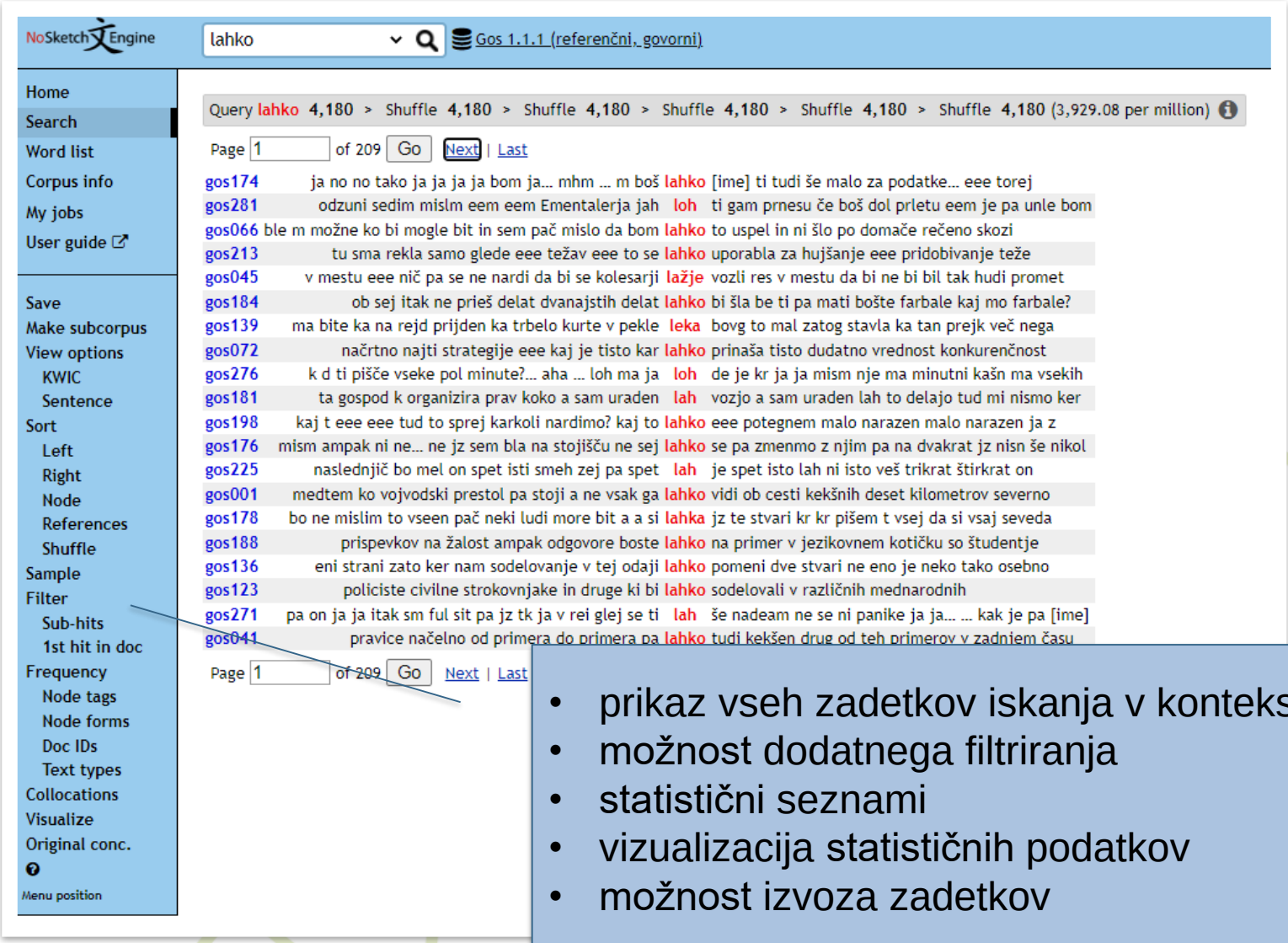


Brskanje po korpusih CLARIN.SI

- poleg prevzema datotek običajno v opisu vira navedene tudi **druge možnosti dostopa** do vsebovanih podatkov (npr. spletni vmesniki, priporočena orodja)

- za **iskanje po besedilnih korpusih** podporo nudi tudi CLARIN.SI preko dveh spletnih konkordančnikov:
  - noSketchEngine: clarin.si/noske
  - Kontext: clarin.si/kontext

Primer iskanja po noSketchEngine



- prikaz vseh zadetkov iskanja v kontekstu
- možnost dodatnega filtriranja
- statistični seznami
- vizualizacija statističnih podatkov
- možnost izvoza zadetkov

ZBIRANJE NOVIH PODATKOV

Zbiranje jezikovnih podatkov

- številne klasične in inovativne metode
  - zbiranje besedil
  - snemanje in transkripcija
  - vprašalniki
  - mobilne aplikacije
  - spletno pajkanje
  - …

Pravni vidiki – koristne povezave

- uporabne pravne informacije o **avtorskih pravicah, licenciranju in varstvu osebnih podatkov**: https://www.clarin.eu/content/legal-information-platform
- orodje za pomoč pri oblikovanju soglasja za sodelovanje v raziskavi (GDPR): https://consent.dariah.eu
- vzorec pogodbe CLARIN.SI za prenos avtorskih pravic za namene izdelave in objave korpusa

Načini zapisa podatkov

- izbira načina zapisa podatkov (formata) je ključna za zagotavljanje njihove nadaljnje uporabnosti
  - izkoristite tehnično podporo
- nekaj priporočil CLARIN.SI:
  - strojna berljivost in pretvorljivost
  - uporaba standardov
  - dobre prakse: .xml, .csv, .tsv, .txt ...
  - odsvetovano: .docx, .xlsx, .pdf
- enako pomemben tudi zapis metapodatkov

Hranjenje podatkov

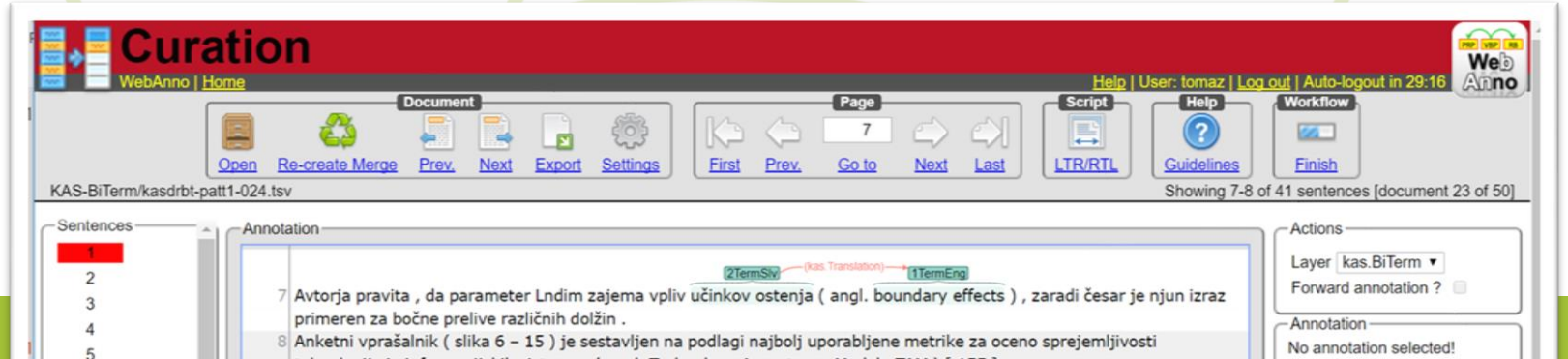
- sprotna in sistematična **dokumentacija**
  - povedna imena datotek, eksplicitne pripone
  - številčenje različic, stiskanje (.zip)
  - beleženje sprememb v obliki dnevnika
- **varnostna kopija** podatkov
  - 1TB prostora v oblaku OneDrive za vse z identiteto UL

Označevanje besedil (opcijsko)

- golim besedilom lahko na ravni besed, stavkov ali celotnih dokumentov strojno ali ročno pripišemo tudi različne oznake oz. kategorije
- nekaj tipičnih primerov:
  - slovnične lastnosti besed
  - sentiment stavkov
  - tema besedila
- številne prednosti označenih besedil
  - **lažje iskanje** (priklic zadetkov po oznakah)
  - **kvalitetnejša analiza** podatkov
  - računalniška uporaba (strojno učenje)

Označevanje besedil (opcijsko)

- za avtomatsko splošno procesiranje smiselno preveriti že obstoječa računalniška orodja
  - CLARIN.SI Repozitorij > toolService
  - CLARIN.SI GitHub: https://github.com/clarinsi
  - CLARIN Language Resource Switchboard: https://switchboard.clarin.eu/
- za ročno označevanje na voljo spletno orodje WebAnno
  - https://www.clarin.si/webanno/



OBJAVA PODATKOV

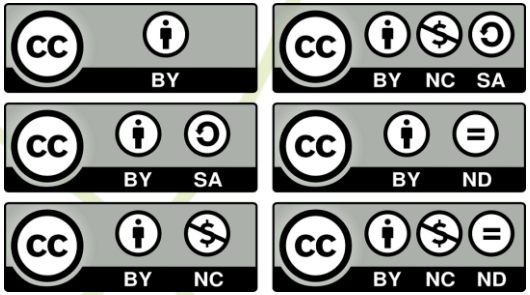
Objava vira v repozitoriju CLARIN.SI

1. vnos ustvari avtor (Ustvarite nov vnos)
2. prijava preko ponudnika identitete (EduGain)
3. urednik pregleda, opozori na popravke in objavi
4. vir postane viden navzven
   - tudi preko Google, VLO, DataCite, arXive …
   - korpusi običajno integrirani tudi v konkordančnike

- pred vnosom nujno prebrati dokumentacijo
  - Kako ustvariti vnos
  - Pogosta vprašanja

Licence

- ob vnosu vira morajo avtorji opredeliti tudi licenco
- velika večina virov je dostopna pod eno od licenc Creative Commons (CC)
- mogoče je izbrati tudi številne druge ali ustvariti novo
- v procesu objave v pomoč '**OPEN License Selector**'



Razpis za projekte CLARIN.SI

- Letni razpis CLARIN.SI, namenjen izdelavi ali nadgradnji virov ali storitev, ki pripomorejo k uresničevanju usmeritev infrasrukture CLARIN(.SI)
  - izdelava ali nadgradnja virov, spletnih storitev ali programske opreme
  - organizacija izobraževalnih dogodkov oz. priprava izobraževalnih gradiv
  - raziskave, ki uporablja obstoječe vire ali storitve CLARIN(.SI)
- rok za prijave običajno **spomladi**
- v 2022 sredstva v višini **2.000-10.000 EUR**

Obrazec NRRP

Tip podatkov in metode njihovega zbiranja in/ali ustvarjanja

1. Katere podatke boste zbirali oziroma pridobivali in/ali ustvarjali?
2. Na kakšen način boste zbirali oziroma pridobivali in/ali ustvarjali nove podatke in kako boste uporabljali že obstoječe za potrebe vaše doktorske disertacije?
3. Ali boste delali z občutljivimi podatki? Če da, kako boste poskrbeli za etično pridobivanje in/ali ustvarjanje podatkov?

- široka uporabnost jezikovnih podatkov
- pregled že obstoječih zbirk v repozitoriju CLARIN.SI in agregatorju VLO
- pred zbiranjem urediti avtorske pravice, varstvo osebnih podatkov in pogoje uporabe

Način hranjenja in zaščita podatkov med raziskovalnim delom za doktorsko disertacijo

1. Na kakšen način boste hranili podatke?
2. Če boste delali z občutljivimi podatki, kako boste skrbeli za njihovo varovanje in zaščito?

- strojno berljivi formati zapisa
- sistematična dokumentacija
- varnostno kopiranje
- anonimizacija

Dolgotrajna dostopnost in hranjenje podatkov

1. Kje oziroma v katerem podatkovnem repozitoriju boste hranili podatke na dolgi rok po zaključku raziskovalnega dela in omogočili njihovo dostopnost?
2. Ali načrtujete za določen čas omejiti dostop do podatkov?

- objava podatkov v repozitoriju CLARIN.SI
- odprta licenca kot privzeta izbira