arnes

NI4OS

National Initiatives for Open Science in Europe

Jezikoslovna infrastruktura za slovenščino: CLARIn.SI in RI-Si - CLARIN

Darinka Verdonik, UM FERI

Mreža znanja 2021, od 23. do 30. novembra

CLARIN.SI

CLARIN.SI je slovenski nacionalni konzorcij v mreži evropske raziskovalne infrastrukture CLARIN.

Raziskovalcem na področju humanistike, družboslovja in drugih, z jezikom povezanih ved zagotavlja:

- jezikovne vire in tehnologije ter
- strokovno podporo in prenos znanja.

Na ta način želi razširiti tehnološke možnosti raziskovanja jezikov, predvsem slovenščine in drugih južnoslovanskih jezikov, ter spodbujati meddisciplinarno sodelovanje.

Računalniško podprto raziskovanje jezika: jezikoslovje in tehniške vede

KORPUSI
- referenčni
- specializirani
- prevodi

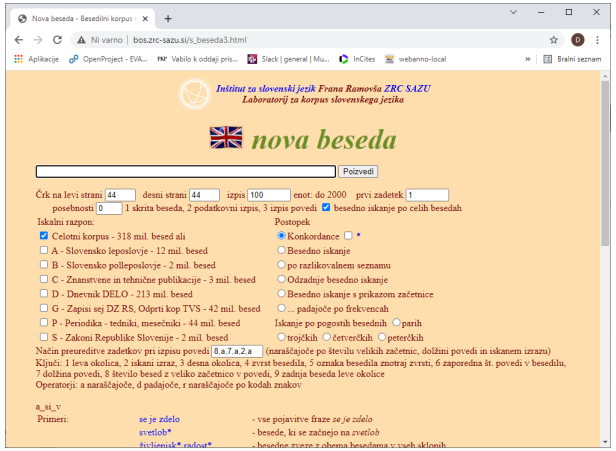
GOVORNE BAZE
- posnetki govora
- zapis govora

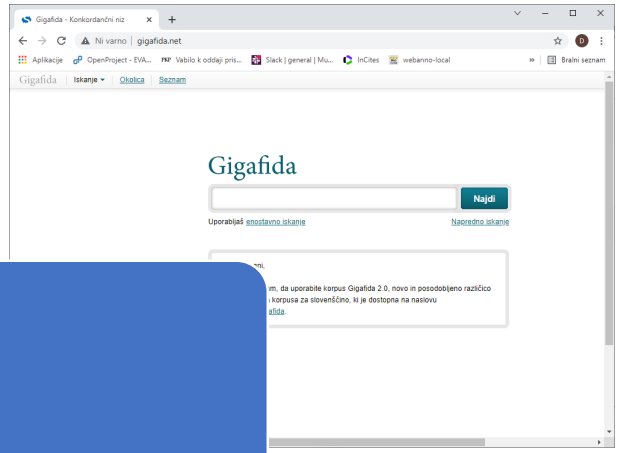
PODATKI O BESEDIŠČU
- seznami besed, sopomenk, kolokacij, koligacij…

ORODJA
- tokenizator in lematizator
- oblikoslovni označevalnik
- skladenjski razčlenjevalnik
- označevanje imenskih entitet
- semantična analiza…

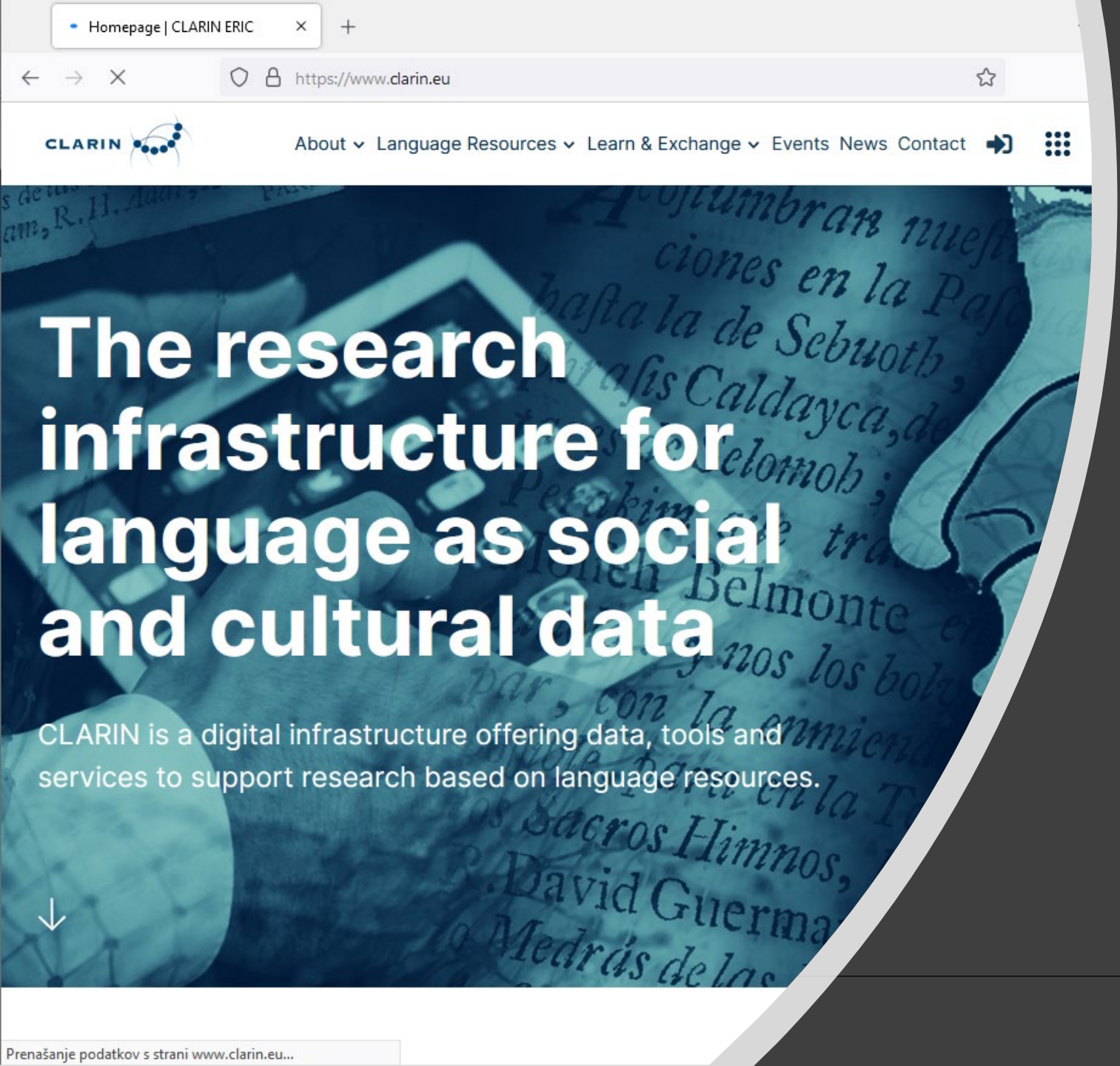
Zgodovinski pregled

Stihijski razvoj -> istovrstni vir re je razvijal vedno znova



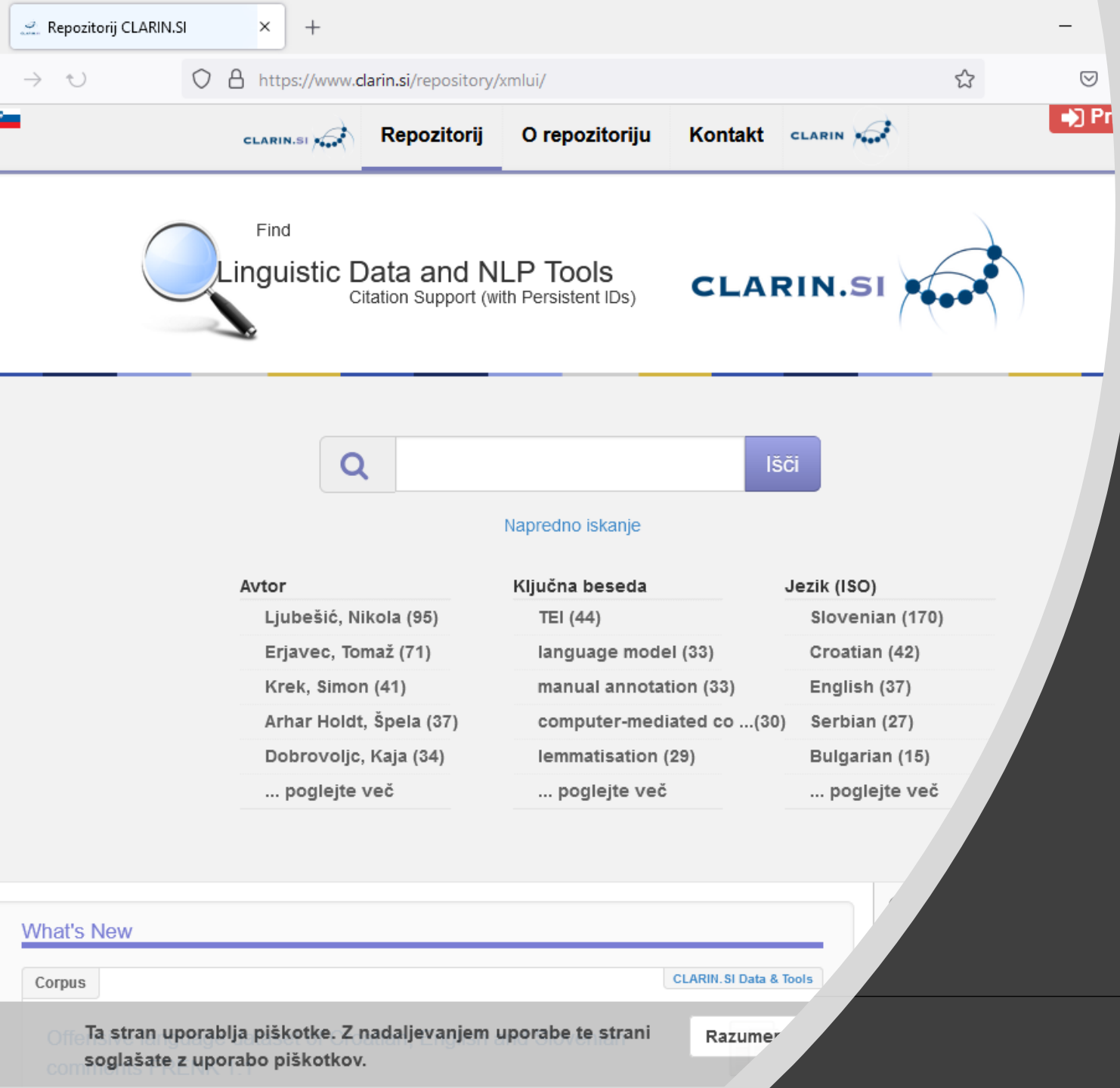


Sodelovanje, prostodostopna in odprta infrastruktura



CLARIN – ERIC

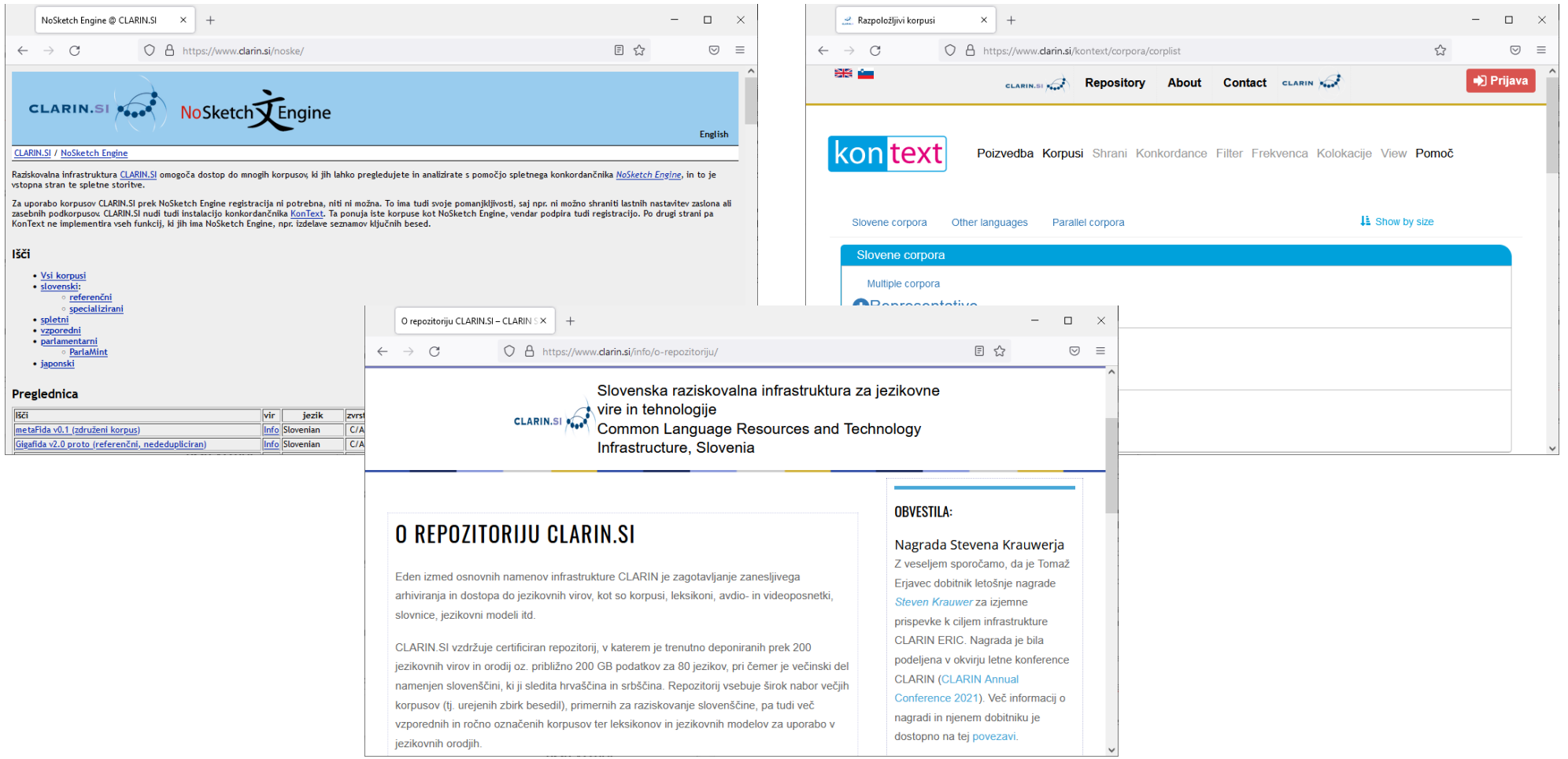
arnes



CLARIN.SI

arnes

REPOZITORIJ CLARIN.SI



RI-SI - CLARIN

CILJI

- zagotoviti nadaljnje delovanje tehničnih storitev infrastruktureCLARIN.SI
- omogočiti hranjenje velikih multimodalnih jezikovnih podatkov
- omogočiti, da CLARIN.SI sledi paradigmi »velepodatkov« (angl. big data)
- omogočiti, da CLARIN.SI ponuja javno dostopne spletne storitve za obdelavo velikih količin slovenskih besedil
- vzpostaviti namensko gručo računalnikov s pospeševalniki GPGPU za potrebe globokega učenja

NABAVLJENA OPREMA

INSITUT JOŽEF STEFAN

- Gruča za spletne storitve
- Strežnik repozitorija
- Diskovno polje za varnostne kopije
- Stikalo za optični kanal

UNIVERZA V MARIBORU

- NVIDIA DGX-1
- Strežniška infrastruktura za obdelavo in hranjenje velepodatkovnih jezikovnih virov

UNIVERZA V LJUBLJANI

- Strežniška rezina

ZAKLJUČEK

- Predstavili smo odprto in prosto dostopno jezikoslovno infrastrukturo CLARIN.SI.

- V jezikoslovju in jezikovnih tehnologijah je bilo odpiranje podatkov in oblikovanje skupnega konzorcija temeljni korak k učinkovitejšemu in ekonomičnejšemu razvoju področja.

Hvala za pozornost

darinka.verdonik@um.si