

Repozitorij CLARIN.SI

Tomaž Erjavec¹ in Katja Meden^{1,2}

¹Odsek za tehnologije znanja, Institut "Jožef Stefan",
²Inštitut za novejšo zgodovino

Predstavitev CLARIN.SI podatkovnim skrbnikom UL
2024-06-11

Kazalo

- 1 Uvod
- 2 Življenjski cikel podatkov
- 3 Vmesnik repozitorija
- 4 Diskusija

Uvod

CLARIN: Common Language Resources and Technology Infrastructure

- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni prek enotne prijave za raziskovalce v humanistiki in družboslovju
- Namenjena je dolgotrajnemu in obsežnemu hranjenju ter dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Sodelovanje pri razvoju virov in orodij, zagotavljanje večkratne uporabnosti in prilagajanja individualnim potrebam



Kaj CLARIN ERIC ponuja slovenskim raziskovalcem?

- S slovensko EduGain prijavo dostop do vseh virov in storitev centrov CLARIN držav članic
- Spletne storitve, npr. virtualni jezikovni observatorij
- Podpora ciljnim projektom, npr. razvoju učnih vsebin, izvedbi delavnic za uporabnike, snovanju evropskih projektnih prijav
- Infrastruktura znanja:



Knowledge centres



Workshops



Digital Humanities Course Registry



Mobility Grants



Oral history & technology



Trainer Network Programme



Annual Conference



Training Suite



VideoLectures



Support for EU-funded projects

CLARIN.SI



- Začetek dela v 2015
- Institut "Jožef Stefan":
 - Odsek za tehnologije znanja (E8)
 - Laboratorij za umetno inteligenco (E3)
 - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 partnerjev:
 - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
 - 4 raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
 - 2 podjetji: Amebis, Alpineon
 - 1 knjižnica: NUK
 - 1 društvo: Slovensko društvo za jezikovne tehnologije

Delovanje CLARIN.SI

Trije stebri

- ① **Repozitorij jezikovnih virov in orodij**
- ② Spletne storitve: konkordančniki itd.
- ③ Podpora digitalni humanistiki in jezikovnim tehnologijam:
 - CLASSLA K-centre za rač. obdelavo južnoslovanskih jezikov
 - financiranje projektov
 - sodelovanje v projektih
 - izobraževanja in predstavitve
 - organizacija in podpora dogodkom

Repozitorij CLARIN.SI

- Trenutno najpomembnejša storitev CLARIN.SI
- Vsebuje okoli 600 jezikovnih virov in orodij (5TB), od tega 300 (tudi) za slovenski jezik:
korpora, slovarji, besedišča, modeli, programi
- Samoarhiviranje + uredniški pregled
- Certificacija: CLARIN in (delno) Core Trust Seal
- Dolgotrajno hranjenje, avtentikacija in avtorizacija, stalni identifikatorji, eksplicitni pogoji uporabe, izbira licence
- Podatki hranjeni po načelih FAIR

Zakaj uporabljati repozitorij CLARIN.SI?

Prevzem

- Širok nabor jezikovnih virov, predvsem za slovenski in druge južnoslovanske jezike
- Večina virov dostopna brez prijave pod licencami CC
- Kvalitetni in preverjeni (meta)podatki
- Uporabniku prijazen vmesnik in iskalnik

Deponiranje

- Zaupanja vreden repozitorij (CTS)
- Večja možnost citiranja podatkov
- Uredniška podpora

Življenjski cikel podatkov

Od stvaritve vnosa do objave

Oddaja vnosa

- Enotna prijava v repozitorij CLARIN.SI
- Ustvarjanje novega vnosa (v skladu z **navodili**)
- Oddaja vnosa v uredniški pregled

Uredniški pregled:

- Urednik vnosa pregleda tehnične zahteve:
 - Zahtevani metapodatki v pravilni obliki
 - Format podatkov (datotek)
 - Validacija podatkov (predvsem XML, CoNNL-U)
- Odločitev urednika:
 - Vnos je tehnično primeren: potrditev in objava vnosa
 - Vnos ni primeren: zavrnitev vnosa, vrnjen avtorju s seznamom potrebnih popravkov
 - Popravki, ponovna oddaja in uredniški pregled

Urejanje objavljenega vnosa

Objavljen vnos

- Podatki in metapodatki postanejo prosto dostopni
 - Omejeni vnosi: prosto dostopni zgolj metapodatki
- Metapodatki so agregirani prek protokola OAI-PMH

Brisanje vnosa

- Vsak lahko zaprosi za umik podatkov
- Metapodatki se ohranijo (PiD, razen v zelo specifičnih primerih)

Spreminjanje vnosa

- Manjše (tipkarske) napake popravimo uredniki (Pomoč uporabnikom)
- Večje spremembe zahtevajo oddajo nove različice vnosa

Nove različice vnosa

- Novo različico vnosa lahko ustvari zgolj oseba, ki je vnos oddala
- Dodajanje/spreminjanje metapodatkov in nalaganje novih podatkov (uredniški pregled)
- Stara in nova različica se avtomatsko povežeta

🔍 Druge različice

[Prikaži polni zapis vnosa](#)

📎 Datoteke v tem vnosu

Seznam vseh različic ▾

- ▶ Multilingual comparable corpora of parliamentary debates ParlaMint 2.1
- Multilingual comparable corpora of parliamentary debates ParlaMint 2.0
- Multilingual comparable corpora of parliamentary debates ParlaMint 1.0

Omejeni vnosi in licence

- Spodbujamo prosti dostop, vendar imajo določeni vnosi omejen dostop (omejevalne licence)
- Metapodatki omejenih vnosov so javno dostopni, za prenos podatkov je potreben e-podpis
- Vodimo evidenco elektronskih podpisov za primere kršitev pravic intelektualne lastnine

Licence

- Licence za javni dostop (Public licences):
 - Creative Commons licence
 - GNU, MIT, Apache licence
 - Ne potrebujejo nobenih dodatnih informacij
- Licence z omejenim dostopom (Academic licences):
 - CLARIN.SI licence
 - MULTEXT-East licence
 - Potrebna prijava in posredovanje dodatnih informacij (npr., država, email, namen uporabe, itd.)
- Če uporabnik ne ve, kakšno licenco bi izbral, lahko uporabi Izbirnik licenc "OPEN Licence Selector"

OPEN Licence Selector

Izbirnik

- Izbirnik uporabniku zastavi vprašanja o naravi podatkov in zahtevah dostopa
- Na podlagi odgovorov mu ponudi najprimernejšo (med tistimi, ki jih ponuja repozitorij)

Izberite licenco za vir



Izbirnik licenc vam nudi vizualno pomoč pri izbiri ustrezne licence za vaše podatke ali programsko opremo. Za vse podprte licence in njihove podrobnosti obiščite stran seznama licenc .

Izbirnik licenc "OPEN License Selector"

- ALI -

Če že veste, pod katero licenco želite distribuirati svoje delo, jo izberite na spustnem seznamu.

Izberite licenco ...

Vmesnik repozitorija

Repozitorij CLARIN.SI - Vmesnik in vnosi

**Repozitorij**[O repozitoriju](#)[Kontakt](#)

Deposit Free and Safe
License of your Choice (Open licenses encouraged)
Easy to Find
Easy to Cite



[Napredno iskanje](#)

Avtor	Ključna beseda	Jezik (ISO)
Ljubešič, Nikola (134)	lexicographic resource (101)	Slovenian (266)
Erjavec, Tomaž (84)	general dictionary (66)	English (90)
Krek, Simon (51)	modern dictionary (64)	Croatian (63)
Arhar Holdt, Špela (48)	monolingual dictionary (64)	Serbian (54)
Dobrovoljc, Kaja (46)	TEI (52)	French (29)
... pogledajte več	... pogledajte več	... pogledajte več

Iskanje in brskanje



parlamint

Išči

Napredno iskanje

Omejite svoje iskanje

Avtor

Ključna beseda

Jezik (ISO)

Vrsta

Prikazovanje 1–10 od 16 zadetkov

1

2

>



Corpus

CLARIN.SI Data & Tools

Multilingual comparable corpora of parliamentary debates ParlaMint 4.1
(CLARIN ERIC / 2024-06-03)
Avtorji:
Erjavec, Tomaž ; et al.
▶ **prikaži vse**
 Ta vnos vsebuje 30 datotek(e) (5.87 GB).

Publicly Available  

Repozitorski vnos 1/3 - enostavni izpis

Multilingual comparable corpora of parliamentary debates ParlaMint 4.1



Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko:

BIBTEX

CMDI

Erjavec, Tomaž; et al., 2024, *Multilingual comparable corpora of parliamentary debates ParlaMint 4.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1912>.



Delite:



CLARIN.SI Data & Tools



Avtorji

Erjavec, Tomaž; et al.

► prikaži vse



Identifikator vnosa

<http://hdl.handle.net/11356/1912>



URL projekta

<https://www.clarin.eu/content/parlamint>



Demo URL

<https://github.com/clarin-eric/ParlaMint/>



Dokumentirano v

<https://doi.org/10.21203/rs.3.rs-4176128/v1>



Datum objave

2024-06-03



Vrsta

corpus, text



Velikost

8073406 utterances, 1231036093 words

Repozitorski vnos 2/3 - podatki vnosa

Ključne besede

parliamentary debates

COVID-19

TEI

Parla-CLARIN

Czech Parliament

Icelandic Parliament

Belgian Parliament

Danish Parliament

Dutch Parliament

Turkish Parliament

Italian Parliament

Hungarian Parliament

Latvian Parliament

Bulgarian Parliament

Croatian Parliament

Polish Parliament

Slovenian Parliament

French Parliament

Austrian Parliament

Bosnian Parliament

Catalonian Parliament

Galician Parliament

Greek Parliament

Norwegian Parliament

Serbian Parliament

Swedish Parliament

Ukrainian Parliament

Finnish Parliament

Spanish Parliament

Estonian Parliament

Basque Parliament

Portuguese Parliament

UK Parliament

Zbirke CLARIN.SI data & tools

Druge različice

Seznam vseh različic ▾

Prikaži polni zapis vnosa

Datoteke v tem vnosu


Prenesi navodila za ukazno vrstico

To je vnos **Publicly Available** z licenco:
Creative Commons - Attribution 4.0 International (CC BY 4.0)

 

Ime ParlaMint-AT.tgz
Velikost 315.2 MB
Format Neznano
Opis Austrian corpus
MD5 e3bfc9b090cffe16cb6bfc36c9672acb



 Prenesi datoteko



Repozitorski vnos 3/3 - polni izpis

dc.language.iso	fin
dc.language.iso	eus
dc.publisher	CLARIN ERIC
dc.relation.isreferencedby	https://doi.org/10.21203/rs.3.rs-4176128/v1
dc.relation.replaces	http://hdl.handle.net/11356/1859
dc.rights	Creative Commons - Attribution 4.0 International (CC BY 4.0)
dc.rights.uri	https://creativecommons.org/licenses/by/4.0/
dc.rights.label	PUB
dc.source.uri	https://www.clarin.eu/content/parlamint
dc.subject	parliamentary debates
dc.subject	COVID-19
dc.subject	TEI
dc.subject	Parla-CLARIN
dc.subject	Czech Parliament
dc.subject	Icelandic Parliament
dc.subject	Belgian Parliament

Diskusija

Negativne izkušnje

- Uporaba podatkov z repozitorija:
kljub temu, da na vrhu vsake strani vnosa piše točno kako je potrebno podatke citirati, večina uporabnikov podatkov ne citira pravilno ali sploh
- Deponiranje podatkov v repozitorij: kljub temu, da imamo natančna navodila kako vnašati metapodatke in kakšni naj bodo podatki, velika večina uporabnikov teh navodil ne prebere
- Zelo koristno opomniti (potencialne) uporabnike o teh dveh postavkah

Pravne prepreke

- Besedilni podatki: avtorske pravice, zaščita zasebnosti, pogoji uporabe
- Odgovornost za pregled pravnih vidikov: avtorji ali uredniki?
- Politika CLARIN.SI: avtorji neposredno odgovorni, uredniki posredno (svetujejo, če mislijo, da bi lahko bili podatki problematčni)
- Restriktivne licence, anonimizacija, vzorčenje, itd.
- Zakonodaja na področju avtorskih pravic (ZASP) proti luščenju podatkov

