# Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

JOTA
FRI UL
2019-05-28

# Uvod

# Kdo potrebuje jezikovne vire?

- Računalniško jezikoslovje / jezikovne tehnologije:
  - nadzorovano in nenadzorovano strojno učenje
  - programi so večinoma jezikovne neodvisni
  - potrebujejo pa učne (ročno označene) jezikovne vire
  - + testne podatke
- Empirično podprte jezikoslovne raziskave:
  - temeljijo na realnih (in po možnosti označenih) besedilih
  - slovaropisje, korpusno jezikoslovje, poučevanje jezikov, ...
- Jezikovne vire potrebujemo za vsak jezik posebej
- Kje lahko dobimo take vire za slovenščino?

# Jezikovni viri

## Korpusi

- enovito kodirana in dokumentirana zbirka besedil
- označena: oblikoskladnja, leme, skladnja, imenske entitete, ...
- referenčni/specializirani; eno/večjezični; pisni/govorni

## Leksikoni

- besedišče jezika
- besede ali večbesedne enote
- oblikoslovje, skladnja, pomen, povezave, prevodi, ...

## Modeli jezka

Podatki za nek program, ki mu omogoči označevanje besedil v nekem jeziku za neko raven označevanja

# Ponovna uporaba

## Klasični pristop

- za vsako raziskavo posebej izdelati jezikovne vire
- viri dostopni samo razvijalcem

## Slabosti

- izdelava jezikovnega vira je lahko zelo draga in dolgotrajna, velika izguba časa in denarja, če se to počne večkrat
- kasnejši raziskovalci ne morejo preveriti ali poboljšati prvih rezultatov
- vzdržuje se monopol raziskovalcev oz. institucij, ki so vire izdelale
- viri ne morejo biti uporabljeni pri razvoju produktov

# Odprt dostop do rezultatov raziskovalnih projektov

## Brez ovir do publikacij in podatkov

- prihranek denarja in časa
- izogibanje ponavljanju dela
- spodbujanje sodelovanja
- večja transparentnost znanstvenega procesa
- spodbujanje inovacij

## Principi "FAIR"

- Findable, Accessible, Interchangeable, Reusable
- Projekti EU za omogočanje odprtega dostopa: EOSC
- FACT: fair, accurate, confidential, transparent

# Problemi pri zagotavljanju odprtega dostopa

- Vir obstaja, a ga je težko najti
- Vir nima opisa (metapodatkov), ali pa je le-ta pomanjkljiv
- Tehnične ovire: posebni in nedokumentirani formati zapisa
- Pravne:
  - avtorske pravice nad besedili
  - varovanje zasebnosti (tudi pravica do pozabe): GDPR
  - pogoji uporabe spletnih portalov (npr. Twitter)
- Več dela za izdelovalce virov!

# CLARIN

# Raziskovalne infrastrukture

## Kaj je RI?

Naprave, viri in storitve, ki jih uporablja znanstvena skupnost za izvajanje vrhunskih raziskav na svojih področjih.

# Raziskovalne infrastrukture ESFRI

- Pobuda EU: ESFRI (European Strategy Forum on Research Infrastructures), osnovan 2002
- Roadmap: predlagali 15 (2016: 21) RI, nekatere že delujejo kot ERIC (EU pravna oseba: European RI Consortium)
- Slovenija sodeluje v 14 RI (npr. CERN)
- Humanitstika:
  - DARIAH ERIC / DARIAH-SI: Digitalna raziskovalna infrastruktura za umetnost in humanistiko (Digital Research Infrastructure for the Arts and Humanities)
  - **CLARIN ERIC / CLARIN.SI: Infrastruktura za skupne jezikovne vire in tehnologije (Common Language Resources and Technology Infrastructure)**

# 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 dolgotrajnemu in obsežnem hranjenju in dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Nova paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnost in prilagajanja individualnim potrebam

# Namen

- Obstoječa orodja in rešitve dati na voljo v enotni infrastrukturi
- Omogočiti svetovalne in učne dejavnosti, kako orodja in vire prilagoditi specifičnim raziskovalnim potrebam
- Prispevati k standardizaciji virov in orodij

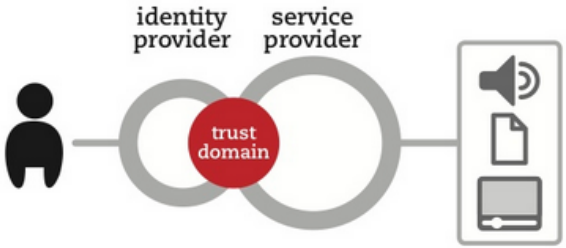
# CLARIN ERIC


- Sedež na Nizozemskem
- 20 držav članic + 4 opazovalke
- Podporno osebje, odbori za vodenje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev

# Kaj ponuja CLARIN?

- Letna konferenca:
  - CLARIN pokrije stroškov 5 udeležencev + avtorjev prispevkov + 1 doktorski študent
- CLARIN Mobility Grants
- Centri znanja (Knowledge Centres):
  - K-Centre for Corpus Linguistics
  - K-Centre for Diachronic Language resources
  - K-Centre for Speech Analysis
  - K-Centre for Terminology Resources and Translation Corpora
  - itd.
- Seznam učnih vsebin s področja digitalne humanistike
- Pregled družin virov (resource families)
- VideoLectures
- Virtual Language Observatory

# Prijava




- Infrastruktura za avtentikacijo in avtorizacijo (AAI)
- Single Sign-On: ločevanje med ponudnikom storitve in ponudnikom identitete
- Za razliko od klasične spletne prijave je tu identiteta uporabnika poznana
- Federacije ponudnikov identitete EduGain: povezuje federacije ponudnikov identitete, da olajša dostop do vsebin, storitev in virov za globalno raziskovalno in izobraževalno skupnost
- Slovenski uporabniki lahko prek EduGain dostopajo do večine servisov CLARIN po EU

# CLARIN.SI

# CLARIN.SI


`http://www.clarin.si/`

- Začetek dela v 2014
- Institut "Jožef Stefan":
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 parterjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 3 raziskovalni inštituti: ZRC SAZU, IJS, INZ
  - 3 društva oz. zavodi: SDJT, Trojina, DDR
  - 2 podjetji: Amebis, Alpineon

# CLARIN.SI storitve

- Podpora dogodkom:
  - Konference „Jezikovne tehnologije in digitalna humanistika“
  - XVIII EURALEX International Congress, Lj., 2018
  - JOTA @ VideoLectures
- CLASSLA: K-Centre for South-Slavic Languages
- Podpora razvoju jezikovnih virov in orodij (za slovenščino):
  - podpora pripravi virov za vključitev v repozitorij
  - večji projekti: prvič v 2018, 30.000 EUR, 7 projektov
- Repositorij: dolgotrajno hranjenje jezikovnih virov (in orodij)
- Dva konkordančnika (spletne analize korpusov)
- Platforma za ročno označevanje korpusov
- Platforma za avtomatično označevanje korpusov
- GitHub & GitLab, pretvorba Word2TEI, . . .

# CLARIN.SI CLASSLA: K-center za južnoslovanske jezike

- Strokovna podpora pri uporabi jezikovnih virov in tehnologij za južnoslovanske jezike.
- Posredovanje informacij prek dokumentacije o razpoložljivih virih in tehnologijah raziskovalcem, študentom, ljubiteljskim znanstvenikom in drugim zainteresiranim posameznikom
- Tehnično podpora pri ustvarjanju, preoblikovanju in objavljanju virov in tehnologij
- Organizacija izobraževanj
- FAQ za slovenščino, hrvaščino, srbščino

# FAQ za slovenščino

## POGOSTA VPRAŠANJA O JEZIKOVNIH VIRIH IN TEHNOLOGIJAH ZA SLOVENŠČINO

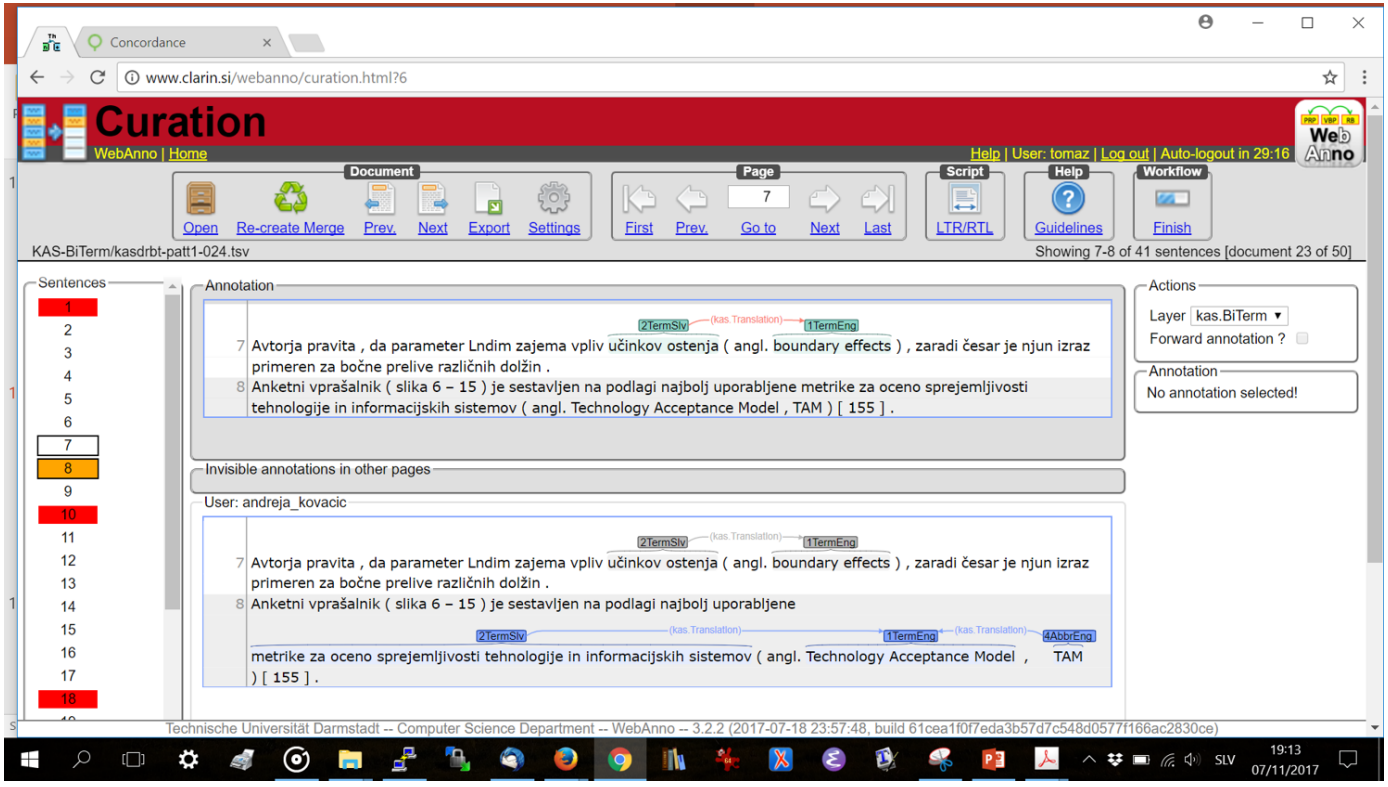
Ta pogosta vprašanja z odgovori (FAQ) so del dokumentacije središča CLASSLA, ki je središče znanja za južnoslovanske jezike v okviru evropske raziskovalne infrastrukture CLARIN. Če opazite manjkajoče ali napačne informacije, prosimo, da nas o tem obvestite v e-poštnem sporočilu z zadevo »FAQ_slovenščina« na naslov helpdesk.classla@clarin.si

Vprašanja v tem razdelku so razdeljena v tri glavne sklope:

1. Spletni jezikovni viri za slovenščino
    1. Kje lahko najdem slovarje slovenskega jezika?
    2. Ali lahko korpuse slovenščine analiziram v spletu?
    3. Katere korpuse slovenščine lahko analiziram v spletu?
    4. Katere označevalne sheme so uporabljene v korpusih slovenščine?
    5. Kje lahko prevzamem vire za slovenščino?
2. Orodja za označevanje slovenskih besedil
    1. Kako lahko izvedem osnovno jezikovno obdelavo slovenskih besedil?
    2. Kako lahko svoja besedila pred obdelavo standardiziram?
    3. Kako lahko v besedilu označim imenske entitete?
    4. Kako lahko skladenjsko razčlenim svoja besedila?
3. Nabor podatkov za učenje označevalnikov za slovenščino
    1. Kje lahko najdem vektorske vložitve besed za slovenščino?

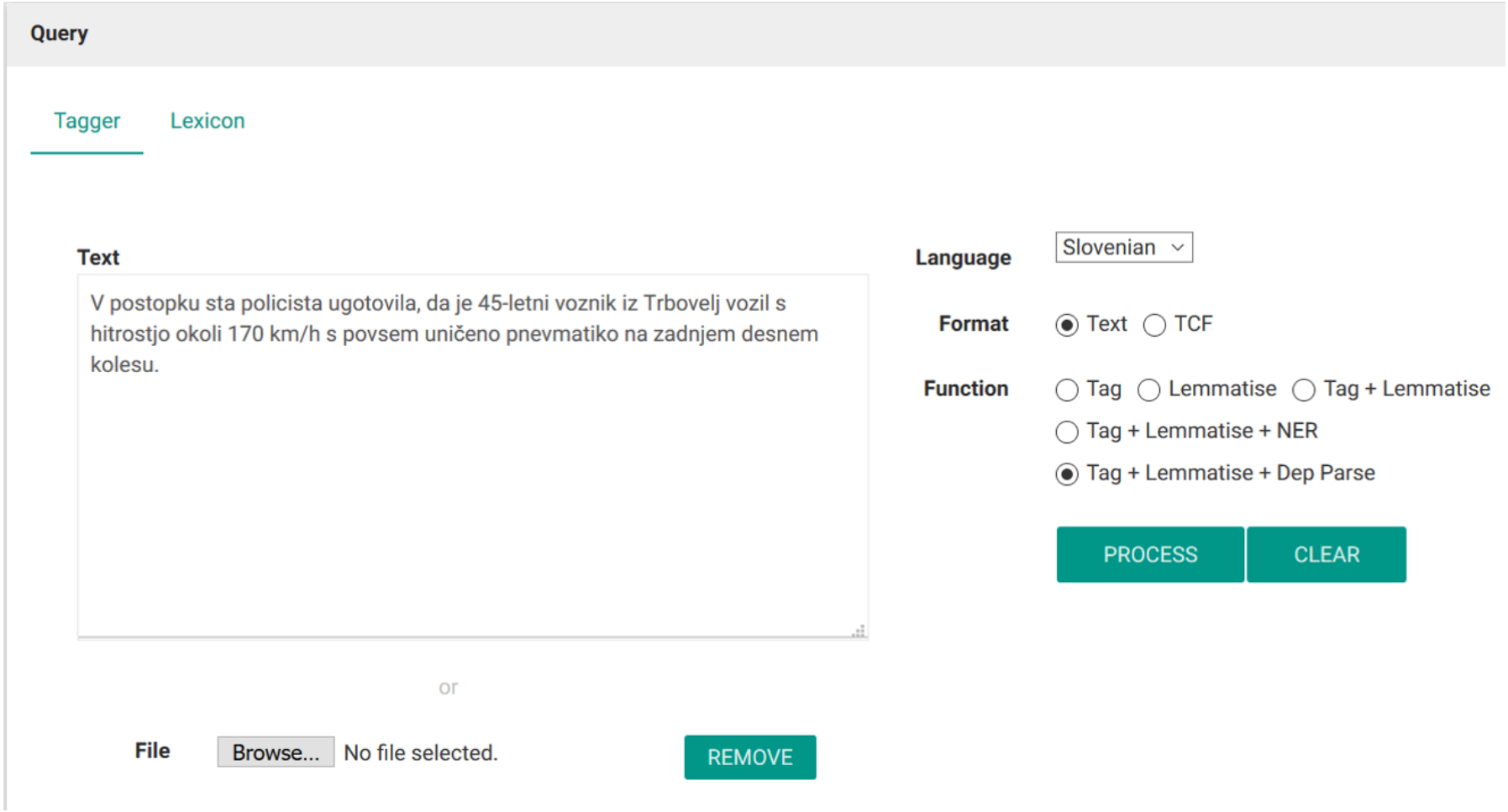
# CLARIN.SI spletne storitve

# WebAnno




- Platforma za ročno označevanje korpusov
- Razvil ga je nemški CLARIN
- Podpira več anotatorjev + uredniški korak
- V CLARIN.SI razvijamo pretvorbo TEI $\rightarrow$ TSV $\rightarrow$ TEI

# Platforma za avtomatsko označevanje besedil




- Slovenščina, hrvaščina, srbščina
- Vnos, datoteka, spletna storitev

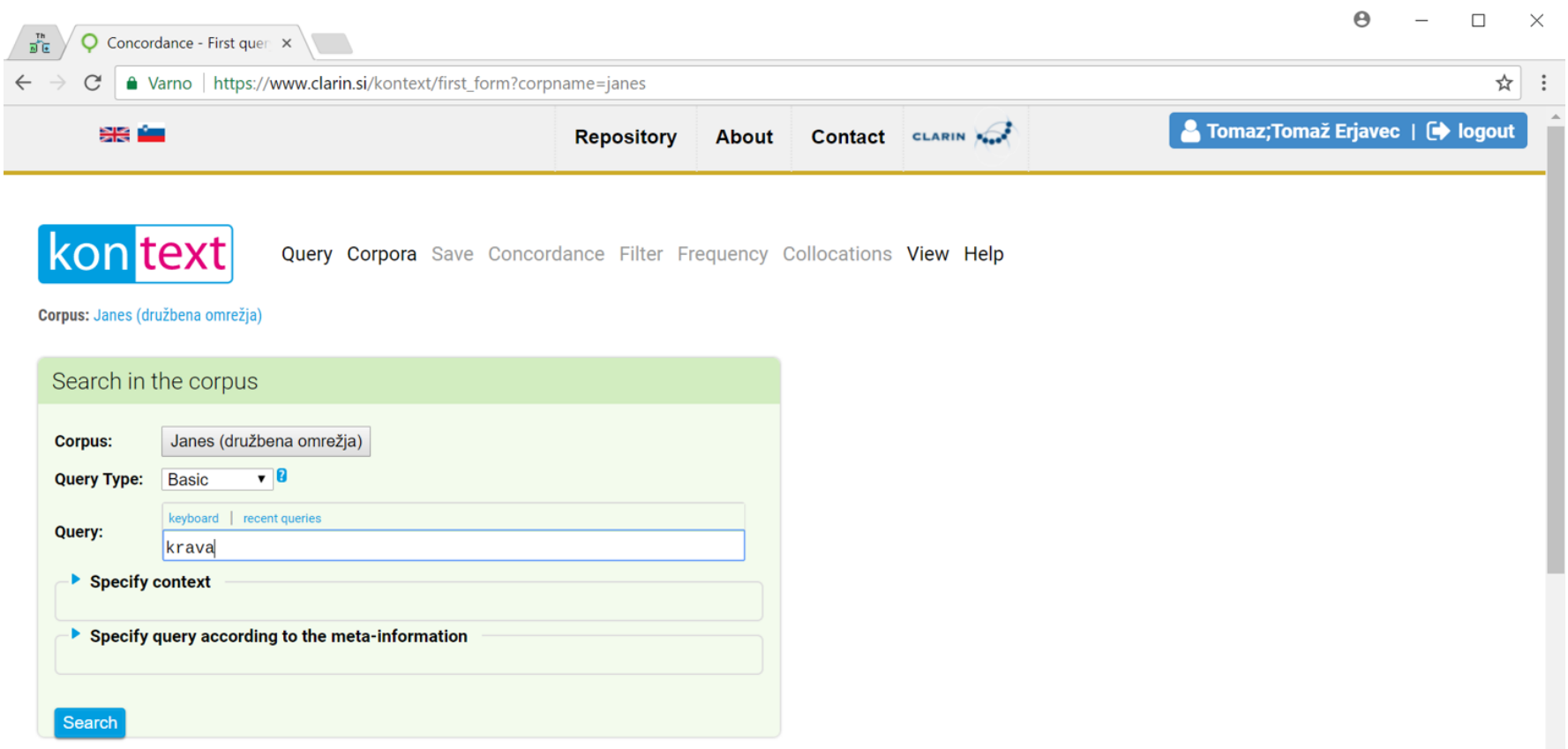
# Rezultat

| | Surface | Tags | Lemma | Dep parse - gov / func |
|---|---|---|---|---|
| **1.** | V | Sl | v | 2 / case |
| **2.** | postopku | Ncmsl | postopek | 3 / nmod |
| **3.** | sta | Va-r3d-n | biti | 0 / root |
| **4.** | policista | Ncmdn | policist | 3 / dobj |
| **5.** | ugotovila | Vmep-dm | ugotoviti | 4 / acl |
| **6.** | , | Z | , | 5 / punct |
| **7.** | da | Cs | da | 13 / mark |
| **8.** | je | Va-r3s-n | biti | 13 / aux |
| **9.** | 45-letni | Agpmsny | 45-leten | 10 / amod |
| **10.** | voznik | Ncmsn | voznik | 13 / nsubj |
| **11.** | iz | Sg | iz | 12 / case |
| **12.** | Trbovelj | Npfpg | Trbovlje | 10 / nmod |

# Konkordančnika

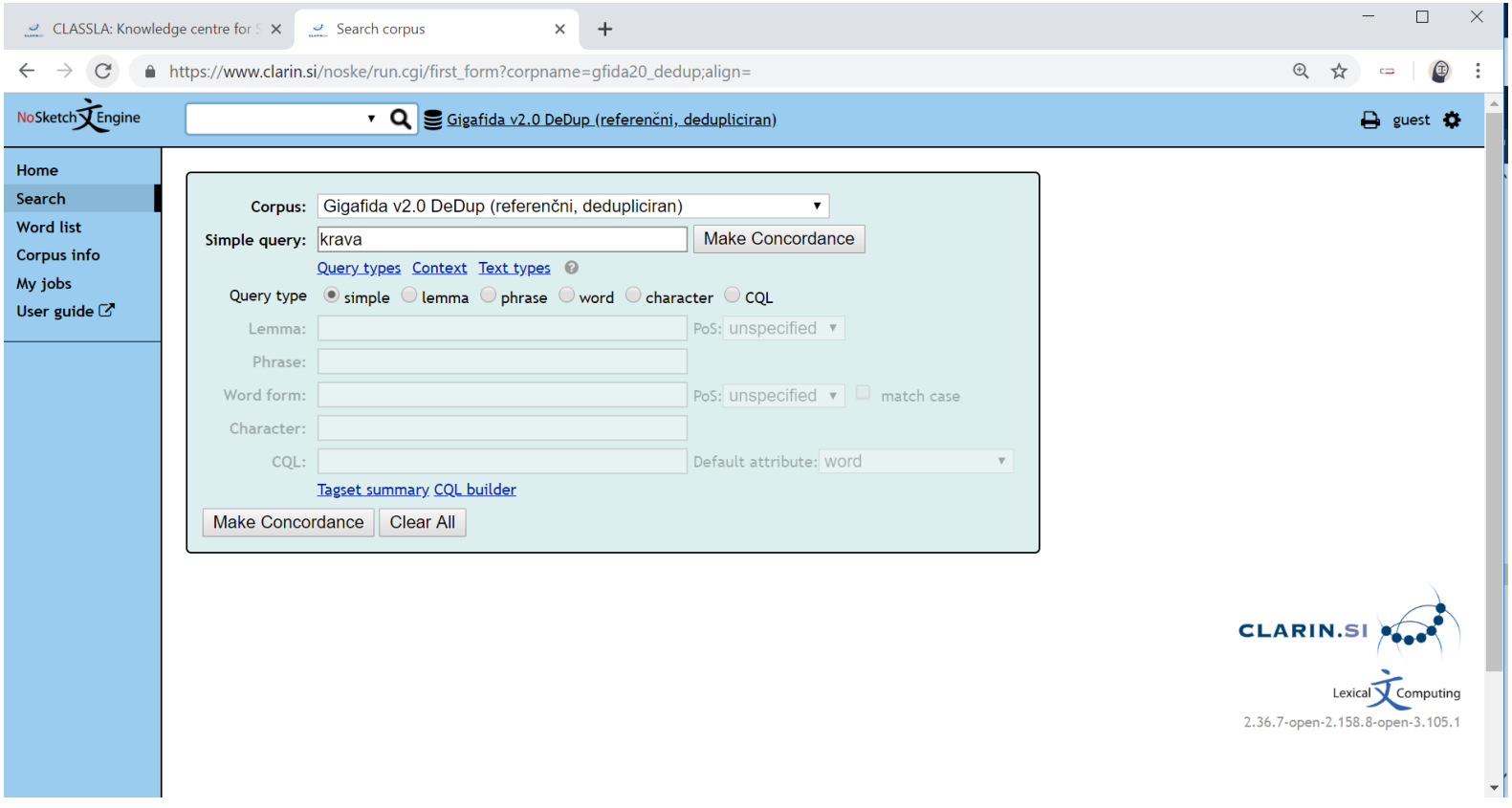
- KonText + noSketch Engine
- Uporabljata isti zaledni program: Manatee
- Delata lahko z velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni:
  - strukture (besedilo, odstavek, termin, . . . )
  - metapodatki (datum, vrsta besedila, spol avtorja, standardnost,. . . )
  - atributi pojavnic (oblikoskladenjska oznaka, lema, normalizirana oblika,. . . )
- Bogat poizvedovalni jezik: CQL
- Raznovrstni izpisi in analize
- RESTful, tj. poizvedbe prek URLjev
- CLARIN.SI noSke & KonText ponujata okoli 50 korpusov v 27 jezikih in s prek 14 milijard besed

# KonText




- Razvil ga je češki CLARIN
- Prijava: shranjene poizvedbe, nastavitve zaslona, podkorpusi
- Nima nekaterih funkcionalnosti noSketch Engine

# noSketch Engine




- Odprtokodna različica Sketch Engine
- Prijava ni niti potrebna niti mogoča

# Repozitorij

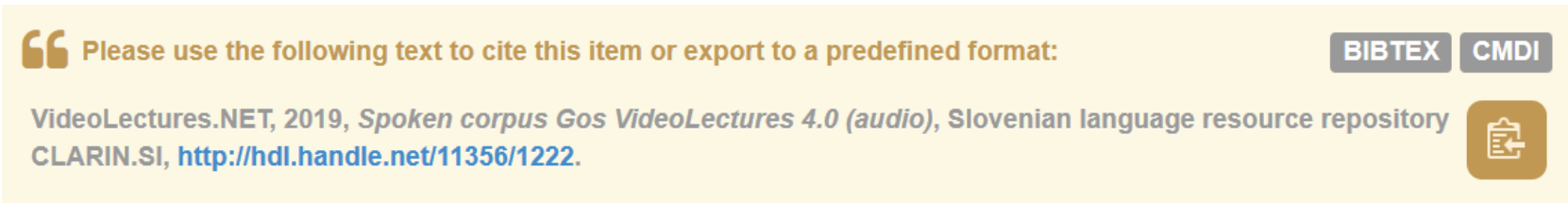
- Zaenkrat najbolj pomembna storitev CLARINa
- Stalna in varna hramba jezikovnih virov (https, Nagios)
- Eksplicitni pogoji uporabe (ToS, licenca), etični kodeks (CoCo)
- Standarden zapis metapodatkov: Component Metadata Infrastructure (CMDI) Dublin Core (DC)
- Žetev metapodatkov (metadata harvesting)
- Večinoma standarden zapis podatkov (XML, TEI)
- Trenutno vsebuje 117 virov

# Programska oprema

- Osnovan na platformi DSpace, namenjena odprtim digitalnim repozitorijem
- DSpace prilagojen za namene CLARIN repozitorijev v češkem CLARIN (ti. DSpace/LINDAT)
- Poleg .si uporablja tudi Češka, Norveška, Poljska, Italija
- Vzdrževanje na GitHub / GitLab
- CLARIN.SI repozitorij CLARIN in DSA certificiran
- Prijava prek AAI
- Žetev metapodatkov: CLARIN, OpenAIRE, re3data itd.

# Stalni identifikatorji

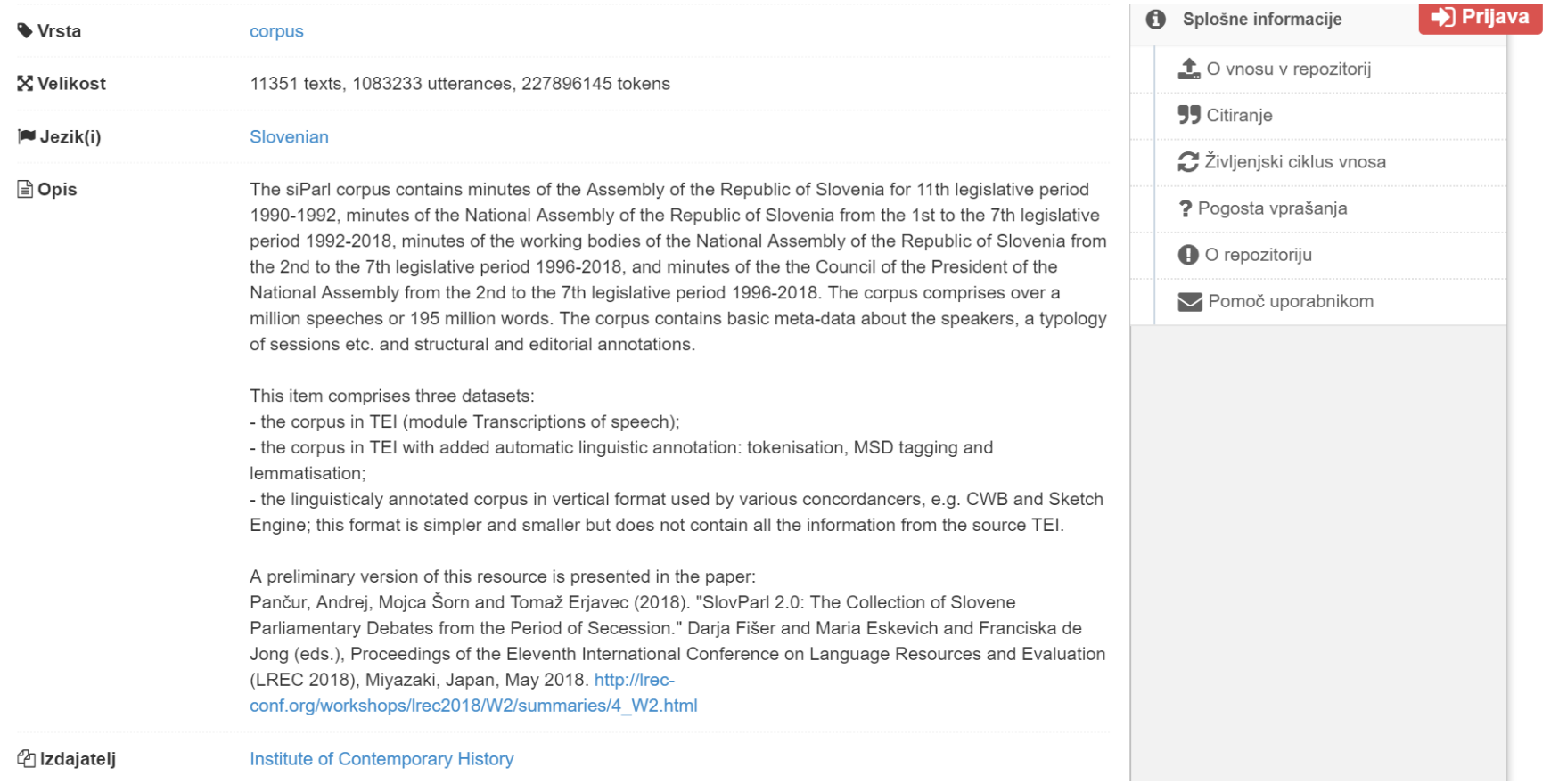
- Problem stabilnih URLjev, ki jih je možno citirati
- Najbolj razširjena rešitev: DOI
- CLARIN uporablja sistem Handle Registracija, dodelitev prefiksa (in plačilo), instalacija lokalne programske opreme
- http://hdl.handle.net/11356/1044 → https://www.clarin.si/repository/xmlui/handle/11356/1044
- Pomembno za pravilno citiranje virov



# Anatomija vnosa, 1


- osnovni metapodatki; citiranje; integracija s storitvami; povezave na projekt, demo & objave
- lokalizacija, orodna vrstica, prijava; iskanje, osnovne informacije o repozitoriju, brskanje; Piwik

# Anatomija vnosa, 2

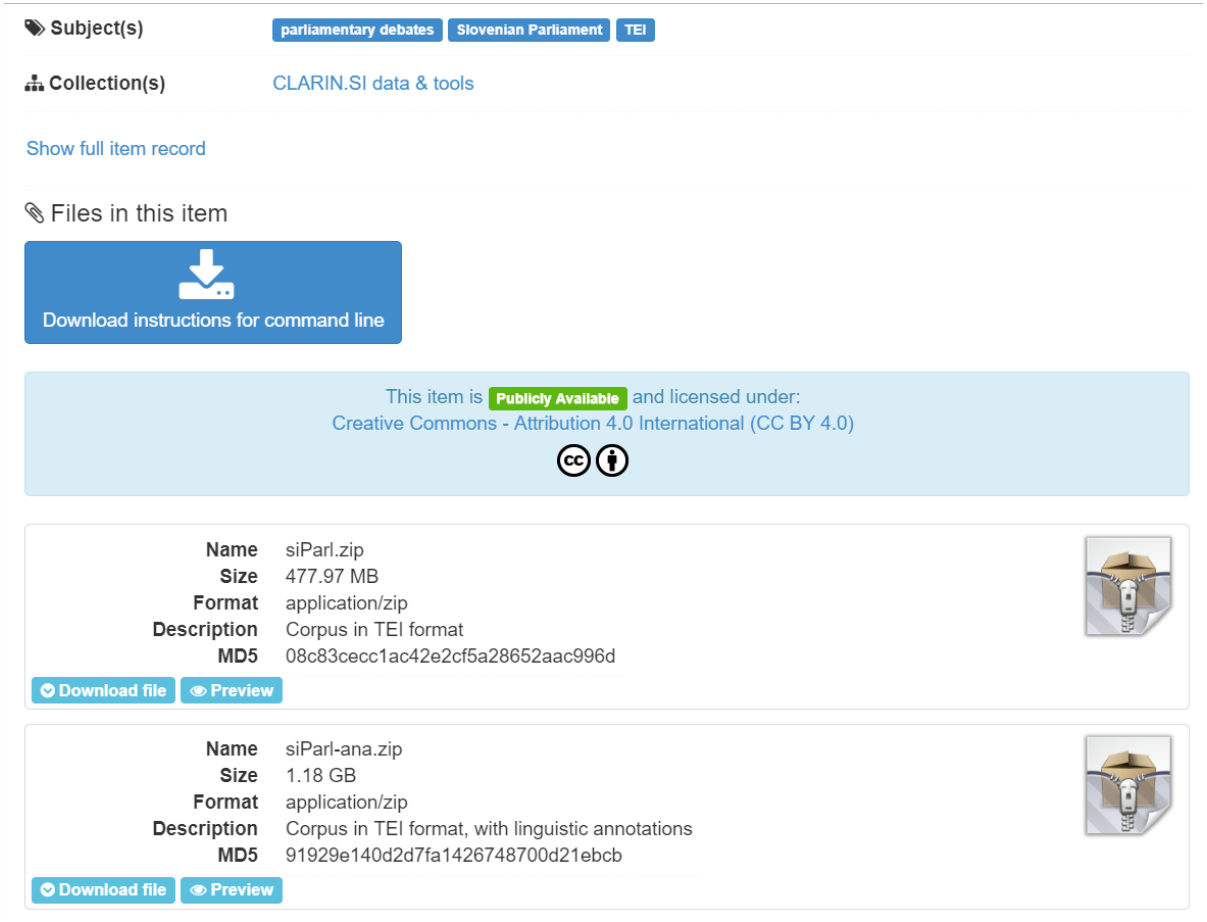


| | |
|---|---|
| Vrsta | corpus |
| Velikost | 11351 texts, 1083233 utterances, 227896145 tokens |
| Jezik(i) | Slovenian |
| Opis | The siParl corpus contains minutes of the Assembly of the Republic of Slovenia for 11th legislative period 1990-1992, minutes of the National Assembly of the Republic of Slovenia from the 1st to the 7th legislative period 1992-2018, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the 2nd to the 7th legislative period 1996-2018, and minutes of the the Council of the President of the National Assembly from the 2nd to the 7th legislative period 1996-2018. The corpus comprises over a million speeches or 195 million words. The corpus contains basic meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations.<br><br>This item comprises three datasets:<br>- the corpus in TEI (module Transcriptions of speech);<br>- the corpus in TEI with added automatic linguistic annotation: tokenisation, MSD tagging and lemmatisation;<br>- the linguisticaly annotated corpus in vertical format used by various concordancers, e.g. CWB and Sketch Engine; this format is simpler and smaller but does not contain all the information from the source TEI.<br><br>A preliminary version of this resource is presented in the paper:<br>Pančur, Andrej, Mojca Šorn and Tomaž Erjavec (2018). "SlovParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession." Darja Fišer and Maria Eskevich and Franciska de Jong (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html |
| Izdajatelj | Institute of Contemporary History |



- vrsta, velikost, jezik, opis, založnik
- več informacih o repozitoriju

# Anatomija vnosa, 3




- ključne besede, vsi metapodatki
- licenca, prevzem podatkov

# Pregled virov

# Učni viri za standardno slovenščino

- Krek, Simon; et al., 2019, Training corpus ssj500k 2.2, http://hdl.handle.net/11356/1210.

  "500,000 tokens manually annotated on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, and lemmatisation. About half of the corpus is also manually annotated with syntactic dependencies, named entities, and verbal multiword expressions. About a quarter of the corpus is annotated with semantic role labels."

- Dobrovoljc, Kaja; et al., 2019, Morphological lexicon Sloleks 2.0, http://hdl.handle.net/11356/1230.

  "Reference morphological lexicon for Slovenian language, developed to be used in NLP applications and language manuals. The lexicon contains approx. 100,000 most frequent Slovenian lemmas, their inflected or derivative word forms and the corresponding grammatical description."

# Korpusi standardne slovenščine

- Logar, Nataša et al., 2013, Written corpus ccGigafida 1.0, http://hdl.handle.net/11356/1035.

  "Paragraph samples from Gigafida with extensive metadata on the source (31,722, 100 million tokens). The corpus is MSD-tagged and lemmatised."

- Logar, Nataša et al., 2013, Written corpus ccKres 1.0, http://hdl.handle.net/11356/1034.

  "The ccKres corpus contains approximately 9% (10 million words) of the Kres corpus, a balanced corpus of Slovene."

- Kadivec, Jože; Robnik-Šikonja, Marko and Vintar, Špela, 2017, ccGigafida ARPA language model 1.0, http://hdl.handle.net/11356/1119.

  "A general language model of contemporary standard Slovenian language that can be used as a language model in statistical machine translation systems."

# Korpusi Janes

- Projekt ARRS "Jezikoslovna analiza nestandardne slovenščine" (2014–2018)
- CMC training corpus Janes-Norm 1.2, CMC training corpus Janes-Tag 2.0
- Twitter corpus Janes-Tweet 1.0, Forum corpus Janes-Forum 1.0, Blog post and comment corpus Janes-Blog 1.0, News comment corpus Janes-News 1.0, Wikipedia talk corpus Janes-Wiki 1.0
- Tweet code-switching corpus Janes-Preklop 1.0, Tweet comma corpus Janes-Vejica 1.0, CMC shortening corpus Janes-Kratko 1.0
- Janes corpus n-grams 1.0
- Dictionary of Twitterese Janes-Dict 1.0

# Jezikovni viri starejše slovenščine IMP

- Erjavec, Tomaž, 2014, Digital library and corpus of historical Slovene IMP 1.1, http://hdl.handle.net/11356/1031.

  "Slovene books and other publications, (658 texts, 45,000 pages, 17,700,000 tokens) from the period 1584-1919. In the corpus each word is marked-up with its modernised form, lemma, and morphosyntactic description (fine grained PoS tag)."

- Erjavec, Tomaž, 2015, Reference corpus of historical Slovene goo300k 1.2, http://hdl.handle.net/11356/1025.

  "Manually annotated reference corpus of historical Slovene. It contains 1,100 pages (about 300,000 tokens) sampled from 89 texts from the period 1584-1899."

- Erjavec, Tomaž, 2014, Lexicon of historical Slovene imp25k 1.1, http://hdl.handle.net/11356/1032.

  "Contains attested and manually verified word forms and their annotations with examples of use. A lexicon entry contains the modern lemma with its part-of-speech and, for archaic words, its gloss (closest modern equivalent(s) or short explanation of their meaning). The lemma is followed by its modern word forms from the corpus (i.e. the complete paradigm of the lemma is not given), and each of these has all its attested historical word forms with examples of usage."

# Zapisniki parlamenta

- Pančur, Andrej; Šorn, Mojca and Erjavec, Tomaž, 2017, Slovenian parliamentary corpus SlovParl 2.0, http://hdl.handle.net/11356/1167.

  "Minutes of the Assembly of the Republic of Slovenia before, during, and after Slovenia became an independent country. Comprises 232 sessions, 58,813 speeches and 10.8 million words; extensive meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations."

- Dobranić, Filip; Ljubešić, Nikola and Erjavec, Tomaž, 2019:
  - ParlaMeter-sl 1.0, http://hdl.handle.net/11356/1208.

    "Minutes of the VIIth mandate of the National Assembly of the Republic of Slovenia. Comprises 41,000,000 tokens. Contains speaker metadata (gender, age, education, party affiliation); transcriptions are MSD tagged, lemmatised, and marked with named entities."
  - ParlaMeter-hr 1.0, http://hdl.handle.net/11356/1209.

    "Minutes of the VIth mandate of the National Assembly of the Republic of Croatia. Comprises 14,000,000 tokens."

- Pančur, Andrej et al., 2019, Slovenian parliamentary corpus siParl 1.0 (1990-2018), http://hdl.handle.net/11356/1236.

  "Minutes of the Assembly of the Republic of Slovenia and comprises over a million speeches or 195 million words. The corpus contains basic meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations."

# Govorni viri

- Zwitter Vitez, Ana et al., 2013, Spoken corpus Gos 1.0, http://hdl.handle.net/11356/1040.

  "Transcripts of 120 hours of speech (one million words) recorded in various situations. Two versions: pronunciation-based spelling and standardized spelling."

- Kačič, Zdravko; et al., 2002, SNABI database for continuous speech recognition 1.2, http://hdl.handle.net/11356/1051.
- Dobrišek, Simon et al., 2017, Speech Database of Spoken Flight Information Enquiries SOFES 1.0, http://hdl.handle.net/11356/1125.
- Verdonik, Darinka; et al., 2019, Spoken corpus Gos VideoLectures 4.0 (transcription), http://hdl.handle.net/11356/1223.

  "A selection of public lectures available through the web portal Videolectures.net. Covers 55 lectures and 22 hours of speech. (180,000 words). Contains speaker meta-data, two hand-produced transcriptions, automatic MSD tagging and lemmatisation and automatically produced word and phone-level alignment with the speech signal."

# Zaključki

# Zaključki

- Namen CLARIN(.SI) je spodbujati raziskave, ki potrebujejo dostop do jezikovnih podatkov
  - digitalna humanistika in družboslovje
  - jezikovne tehnologije
  - vse ostale vede, kjer je jezik pomemben
- Odprt dostop do virov, orodij in storitev
- Avtentikacija AAI, citiranje (handle)
- Prek 50 korpusov na konkordančnikih
- Prek 100 virov v repozitoriju, od tega 88 (tudi) v slovenščini
- Slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN

# Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"

JOTA
FRI UL
2019-05-28