Predstavitev raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI

Tomaž Erjavec

Nacionalni koordinator CLARIN.SI
Odsek za tehnologije znanja, Institut "Jožef Stefan"

Kako bo mogoče? – posvet o digitalni prihodnosti slovenščine
2019-09-27

Raziskovalne infrastrukture

Kaj je RI?

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

Raziskovane infrastrukture ESFRI

- *European Strategy Forum on Research Infrastructures* (ESFRI) je predlagal 15 (2016: 21) RI, nekatere že delujejo kot pravne osebe *European RI Consortium* (ERIC)
- Slovenija sodeluje v 14 ESFRI RI (npr. CERN)
- 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 uporabe in prilagajanja individualnim potrebam

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

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 partnerjev:
  - 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

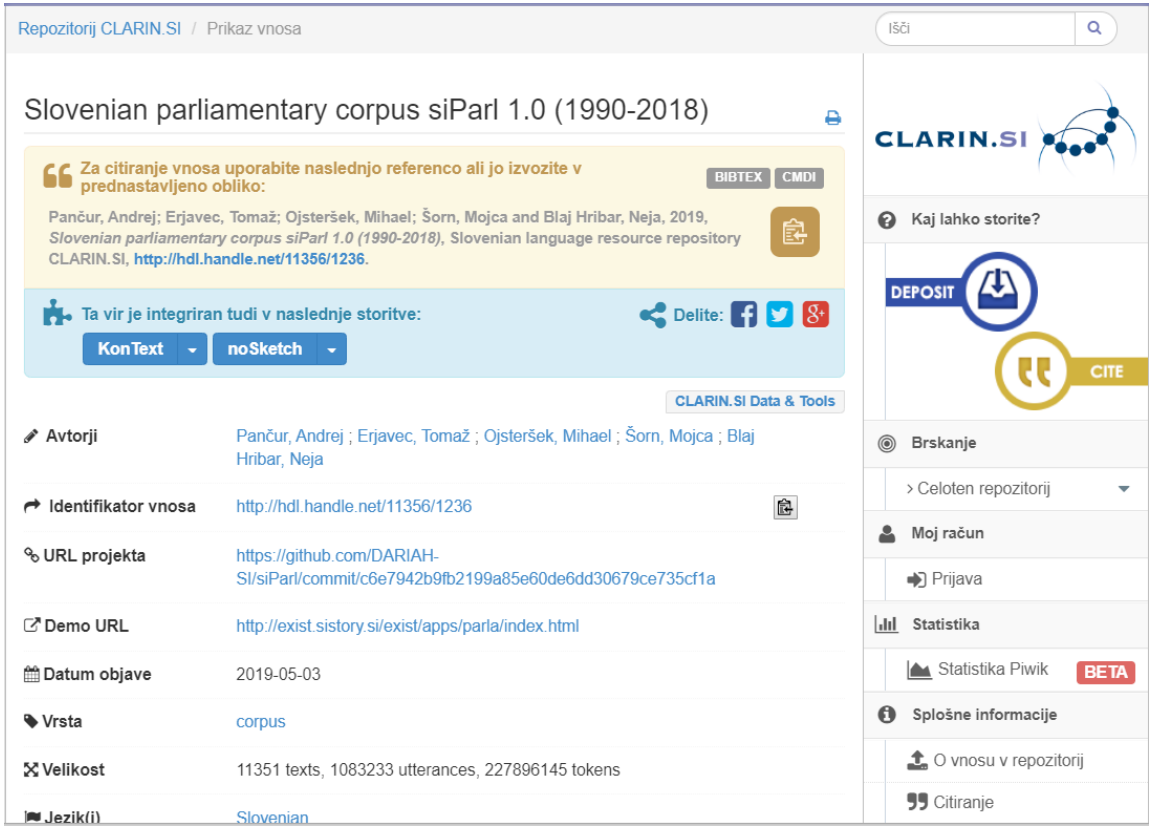
CLARIN.SI storitve

Trije stebri:

- Certificiran repozitorij jezikovnih virov in orodij
  - dolgotrajno hranjenje
  - avtentikacija in avtorizacija
  - stalni identifikatorji
  - eksplicitni pogoji uporabe in licence
  - principi FAIR
- Spletne storitve
  - dva konkordančnika (spletne analize korpusov)
  - orodja za označevanje besedil
  - itd.
- Podpora:
  - financiranje priprave virov za vključitev v repozitorij
  - večji projekti: 30.000 EUR letno, 7 projektov v 2018
  - dogodki: JOTA, JT-DH 2018, EURALEX 2018, TSD 2019, ...
  - center znanja za računalniško obdelavo južnoslovanskih jezikov

Repozitorij

- Najpomembnejša storitev CLARIN.SI
- Danes 129+ jezikovnih virov, od tega 96 slovenskih: korpusi, slovarji, besedišča, modeli, programi
- Velika večina pod eno od licenc Creative Commons



Problemi

- Pomanjkanje kadra
- Premajhna informiranost skupnosti o obstoju CLARIN.SI (posebej še repozitorija)
- Zakonodaja, ki ovira redistribucijo korpusov

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 in brezplačen dostop do virov, orodij in storitev
- Slovenski raziskovalci imajo dostop tudi do storitev CLARIN ERIC in drugih nacionalnih konzorcijev CLARIN

Predstavitev raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI

Tomaž Erjavec

Nacionalni koordinator CLARIN.SI
Odsek za tehnologije znanja, Institut "Jožef Stefan"

Kako bo mogoče? – posvet o digitalni prihodnosti slovenščine
2019-09-27