

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut „Jožef Stefan“

FRI, 10. 11. 2017

Pregled predavanja

1. Uvod
2. Raziskovalna infrastruktura CLARIN(.SI)
3. Storitve CLARIN.SI

Uvod

- Jezikovne tehnologije
 - glavna paradigma: nadzorovano strojno učenje
 - programi so večinoma jezikovno neodvisni
 - potrebujejo pa učne (ročno označene) jezikovne vire
 - + testne podatke
- Empirično podprte jezikoslovne raziskave:
 - temeljijo na realnih (in po možnosti označenih) besedilih
- Označene jezikovne vire potrebujemo za vsak jezik posebej
- Kje lahko dobimo take vire za slovenščino?

Jezikovni viri

1. Korpusi:

- enovito kodirana in dokumentirana zbirka besedil
- besedila (i)zbrana po vnaprej določenih kriterijih
- označena
- referenčni/specializirani; eno/večjezični; pisni/govorni

2. Leksikoni:

- besedišče jezika
- besede / fraze
- oblikoslovje, skladnja, pomen, povezave, prevodi

3. Modeli:

- podatki za nek program, ki mu omogoči označevanje besedil v nekem jeziku za neko raven označevanja
- npr. CLOG model za lematizacijo slovenščine; Moses model za prevajanje slovenščina → angleščina

Ponovna uporaba

- Klasični pristop:
 - za vsako raziskavo posebej izdelati jezikovne vire
 - viri nedostopni drugim raziskovalcem
- 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
- Zelo močan trend v EU (H2020), tudi v Sloveniji
- Problemi pri omogočanju odprtega dostopa do jezikovnih virov:
 - avtorske pravice nad besedili
 - varovanje zasebnosti (tudi pravica do pozabe)
 - pogoji uporabe spletnih portalov (npr. Twitter)

Raziskovalne infrastrukture

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

[| A to Z](#) | [Sitemap](#) | [About this site](#) | [Legal notice](#) | [Cookies](#) | [Contact](#) | [Search](#) | English (en) ▾



European Commission

RESEARCH & INNOVATION

Infrastructures

[European Commission](#) > [Research & Innovation](#) > [Research infrastructures](#) > [ESFRI](#)



Research Infrastructures

- HOME
- WHAT ARE RIs ?
- MAPS of RIs
- THE EUROPEAN LANDSCAPE
- EU FINANCIAL SUPPORT
- ERIC-LEGAL FRAMEWORK
- SYNERGIES - EU INITIATIVES
- INTERNATIONAL COOPERATION



The ESFRI Roadmap 2016

The [ESFRI Roadmap](#) 2016 identifies the new Research Infrastructures (RI) of pan-European interest corresponding to the long term needs of the European research communities, covering all scientific areas, regardless of possible location.



The 2016 Roadmap consists of 21 ESFRI Projects with a high degree of maturity - including 6 new Projects - and 29 ESFRI Landmarks - RIs that reached the implementation phase by the end of 2015.

The ESFRI Roadmap 2016 was launched on 10 March 2016, in Amsterdam. The event was organized under the [Dutch Presidency](#) by the Royal Netherlands Academy of Arts and Sciences (KNAW) in close cooperation with ESFRI, the European Commission and the Dutch Ministry of Education, Culture and Science. Discussions focussed on strategic roadmapping, long-term sustainability and the socio-economic impact of research infrastructures.

See Event [Agenda](#) and [Live Stream](#)

ESFRI

Highlights



An on-line map to locate the ESFRI infrastructures and their partner facilities

About 400 facilities are part of these distributed

Raziskovalne infrastrukture

- 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)
- Humanistika:
 - 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)



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, obsežnemu in lahko dostopnemu hranjenju jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske dediščine
- Nova paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnost in prilagajanja individualnim potrebam

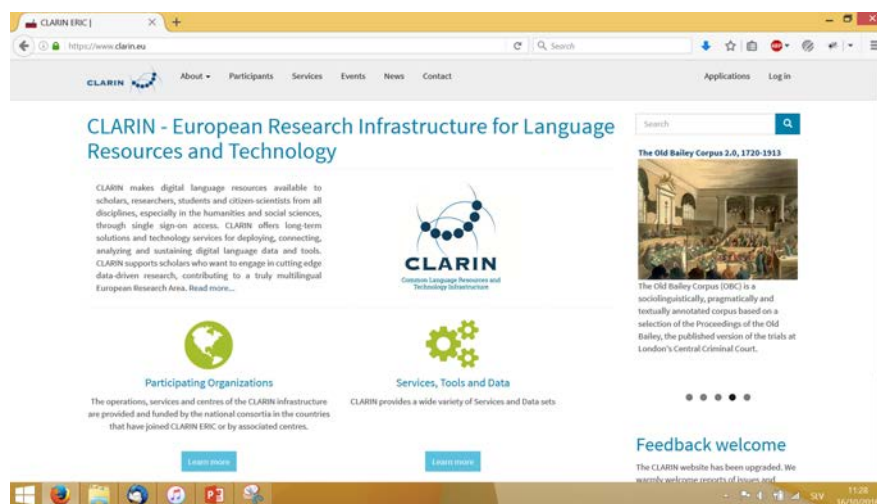


Namen:

- 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

- 19 držav članic + 2 opazovalki
- Sedež na Nizozemskem
- Podporno osebje, odbori za vodenje, delovne skupine
- Letne konference
- Virtual Language Observatory
- Večina dela se odvija v okviru nacionalnih konzorcijev



CLARIN.SI



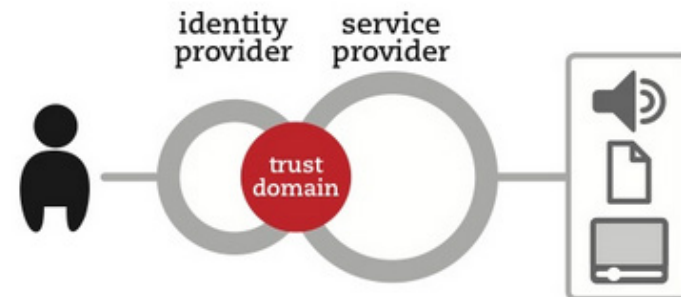
- Začetek dela 2014
- Sedež IJS:
 - Odsek za tehnologije znanja
 - Laboratorij za umetni inteligenco
 - Center za mrežno infrastrukturo
- Organiziran kot konzorcij 12 partnerjev:
 - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
 - 3 inštituti: ZRC SAZU, IJS, INZ
 - 3 društva: SDJT, DDR, Trojina
 - 2 podjetji: Amebis, Alpineon

CLARIN.SI



- **Podpora dogodkom:**
 - Seminar ReLDI “Empirični podatki v jezikoslovju: Od zasnove raziskave do analize podatkov” (FF, 21.–23. 6. 2017)
 - Redne konference „Jezikovne tehnologije in digitalna humanistika“ (... , 29. 9.–1. 10. 2016, 20.–21.9. 2018, ...)
 - 4th Conference on CMC and Social Media Corpora for the Humanities (FF, 27.–28. 9. 2016)
- **Spletne storitve:**
 - konkordančniki (analize korpusov)
 - ročno označevanje (izdelava učnih korpusov)
 - delotoki (avtomatizirano označevanje)
- **Repozitorij:**
 - Dolgotrajno hranjenje jezikovnih virov in orodij

Prijava

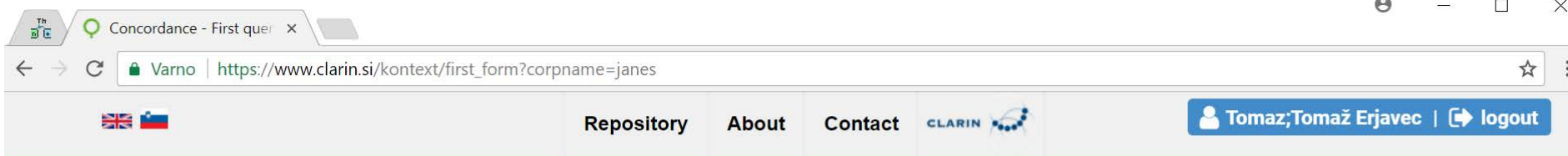


- Infrastruktura za avtentikacijo in avtorizacijo (AAI)
- Single Sign-On: ločevanje med ponudnikom storitve in ponudnikom identitete
- Za razliko od klasične spletne prijave je tu identiteta uporabnika poznana
- Federacije ponudnikov identitete
- EduGain: povezuje federacije ponudnikov identitete, da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

Konkordančniki

- KonText + noSketch Engine
- oba uporabljata isto zaledje: Manatee
- delujeta lahko z velikimi korpusi (več milijard besed)
- korpusi so lahko bogato označeni:
 - strukture in metapodatki (datum objave, vrsta besedila, spol avtorja, standardnost besedila, ...)
 - atributi besed (oblikoskladenjska oznaka, lema, normalizirana oblika, ...)
- bogat poizvedovalni jezik CQL
- nastavitve pogleda
- raznovrstni izpisi in analize
- CLARIN.SI trenutno ponuja okoli 50 korpusov

KonText (CLARIN-CZ)



Concordance - First quer

Varno | https://www.clarin.si/kontext/first_form?corpname=janes

Repository About Contact CLARIN

Tomaz;Tomaž Erjavec | logout



Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: [Janes \(družbena omrežja\)](#)

Search in the corpus

Corpus:

Query Type: ?

Query: [keyboard](#) | [recent queries](#)

▶ Specify context

▶ Specify query according to the meta-information

Search

Konkordance

The screenshot shows the KonText web interface. At the top, there is a navigation bar with links for Repository, About, and Contact, along with a user profile for Tomaz;Tomaž Erjavec and a logout button. The main header features the KonText logo and a menu with options like Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, and Help.

The search results section displays the following information:

- Corpus: Janes (družbena omrežja) | Query: krava (6,958 hits)
- Hits: 6,958 | i.p.m.: 27.51 (related to the whole "janes") | ARF: 3,261.05 | Result is sorted
- Line selection: simple | Display options

The results are presented in a table with columns for document ID, context snippet, and the search term 'krava'. The search term is highlighted in red in the original image.

Document ID	Context Snippet	Search Term
wiki,comment,slv,negative,T...	prekmurščini, ker za [per] je internet "indijska sveta	krava
wiki,comment,slv,negative,T...	spet ne smemo. V Novi vasi so navrhu kosti	krav
wiki,comment,slv,negative,T...	o nekem bivšem političnem veljaku, ki je skušal navaditi	krave
wiki,comment,slv,negative,T...	, da bi se prehranjevale z oblanci. In neumne	krave
wiki,comment,slv,neutral,T1...	lahko preživi (vsaj nekaj časa), tudi tista	krava
wiki,comment,slv,negative,T...	Kot začetnik sem pri opisovanju svoje vasi imel rdečo besedo	krava
wiki,comment,slv,negative,T...	v modro. Pa je začel nastajati članek Domače govedo	Krava
wiki,comment,slv,negative,T...	Bil sem ponosen, da sem imel tudi lastno fotografijo	krav
wiki,comment,slv,negative,T...	ponosen, da sem imel tudi lastno fotografijo krav.	Krave
wiki,comment,slv,neutral,T2...	[per], (nekoč so rekli, da če skupaj	krave
wiki,comment,slv,negative,T...	knjigi rekordov. Pa čeprav bi bila prva. thumbSlika	krave
wiki,comment,slv,negative,T...	na govedo, a na sliki po mojem ni cikasta	krava
wiki,comment,slv,negative,T...	majhnimi otroci (po domače rečeno, kot da smo	krave
wiki,comment,slv,positive,T...	ker je [per] vse članke ožigosal kot junca, ali	kravo

The bottom of the screenshot shows a Windows taskbar with various application icons and a system tray displaying the time as 16:22 on 07/11/2017.

WebAnno (CLARIN-DE)

Concordance

www.clarin.si/webanno/curation.html?6

Curation

WebAnno | Home

Help | User: tomaz | Log out | Auto-logout in 29:16

Document: Open, Re-create, Merge, Prev., Next, Export, Settings

Page: First, Prev., Go to (7), Next, Last

Script: LTR/RTL

Help: Guidelines

Workflow: Finish

KAS-BiTerm/kasdrbt-patt1-024.tsv

Showing 7-8 of 41 sentences [document 23 of 50]

Sentences

Annotation

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

Invisible annotations in other pages

User: andreja_kovacic

7 Avtorja pravita , da parameter Lndim zajema vpliv učinkov ostenja (angl. boundary effects) , zaradi česar je njun izraz primeren za bočne prelive različnih dolžin .

8 Anketni vprašalnik (slika 6 – 15) je sestavljen na podlagi najbolj uporabljene metrike za oceno sprejemljivosti tehnologije in informacijskih sistemov (angl. Technology Acceptance Model , TAM) [155] .

Actions

Layer kas.BiTerm

Forward annotation ?

Annotation

No annotation selected!

Technische Universität Darmstadt -- Computer Science Department -- WebAnno -- 3.2.2 (2017-07-18 23:57:48, build 61cea1f0f7eda3b57d7c548d0577f166ac2830ce)

19:13 07/11/2017

WebLicht (CLARIN-DE)

The screenshot shows the WebLicht web interface in a browser window. The address bar displays the URL <https://weblight.sfs.uni-tuebingen.de/weblight/>. The interface includes a navigation bar with "Main Page", "Chain 1 x", and "+ New Chain" buttons. A "View Tool List" button and a "HELPDESK" button are also visible.

Below the navigation bar, there is a section for tool status selection: "Show tools with status: dev development production superseded withdrawn".

A "Next Choices" section is present, with a note: "Next Choices (Double-click on an icon to add it to the chain)".

The main area is titled "Input and Chain Selection". It contains a text input field with the text "Prazna stanovanja v Celovških dvorih." and a "Run Tools" button. To the right of the input field are "Clear Results" and "Download chain" buttons.

Below the input field, there are three tool cards:

Prazna stanovan [Plain Te]	SFS: To TCF Converter	JSI: ReLDI tag+lemma
Prazna stanovanja v Celovških dvorih.	Language: Slovenian Document Type: TCF TCF Version: 0.4 Text	Sentences Lemmas Tokens Part of Speech: mte-v5-sl

At the bottom of the interface, a status bar indicates "Done running tools." The Windows taskbar at the bottom shows the system time as 19:17 on 07/11/2017.

Nacionalni repozitoriji

- Zaenkrat najbolj pomembna storitev CLARIN-a
- Stalna in varna hramba jezikovnih virov ([https](https://nagios.org/), Nagios)
- Eksplicitni pogoji uporabe (ToS, licenca)
- Etični kodeks (CoCo)
- Standarden zapis metapodatkov:
 - Component Metadata Infrastructure (CMDI)
 - Dublin Core (DC)
- Izvoz metapodatkov (metadata harvesting)

Repozitorij CLARIN.SI

- Osnovan na platformi DSpace, namenjena odprtim digitalnim repozitorijem
- DSpace prilagojen za namene CLARIN repozitorijev v češkem CLARIN (t.i. DSpace/LINDAT)
- Poleg Slovenije ga uporablja tudi Češka, Norveška, Poljska, Italija, ...
- Vzdrževanje Dspace/LINDAT na GitHub, veja za CLARIN.SI na GitLab@IJS



- Ceritificiran B center CLARINa
- Certifikat DSA
- Ponudnik metapodatkov prek OAI-MHP
- Zaveden v registru re3data
- Povezan v sledenje rezultatom EU projektom OpenAIRE



Stalni identifikatorji in citiranje virov

- Problem stabilnih identifikatorjev / URLjev, ki jih je možno dolgoročno citirati
- Najbolj razširjena rešitev: DOI
- CLARIN uporablja sistem Handle
- **<http://hdl.handle.net/11356/1044>** → <https://www.clarin.si/repository/xmlui/handle/11356/1044>

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

BIBTEX

CMDI

Verdonik, Darinka; Potočnik, Tomaž; Sepesy Maučec, Mirjam and Erjavec, Tomaž, 2017, *Spoken corpus Gos VideoLectures 2.0 (transcription)*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1158>.



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
- Kjer je potrebna avtentikacija, se uporablja AAI
- Stabilno citiranje skozi sistem Handle