Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec[†], Jan Jona Javoršek[‡], Simon Krek[♠]

[†] Odsek za tehnologije znanja
[‡] Center za mrežno infrastrukturo
[♠] Laboratorij za umetno inteligenco
Institut »Jožef Stefan«
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si, jona.javorsek@ijs.si, simon.krek@ijs.si

Povzetek

V prispevku predstavimo slovensko jezikoslovno raziskovalno infrastrukturo CLARIN.SI, katere dolgoročni namen je, da v povezavi z evropsko infrastrukturo CLARIN ERIC spodbuja raziskave na področju humanistike in družboslovja s tem, da omogoči raziskovalcem enovit avtoriziran dostop do platforme, ki integrira jezikovne vire slovenskega jezika in napredna orodja za obdelavo slovenščine. Prispevek predstavi evropsko infrastrukturo CLARIN in njena temeljna načela ter povzame dosedanjo zgodovino vzpostavitve CLARIN.SI, nato pa podrobneje obdela trenutno stanje izgradnje te slovenske infrastrukture s poudarkom na repozitoriju jezikovnih virov in jezikoslovnih storitvah in orodjih.

The research infrastructure CLARIN.SI

The paper introduces the Slovene research infrastructure CLARIN.SI, whose long term objective is, in connection with the European research infrastructure CLARIN ERIC, to facilitate research in the humanities and social sciences by enabling researchers a uniform and authorised access to its platform, which will integrate Slovene language resources and advanced tools for processing of Slovene. The paper introduces CLARIN ERIC and its mission and summarises the history of the establishment of CLARIN.SI. It then discusses the current state of its development with a focus on the repository of language resources and on linguistic services and tools.

1. Uvod

CLARIN[1] (Váradi in dr., 2008) je ena izmed evropskih raziskovalnih infrastruktur, ki jih je izbral ESFRI, Evropski strateški forum o raziskovalnih infrastrukturah za Program evropskih raziskovalnih infrastruktur. CLARIN je distribuirana podatkovna infrastruktura, ki vključuje predvsem evropske univerze in raziskovalne inštitute. Od 2012 je CLARIN prijavljen kot evropska pravna oseba (CLARIN ERIC, European Research Infrastructure Consortium) in ima trenutno osem držav članic (Avstrija, Bolgarija, Češka, Nemčija, Danska, Estonija, Nizozemska in Poljska), deveta članica je meddržavno telo »Dutch Language Union«, članici pa bosta predvidoma kmalu postali tudi Norveška in Portugalska.

Kot piše na spletnih straneh CLARIN ERIC,[2] je dolgoročni namen te raziskovalne infrastrukture, da spodbuja raziskave na področju humanistike in družboslovja tako, da omogoči raziskovalcem enovit avtoriziran dostop do distribuirane platforme, ki integrira jezikovne vire in napredna orodja na evropski ravni. Ta vizija temelji na naslednjih stebrih:

1. ***Pokritje:*** V perspektivi naj bi vsak raziskovalec v humanistiki in družboslovju v EU in pridruženih članicah imel z enotnim overjanjem neposreden dostop do vseh zbirk digitalnih podatkov, ki vsebujejo na jeziku temelječa gradiva in so last oz. so dane v dostop s strani javnih ustanov.
2. ***Pravo:*** Za raziskave naj bi pri dostopu do podatkov ne bilo omejitev, razen tistih, ki izvirajo iz zaupnosti podatkov, pravice do zasebnosti ali etičnih zadržkov. Pravice in legitimni interesi lastnikov podatkov morajo biti zaščiteni.
3. ***Integracija podatkov:*** Iskanje po metapodatkih in vsebinah naj bi raziskovalcem omogočilo, da najdejo želene podatke. Lahko bodo gradili virtualne zbirke podatkov, ki prihajajo iz različnih virov in držav, in jih uporabljali, kot da bi bili vsi na istem mestu in z enakimi standardi zapisa.
4. ***Integracija storitev:*** Dodatno naj bi imeli raziskovalci tudi dostop do naprednih jezikovnotehnoloških storitev v obliki spletnih storitev, ki bi jim omogočili označevanje, raziskovanje, izkoriščanje, izboljšanje, upravljanje in vizualizacijo podatkov za podporo raziskovanju. Spletne storitve naj bi delovale na podatkih iz raznovrstnih virov, mogoče bi jih bilo sestavljati v kompleksne verige in strukture za izvedbo zahtevnih operacij.
5. ***Hramba:*** Rezultate raziskovalnih projektov in rezultate, dobljene z uporabo storitev, naj bi bilo možno shraniti kot nove podatkovne zbirke, tako da bi jih lahko uporabili tudi drugi raziskovalci. Podatki in rezultati naj bi bili trajnostno hranjeni in opremljeni s trajnimi identifikatorji, tako da bi bilo do njih možno dostopati za namen repliciranja rezultatov ali za izvajanje novih raziskav. Dodatno naj bi obstajale tudi trajne povezave do publikacij, ki uporabljajo ali dokumentirajo te vire.
6. ***Dostop:*** Raziskovalci naj bi razumeli in uporabljali infrastrukturo CLARIN brez tehničnih zadreg.
7. ***Brez meja:*** Infrastruktura CLARIN naj bi bila umeščena v globalno raziskovalno krajino in naj bi aktivno spodbujala preseganje meja med znanstvenimi področji, drugimi infrastrukturami, državami in kontinenti, kot tudi preseganje meja med akademskim in poslovnim svetom.

[1] http://www.clarin.eu

[2] http://www.clarin.eu/content/mission

Ta zelo ambiciozen načrt se je začel izvajati že leta 2008, do njegove uresničitve pa bo minilo še dosti časa.

V prispevku bomo predstavili, kako je z infrastrukturo CLARIN v Sloveniji, kjer se bomo navezali tudi na druga slovenska vozlišča evropskih humanističnih in družboslovnih raziskovalnih infrastruktur in na delo infrastrukture CLARIN v drugih evropskih državah. Prispevek v razdelku 2 obravnava zgodovino in ureditev infrastrukture CLARIN v Sloveniji, v razdelku 3 delo na vzpostavitvi repozitorija jezikovnih virov, v razdelku 4 spletne storitve, v razdelku 5 pa podamo nekaj zaključkov.

2. CLARIN v Sloveniji

Vlada RS je leta 2011 sprejela načrt razvoja slovenskih infrastruktur ESFRI, ki za humanistiko in družboslovje predvideva vzpostavitev slovenskih infrastruktur za DARIAH (Digital Research Infrastructure for the Arts and Humanities), ki je namenjena spodbujanju digitalno podprtih raziskav in poučevanja v humanističnih vedah in umetnosti, za CESSDA (Consortium of European Social Science Data Archives), ki opravlja podobno nalogo za družboslovje ter za CLARIN. Slovenski DARIAH in CESSDA sta bili ustanovljeni kmalu po sprejetju načrta, prejeli sta tudi financiranje in lahko pokažeta konkretne rezultate. Na Inštitutu za novejšo slovensko zgodovino (INZ) so v sodelovanju z Znanstvenoraziskovalnim centrom Slovenske akademije znanosti in umetnosti (ZRC SAZU) postavili spletno infrastrukturo SI-DIH,[3] ki omogoča iskanje podatkov po različnih repozitorijih oziroma arhivih institucij ali društev v humanistiki in umetnosti. Na Fakulteti za družbene vede Univerze v Ljubljani pa so vzpostavili spletno infrastrukturo ADP[4] (Arhiv družboslovnih podatkov), ki hrani zbirko podatkov, zanimivih za družboslovne analize, s poudarkom na problemih, povezanih s slovensko družbo.

Za razliko od DARIAH in CESSDA Slovenija ni bila vključena v prvo, pilotno fazo vzpostavljanja evropske infrastrukture CLARIN (2008–2011), zato je tudi realizacija vzpostavljanja infrastrukture v Sloveniji potekala zelo počasi, saj je minimalno financiranje steklo šele konec 2013 z Institutom »Jožef Stefan« (IJS) kot sedežem infrastrukture, pri čemer si upravljanje delita Odsek za tehnologije znanja E8 in Laboratorij za umetno inteligenco E3.

Začetek financiranja slovenske infrastrukture CLARIN je sovpadel s koncem velikega slovenskega projekta Sporazumevanje v slovenskem jeziku (SSJ), v okviru katerega je bilo v petih letih trajanja projekta zgrajenih večje število temeljnih jezikovnih virov in storitev za slovenski jezik (Arhar Holdt in dr., 2012; Krek in dr. 2012; Logar Berginc in dr., 2009). Spletišče projekta,[5] na katerem je možno uporabljati spletne storitve in prevzeti odprte jezikovne vire, je gostovalo pri podjetju Amebis, d.o.o., vendar je ob koncu projekta usahnilo financiranje za vzdrževanje strojne in programske opreme, kar je postavilo pod vprašaj nadaljnjo usodo dostopa do rezultatov projekta. Zato smo kot urgentno prvo nalogo infrastrukture postavili prenos spletišča na strežnike IJS, kar je vsebovalo nakup razmeroma zahtevne strojne in programske opreme, ki obsega 3 medsebojno povezane strežnike, od tega enega pod operacijskim sistemom Windows, dva pa GNU/Linux, ter prenos in namestitev programske opreme projekta. Čeprav navzven ni opaziti, razen hitrejšega delovanja, nobene razlike, je tako od začetka 2014 spletišče postavljeno na IJS, kjer se bo tudi naprej vzdrževalo.

V 2014 smo se tudi lotili vzpostavljanja formalnega statusa slovenske infrastrukture CLARIN, ki smo jo poimenovali CLARIN.SI. Sestanki na Ministrstvu za izobraževanje, znanost in šport Republike Slovenije ter s potencialnimi zainteresiranimi inštitucijami v Sloveniji so obrodili dobre rezultate: v začetku junija 2014 je devet partnerjev podpisalo Sporazum o ustanovitvi konzorcija CLARIN.SI. Konzorcij vključuje vse večje javne institucije kot tudi podjetja in društva, ki se ukvarjajo z jezikoslovjem in jezikovnimi tehnologijami v Sloveniji: Alpineon d. o. o.; Amebis, d. o. o.; Institut "Jožef Stefan"; Slovensko društvo za jezikovne tehnologije; Trojino, zavod za uporabno slovenistiko; Univerzo v Ljubljani; Univerzo v Mariboru; Univerzo na Primorskem in ZRC SAZU. S sporazumom je bil ustanovljen upravni odbor CLARIN.SI, v katerem ima vsaka članica enega zastopnika z namestniki in en glas pri glasovanju, s katerim se odloča o delovanju konzorcija. Upravni odbor je doslej imel en sestanek, na dopisnem glasovanju pa se je odločil, da med partnerje sprejme še INZ in Društvo za domače raziskave, snovalce in razvijalce spletnega slovarja Razvezani jezik.

Z vzpostavitvijo konzorcija je omogočeno, da Slovenija lahko zaprosi za včlanjenje v CLARIN ERIC in tako enakopravno sodeluje v delu evropskega konzorcija in izkorišča ugodnosti, ki jih nudi članstvo, npr. financiranje medsebojnih obiskov, sodelovanje v letnih srečanjih itd. Pogoj za vključitev je poleg zagotovitve tehničnih in pravnih pogojev tudi redno letno plačevanje članarine Slovenije, za katero je pristojno Ministrstvo za izobraževanje, znanost in šport.

3. Repozitorij jezikovnih virov

Eden od osnovnih storitev infrastrukture CLARIN je zagotavljanje zanesljivega arhiviranja in dostopa do jezikovnih virov, kot so korpusi, leksikoni, avdio in video posnetki, slovnice, jezikovni modeli itd. Za dolgoročno hranjenje jeziko(slo)vnih raziskovalnih podatkov so mnogi centri CLARIN po Evropi že vzpostavili storitve za deponiranje, ki vključujejo tudi pomoč pri tehničnih in organizacijskih zadregah, povezanih z deponiranjem. Storitve za deponiranje CLARIN naj bi imeli naslednje značilnosti:

1. *dolgoročno arhiviranje*: zagotovljen je dostop za daljše obdobje;
2. vire je mogoče enostavno citirati s *trajnimi identifikatorji*;
3. viri in njihovi metapodatki so integrirani v infrastrukturo, kar omogoča *učinkovito iskanje* po katalogih;
4. dostop do zaščitenih virov je omogočen preko *enotnega overjanja identitete uporabnikov*;
5. vire, integrirane v infrastrukturo CLARIN je možno analizirati in obogatiti z raznovrstnimi *jezikoslovnimi orodji*.

[3] http://www.sidih.si
[4] http://www.adp.fdv.uni-lj.si
[5] http://www.slovenscina.eu

Trenutno je v okviru CLARIN aktivnih dvanajst centrov za deponiranje in arhiviranje jezikovnih virov, pri čemer jih je sedem v Nemčiji, dva na Nizozemskem in po eden v Avstriji in na Češkem. Kljub enakim zunanjim tehničnim zahtevam so različni centri ubrali različne poti pri implementaciji arhivov, začenši z osnovno platformo za njihovo izgradnjo.

3.1. Repozitorij LINDAT

Za Slovenijo je bil najbolj zanimiv pristop, ki so ga ubrali na Češkem, kjer so v okviru Instituta za formalno in uporabno jezikoslovje Karlove univerze v Pragi (UFAL[6]) postavili servis LINDAT,[7] ki ima enostaven in uporabniku prijazen vmesnik in prinaša večino funkcij, ki jih želimo vključiti v sodoben repozitorij v okviru omrežja CLARIN. Za razvoj in vzdrževanje servisa LINDAT skrbi razmeroma velika ekipa, pri tem pa je češka različica tudi že pridobila »Data Seal of Approval«,[8] torej potrdilo, da izpolnjuje pogoje za trajen in zaupanja vreden digitalni repozitorij. LINDAT je odprtokodno dostopen in kolegi z UFAL so nam prijazno pomagali pri namestitvi repozitorija LINDAT na IJS.

LINDAT je osnovan na platformi za gradnjo digitalnih repozitorijev DSpace[9] (Branschofsky in dr., 2002), ki je odprtokodni projekt z velikim številom namestitev. DSpace je eden uspešnejših projektov za razvoj institucionalnih digitalnih repozitorijev, ki so v zadnjem desetletju in pol nastali kot odgovor na vse večje potrebe po organiziranem objavljanju, arhiviranju, bibliografski obdelavi in kuratorstvu digitalnih dokumentov v akademskem okolju. V raziskovalnem in akademskem okolju nove publikacije (članki in knjige) ne le nastajajo, temveč so vedno bolj pogosto tudi uporabljene ali vsaj distribuirane v elektronski obliki (Crow, 2002). DSpace temelji na konceptu »trajnih dokumentov« (durable document space) in se naslanja na priporočila referenčnega modela Open Archival Information Systems (OAIS, CCSDS 650.0-R-2, 2001) in priporočil FEDORA (2002), na osnovi katerih je nastal sistem Fedora Commons. Če primerjamo sistem DSpace s splošno sprejetimi zahtevami za takšne sisteme (Kenney in McGovern, 2003) ter s sorodnimi sistemi, zlasti odprtokodnim projektom Fedora Commons (prim. Lagoze in dr., 2006), ki smo ga že uporabili kot osnovni gradnik za postavitev repozitorija za digitalne dokumente, ter uveljavljenim sistemom GNU ePrints (Nixon, 2003; Kim, 2005), ima DSpace, še zlasti v prilagojeni različici LINDAT, kot repozitorij virov v okviru slovenskega vozlišča CLARIN nekaj očitnih prednosti.

Repozitorij omogoča ločeno obravnavo zahtev in avtorizacije več skupin uporabnikov. Vsak dokument je sestavljen iz metapodatkov v standardnem zapisu Dublin Core (Powell in Johnston, 2003) in enega ali več paketov (bundles), ki lahko vsebujejo enega ali več bitnih tokov (bit streams). DSpace datoteke shranjuje kot bitne tokove, paketi pa omogočajo združevanje datotek v logične skupine (npr. dokument v zapisu HTML s pripadajočimi slikami je logično en paket).

Za stabilen dostop do posameznega dokumenta oz. drugih deponiranih virov in njihovo navajanje je poskrbljeno z neodvisnim sistemom stabilnih identifikatorjev na osnovi sistema kazalcev (handles), ki ga razvija in vzdržuje Corporation for National Research Initiative (CNRI).[10] Sistem je povsem integriran z repozitorijem in poskrbi za dodeljevanje, upravljanje in razreševanje trajnih identifikatorjev za digitalne objekte in druge vire na internetu. Ker je sistem s katalogom kazalcev oz. identifikatorjev neodvisen od repozitorija, je torej mogoče v primeru spremembe domenskega sistema, zamenjave uporabljene arhitekture ipd. posodobiti naslove, kamor kažejo kazalci, in tako zagotoviti trajno veljavnost povezav na spletnih straneh in navedkov v objavljenih publikacijah. Uporaba takšnega sistema je pomembna zahteva za repozitorije v omrežju CLARIN, saj je mogoče na ta način zagotoviti trajno dostopnost virov in ponovljivost eksperimentov.

DSpace prinaša vrsto vtičnikov za registracijo in overjanje uporabnikov, ki preko mehanizma skupin omogočajo različne stopnje avtorizacije in različne vloge za uporabnike. Tako je mogoče določiti urednike ali administratorje posameznih zbirk, ki preko delotokov z uporabniki sodelujejo pri deponiranju virov in poskrbijo za uporabo ustreznih formatov in metapodatkov. V različici LINDAT je uporabljen prilagojen vtičnik, ki uporablja protokol SAML (Security Assertion Markup Language), ki se uporablja v sistemih za enotno spletno overjanje AAI (Authentication and Authorization Infrastructure). Vsaka organizacija tako lahko postane varuh osebnih podatkov svojih članov, ponudnikom aplikacij pa se ni treba ukvarjati z dodeljevanjem uporabniških imen ter kočljivim zbiranjem in preverjanjem podatkov o uporabnikih. Hkrati sistem omogoča posredovanje atributov uporabnika, tako da je mogoče članstvo v skupinah določati tudi na osnovi podatkov o uporabniku, ki jih posreduje njegova matična organizacija, npr. ARNES[11].

AAI je postal ena od ključnih tehnologij evropskih akademskih omrežij in skupnega evropskega raziskovalnega prostora, ker omogoča vzpostavitev nacionalnih (in širših) federacij, v katerih AAI povezuje uporabnike in storitve v celoto ter pridruženim organizacijam omogoča dodeljevanje enotnega uporabniškega imena, ki lahko uporabnikom služi za vse vrste aplikacij, tako v domači kot v drugih organizacijah v isti federaciji. Trenutni razvoj tehnologije (v okviru iniciative eduGAIN[12]) že omogoča tudi podporo za gostujoče uporabnike iz drugih nacionalnih federacij, (podobno kakor pri sorodni tehnologiji za overjanje v omrežju Eduroam), vendar ta sistem še ne deluje povsod. Zato vtičnik repozitorija LINDAT omogoča hkratno uporabo več kot ene identifikacijske federacije, kar je posebej pomembno, ker ima CLARIN lastno federacijo AAI, ki je nastala še pred iniciativo eduGAIN in tako omogoča dostop tudi uporabnikom, ki niso člani federacije AAI.

DSpace ima modularno arhitekturo za spletne vmesnike. LINDAT uporablja izvedbo spletnega vmesnika

[6] http://ufal.mff.cuni.cz
[7] https://lindat.mff.cuni.cz
[8] http://datasealofapproval.org
[9] http://www.dspace.org
[10] http://en.wikipedia.org/wiki/Handle_System
[11] https://aai.arnes.si
[12] http://www.geant.net/service/eduGAIN

XMLUI, ki uporablja podatkovne tokove XML in je zgrajena na osnovi odprtokodnega javanskega sistema za razvoj spletnih aplikacij Apache Cocoon.[13] Uveljavljena in fleksibilna tehnologija sicer dodaja nekaj kompleksnosti, vendar je razvoj repozitorija LINDAT dokaz, da je mogoče hitro in učinkovito razviti uporaben in uporabniku prijazen vmesnik.

Za CLARIN.SI trenutno poteka prilagajanje vmesnika s pomočjo mehanizmov, ki so jih razvijalci v okviru razvoja servisa LINDAT predvideli za prilagoditev na uporabo v drugih institucijah, ter lokalizacija vmesnika za uporabo v slovenščini. Zaradi narave implementacije vmesnika je ta naloga nekaj zahtevnejša, saj je treba vzporedno nekatere segmente lokalizirati v programskih paketih v jeziku Java (preko nastavitvenih datotek), druge pa je treba lokalizirati v podatkovnih tokovih XML v okviru spletnega vmesnika.

Zaradi hitrega razvoja na področjih podpore za overjanje AAI, prilagoditve na slovensko vozlišče CLARIN, lokalizacije, metodologije deponiranja virov in integracije novih razvojnih različic LINDAT je sicer pričakovati še občasne hitre in velike spremembe, vendar pa trenutna pilotna postavitev na osnovi razvojne različice izvedbe repozitorija LINDAT že deluje[14] in je primerna za testiranje in preizkušanje servisa. Do konca 2014predvidevamo prehod na stabilne različice in postopen začetek testne uporabe s pravimi podatki.

3.2. Deponiranje virov

Ko bo repozitorij CLARIN.SI postal operativen, bo treba zagotoviti, da bo ponujal dovolj kvalitetnih in za raziskovalce zanimivih jezikovnih virov. V prvi fazi nameravamo v repozitorij prenesti odprto dostopne vire projekta SSJ, torej take vire, ki jih je mogoče prevzeti (*download*) na lasten računalnik. Ti viri naj bi naknadno tudi služili kot primer dobre prakse, vključno s postopkom validacije prevzetih virov. Istočasno nameravamo v repozitorij vključiti odprtodostopne vire, ki jih na IJS trenutno ponujamo na spletnih straneh posameznih projektov, ki so omogočili njihov nastanek; primera sta ročno označena korpusa slovenskega jezika projekta JOS (Jezikoslovno označevanje slovenskega jezika)[15] (Erjavec in Krek, 2008) in korpusa ter besedišče starejše slovenščine projekta IMP[16] (Erjavec in dr., 2011). Na IJS imamo še več manjših (specializiranih, večjezičnih) korpusov in drugih jezikovnih virov, vendar pa se bo kmalu potrebno zazreti tudi navzven, v prvi meri h konzorcijskim partnerjem CLARIN.SI, ki so izdelali že večje število raznovrstnih virov slovenskega jezika, od korpusov do slovarjev.

Po sklepu upravnega odbora CLARIN.SI bodo do konca leta 2014 člani konzorcija pripravili seznam jezikovnih virov, ki bi jih bili pripravljeni prispevati v repozitorij skupaj z licenco oz. informacijo, kakšni so pogoji njihove uporabe.

Pri vključevanju teh virov v platformo pričakujemo dva problema. Prvi je tehnične narave, saj so jezikovni viri zapisani v raznovrstnih formatih, ki niso vedno dokumentirani. Rešitev vidimo v tehnični in finančni podpori bodisi za dokumentiranje zapisa virov (tu bomo morali biti posebej pozorni na dober izbor metapodatkovne sheme) in, kjer bo to le mogoče, konverzijo v enega od standardnih formatov, npr. TEI[17] (TEI, 2007) ali LMF[18] (Francopoulo, 2013).

Večji problem bodo verjetno predstavljale omejitve pri nadaljnjem razširjanju virov, ki so v lasti posameznih ustanov, kjer bodo problematične avtorske pravice in nenaklonjenosti ideji, da bi se viri hranili na javnem repozitoriju in bili na voljo kateremukoli raziskovalcu ali celo v komercialne namene. Tu bo vsaj v začetku verjetno potrebno vsak problem reševati posebej, nato pa se bodo sčasoma nabrale izkušnje in primeri dobrih praks, tako v Sloveniji kot tudi v ostalih nacionalnih repozitorijih CLARIN.

Pri virih, ki bodo nastali v prihodnje, bo situacija, upamo, bolj enostavna, vsaj če se bo sprejet (in se bo izvajal) predlog Akcijskega načrta za jezikovno opremljenost, ki predvideva korake za večjo odprtost izdelanih jezikovnih virov, ki nastanejo kot rezultat javnega financiranja.

Repozitorij CLARIN.SI bi lahko poleg jezikovnih virov vseboval tudi odprtokodne programe za jezikoslovne obdelave oz. modele zanje, ki podpirajo (tudi) slovenski jezik. Za razliko od jezikovnih virov verjetno ne bi bilo smiselno ponujati (le) prevzema programov, temveč tudi njihovo evidentiranje, skupaj s kazalko na sistem za upravljanje izvorne kode, kot sta GIT ali SVN, kjer poteka razvoj.

4. Jezikoslovna orodja in storitve

Poleg vzpostavitve repozitorija je naloga infrastrukture CLARIN tudi vzpostavitev sistema spletnih storitev, kar je (še) bolj dolgotrajen in kompleksen proces. Spletne storitve lahko razdelimo na take, ki so namenjeni vizualizaciji vsebine jezikovnih virov (konkordančniki za korpuse in pregledovalniki različnih vrst slovarjev oz. leksikalnih baz) in tiste, katerih namen je obdelati neki jezikovni vir, predvsem označiti korpus za nadaljnje analize.

4.1. Spletni dostop do korpusov

Na platformi CLARIN.SI predvidevamo povezave do obstoječih spletnih konkordančnikov za slovenščino in, kjer bo to mogoče, njihovo agregiranje. Tudi CLARIN.SI bo ponujal svoj konkordančnik oz. konkordančnike, kjer bomo kot tehnološko in korpusno osnovo vzeli že obstoječa konkordančnika nl.ijs.si, in sicer noSketchEngine[19] in CUWI[20], ki že sedaj ponujata preko 20 korpusov (Erjavec, 2013). Drugi dobro obiskani vmesniki do korpusov, ki bi jih tudi bilo smiselno vključiti v agregiran oz. enovit dostop, so v Sloveniji vsaj še:

- spletišče SSJ, z dostopom do korpusov Gigafida (reprezentativen), KRES (uravnotežen), Gos (govorni) in Šolar (učenci slovenščine s popravki napak);
- spletni vmesnik do Nove Beseda ZRC SAZU, ki ostaja zelo cenjen korpusni vir;

[13] http://cocoon.apache.org
[14] https://www.clarin.si/repository/xmlui/
[15] http://nl.ijs.si/jos
[16] http://nl.ijs.si/imp
[17] http://www.tei-c.org
[18] http://www.lexicalmarkupframework.org
[19] http://nl.ijs.si/noske
[20] http://nl.ijs.si/cuwi

- večjezični Evrokorpus Službe vlade RS za evropske zadeve, ki je povezan s terminološkimi slovarji.

Povezovanje zelo raznorodnih iskalnikov oz. korpusov je zanimiv in verjetno ne dokončno rešljiv problem, je pa v perspektivi zelo smiseln, saj bi uporabniku, ki bi rad izvedel nekaj o slovenskem jeziku, radi ponudili čim bolj reprezentativne podatke na enem mestu.

4.2. Slovarski in terminološki portali

Na institucijah, ki so članice infrastrukture CLARIN.SI, že obstajajo portali, ki ponujajo različne slovarske in terminološke vire. Taki primeri so:

1. viri Inštituta za slovenski jezik Frana Ramovša na bos.zrc-sazu.si:
   - Slovar slovenskega knjižnega jezika
   - Slovar novejšega besedja slovenskega jezika
   - Slovenski pravopis 2001
   - Pleteršnikov Slovensko-nemški slovar
   - Besede slovenskega jezika (združeno besedišče iz SSKJ, BSJ, korpusa Nova beseda in spletnega iskalnika NAJDI.SI)
   - Odzadnji slovar slovenskega jezika
   - Besedišče slovenskega jezika (BSJ - besede, ki niso bile sprejete v SSKJ)
2. slovarski viri na nl.ijs.si, izhajajoči iz različnih raziskovalnih projektov:
   - Japonsko-slovenski slovar za učence japonščine
   - Besedišče starejše slovenščine (iz ročno označenega korpusa starejše slovenščine IMP)
   - slovenski WordNet oz. sloWNet v spletni aplikaciji sloWTool
3. slovarski ali leksikonski viri SSJ na portalu www.slovenscina.eu:
   - leksikon besednih oblik Sloleks
   - leksikalna baza za slovenščino
4. portal Termania, www.termania.net podjetja Amebis, ki trenutno vsebuje 43 slovarjev, od terminoloških, dvojezičnih, splošnih itd.
5. Terminologišče, isjfr.zrc-sazu.si/terminologisce terminološke sekcije na Inštitutu za slovenski jezik Frana Ramovša, na katerem je mogoče dostopati do devetih terminoloških slovarjev.

Zaenkrat smo se odločili za prenos portala Termania na strežnike CLARIN.SI, saj kljub temu, da niti program niti vsebovani slovarji niso odprti, zagotavlja ta koristen in prostodostopen servis. V načrtu je tudi preverjanje možnosti agregiranega iskanja po zgoraj omenjenih portalih, s čimer bi bilo uporabnikom omogočeno poenoteno iskanje in skupen prikaz rezultatov. V daljši perspektivi bi pa bilo smiselno zasnovati splošni vmesnik do slovarskih podatkov, ki bi nato gostil čim večje število slovarjev.

4.3. Orodja in storitve za označevanje

Infrastrukture posameznih evropskih centrov že ponujajo dostop do orodij in spletnih storitev. Češki LINDAT ponuja za češčino (in druge jezike) poleg iskalnika po skladenjsko označenih korpusih (drevesnicah) tudi storitve za oblikoskladenjsko označevanje, označevalnik imenskih entitet, strojno prevajanje med češčino in slovaščino itd.

Tudi za slovenščino že obstaja nekaj orodij in spletnih storitev, npr. za oblikoskladenjsko označevanje in lematizacijo sta na voljo ToTaLe[21] projekta JOS in Obeliks[22], za skladenjsko analizo pa označevalnik SSJ[23]. V prihodnosti bi bilo dobro te in novo razvite označevalnike združiti v platformo CLARIN.SI in jih ponujati pod skupnim vmesnikom in omogočiti dostop do njih preko spletnih protokolov za izvajanje programov, kakršen je WSDL.

4.4. Spletni delotoki

Pri vedno večjem številu označevalnih in drugih spletnih storitev se kmalu pojavi potreba po njihovem dinamičnem kombiniranju; temu služijo platforme za izdelavo in izvajanje spletnih delotokov. Njihov razvoj je v zadnjem času postali zelo popularen, tudi na področju označevanja besedil. V okviru nacionalnih infrastruktur CLARIN so najdlje prišli v Nemčiji, kjer so na Univerzi v Tübingenu razvili sistem WebLicht[24], ki omogoča (predvsem za nemščino) izdelavo verige označevalnikov, pri kateri za posamezne korake lahko izbiramo med več sistemi.

Tudi na IJS že več let razvijamo platformo za delotoke ClowdFlows[25] (Kranjc in dr., 2012), ki je trenutno sicer usmerjena predvsem v podatkovno rudarjenje, v perspektivi pa bi lahko služila tudi kot podlaga za obdelavo besedil; pilotno smo sistem že preizkusili z orodjem ToTrTaLe (Pollak in dr., 2012).

Pri CLARIN.SI bi z implementacijo svoje platforme za izdelavo in izvajanje spletnih delotokov počakali na zadostno število virov in spletnih storitev, da ima izdelava sistema za njihovo dinamično kombiniranje smisel. Kompleksne operacije nad velikimi podatkovnimi množicami potrebujejo tudi velike računalniške kapacitete, kjer pa bi CLARIN.SI verjetno lahko uporabil slovenski nacionalni grid, SLING.[26]

Ker imajo v Nemčiji, kmalu pa mogoče tudi v drugih nacionalnih centrih CLARIN, takšne platforme z ustreznimi kapacitetami že na voljo, se tu pojavi tudi vprašanje, ali je smiselno vlagati v razvoj platforme CLARIN.SI za ustvarjanje in izvajanje delotokov, saj bi bilo verjetno bolj smiselno, da bi posamezne storitve, orodja ali modele za slovenščino enostavno ponudili v uporabo drugim platformam.

5. Zaključki

V prispevku smo predstavili prve korake slovenske raziskovalne infrastrukture CLARIN.SI[27] in načrte za nadaljnje delo po posameznih področjih, predvsem pri vzpostavljanju računalniške platforme in zagotavljanju virov in orodij, ki jo bodo osmislila.

V nadaljevanju pa bo seveda potrebno poskrbeti tudi za promoviranje CLARIN.SI, tako da se bodo tuji, predvsem pa domači raziskovalci, učitelji, študentje in drugi

[21] http://nl.ijs.si/analyse
[22] http://www.slovenscina.eu/tehnologije/oznacevalnik
[23] http://www.slovenscina.eu/tehnologije/razclenjevalnik
[24] http://weblicht.sfs.uni-tuebingen.de
[25] http://clowdflows.org
[26] http://www.sling.si
[27] http://www.clarin.si

potencialni uporabniki zavedali, da platforma obstaja in da jim lahko – upajmo – pomaga pri raziskovanju slovenskega jezika.

Zahvala

Avtorji se zahvaljujejo anonimnima recenzentoma za koristne pripombe. CLARIN.SI financira Ministrstvo za izobraževanje, znanost in šport Republike Slovenije.

Literatura

Branschofsky, M., Chudnov, D, 2002. "DSpace: Durable Digital Documents." V *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM), str. 372. doi:10.1145/544220.544319.
URL: http://hdl.handle.net/1721.1/26703

Crow, R., 2002. The Case for Institutional Repositories: A SPARC Position Paper. URL: http://www.sparc.arl.org/sites/default/files/media_files/instrepo.pdf (4. 7. 2014)

CCSDS 650.0-R-2: *Reference Model for an Open Archival Information System (OAIS)*. Red Book. Issue 2. June 2001.
URL: http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf.

Erjavec, T., Krek, S., 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. V T. Erjavec, J. Žganec Gros (ur.), *Zbornik šeste konference Jezikovne tehnologije*: Ljubljana: Institut Jožef Stefan, 49–53.

Erjavec, T., 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. V zborniku: *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland.

Erjavec, T., Jerele, I., Kodrič, M., 2011. Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT. V: KRANJC, Simona (ur.). *Meddisciplinarnost v slovenistiki, (Obdobja, Simpozij, = Symposium, 30)*. Ljubljana: Znanstvena založba Filozofske fakultete, 41-47.

Erjavec, T., 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1 (1): 24–49. URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf

Flexible and Extensible Digital Object and Repository Architecture (FEDORA): URL: http://www.cs.cornell.edu/cdlrg/fedora.html

Francopoulo, G. (ur.), 2013. *LMF Lexical Markup Framework*. Wiley-ISTE.

Kenney, A. R., McGovern, N.Y., 2003. The Five Organizational Stages of Digital Preservation. V *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee*, ur. Patricia Hodges, Mark Sandler, Maria Bonn, in John Price Wilkin, 122–53. Ann Arbor: The Scholarly Publishing Office, University of Michigan Library.

Kim, J., 2005. Finding Documents in a Digital Institutional Repository: DSpace and Eprints. V: *68th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, Charlotte, ZDA, 28. 10. – 2. 11. 2005.

Kranjc, J., Podpečan, V., Lavrač, N., 2012. ClowdFlows: A Cloud Based Scientific Workflow Platform / Machine Learning and Knowledge Discovery in Databases. V *Lecture Notes in Computer Science*. Volume 7524. Springer, pp 816-819.

Krek, S., Grčar, M., Dobrovoljc, K., 2012. Označevalnik za slovenski jezik Obeliks. *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012*, Ljubljana, Slovenia. 89-94.

Lagoze, C., Payette, S., Shin, E., Wilper C., 2006. "Fedora: An Architecture for Complex Objects and Their Relationships." International Journal on Digital Libraries 6(2): 124–38. doi:10.1007/s00799-005-0130-3.

Lewis, K. D., Lewis J. E., 2009. »Web Single Sign-On Authentication using SAML«. *IJCSI International Journal of Computer Science Issues*, zv. 2.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S., 2012. *Korpusi slovenskega jezika* Gigafida, KRES, ccGigafida *in* ccKRES*: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Nixon, W., 2003. DAEDALUS: Initial Experiences with Eprints and DSpace at the University of Glasgow. Ariadne, št. 37.
URL: http://www.ariadne.ac.uk/issue37/nixon (4. 7. 2014)

Pollak, S., Trdin, N., Vavpetič, A., Erjavec, T., 2012. NLP web services for Slovene and English: morphosyntactic tagging, lemmatisation and definition extraction. *Informatica*, 36/4, str. 441-449.

Powell, A., Johnston, P. 2003. *Guidelines for Implementing Dublin Core in XML*. URL: http://dublincore.org/documents/dc-xml-guidelines/ (4. 7. 2014)

Smith, MacKenzie, Bass, M., McClellan, G., Tansley, R., Barton, M., Branschofsky, M., Stuve, D., Harford Walker J., 2003. DSpace: An Open Source Dynamic Digital Repository, *D-Lib Magazine*, 9. zv., jan. 2003. URL: http://dlib.org/dlib/january03/smith/01smith.html (4. 7. 2014)

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., Smith, M., 2003. The DSpace institutional digital repository system: current functionality. *JCDL '03, Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 87-97.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. URL: http://www.tei-c.org/Guidelines/P5.

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. *6th International Conference on Language Resources and Evaluation (LREC 2008)*.