

## PLAN UPRAVLJANJA ISTRAŽIVAČKIM PODACIMA (PUP)

Opće informacije		
	Ime i prezime predlagatelja	Matea Andrea Birtić
	Matična organizacija	Institut za hrvatski jezik
	Naziv projekta	Sintaktička i semantička analiza dopuna i dodatka u hrvatskome jeziku SARGADA (IP 2019-04-7896)
	Upravitelj podacima	Siniša Runjaić
1. Prikupljanje podataka i dokumentacija		
	Koje ćete podatke prikupljati, obrađivati, stvarati ili se ponovno njima koristiti? (navedite formate, vrste i opseg svih podataka s kojima ćete raditi, a ne samo krajnji skup podataka koji će biti rezultat istraživanja)	Projekt je teorijske lingvističke prirode, stoga će podatci biti tekstualni. Projekt ima zahtjevniji teorijski dio, a sama primjena (baza podataka s analizama rečenica –repositorij SARGADA) izravan je proizvod zaključaka istraživanja i nije zahtjevan opsegom.
	Kako će se podaci prikupljati, obrađivati ili stvarati? (ukratko navedite metodologiju i procese osiguranja kvalitete, načine organiziranja podataka te alate i instrumente kojima ćete se koristiti za prikupljanje i obradu)	Projekt uključuje stvaranje baze podataka, a podatci u njoj su mahom tekstualni, ali izvlači se i elementarna statistika. Temelj su baze analize javno dostupnih rečenica na hrvatskom jeziku koje sadržavaju upitne sveze riječi u kojima teorijski nije jasno određeno je li riječ o sintaktičkim pojmovima dopune ili dodatka. Analize (rastavljanje rečenica i pripreme za označivanje dijelova teksta (tzv. <i>tagiranje</i> )) prije unosa u samu bazu podataka spremaju se u zajedničke datoteke u oblaku.
	Koju ćete dokumentaciju i metapodatke izraditi osim podataka? (dokumentacija mora sadržavati informacije i standarde potrebne korisnicima kako bi mogli samostalno čitati i interpretirati podatke u budućnosti, primjerice, kodne knjige, <i>ReadMe</i> datoteke i sl.)	Podatci iz analiziranih rečenica uneseni su i obrađeni u bazi podataka i stvoren je grafički predložak za interakciju s vanjskim korisnicima i pretragu unesenih podataka na mrežnim stranicama projekta, s uputama za razumijevanje i pretraživanje. Metapodatci o članovima projekta koji unose podatke, vremenu unosa i ostalo zapisuju se u projekt softvera <i>Laravel</i> koji služi kao pozadinski sustav za izradu aplikacije i sučelja za unos kojim se članovi projekta služe.

2.	Pravna i sigurnosna pitanja	
	Jeste li ograničeni sporazumom o povjerljivosti? Imate li potrebna dopuštenja za prikupljanje, obradu, čuvanje i dijeljenje podataka? Jesu li osobe čiji se podaci obrađuju informirani o tome i jesu li dali privolu? Kojim ćete se metodama koristiti u svrhu zaštite osjetljivih podataka (GDPR - posebne kategorije osobnih podataka, navesti metode anonimizacije podataka)?	Ne postoji poseban ugovor o povjerljivosti jer nemamo osjetljive podatke ili podatke koji se tiču kategorija osobnih podataka, odnosno riječ je isključivo o analizi tekstualnih podataka.
	Kako će se regulirati pristup podacima i njihova sigurnost? Koji su potencijalni rizici koje treba uzeti u obzir? Kako ćete osigurati sigurnost pohrane osjetljivih podataka?	Podatci se atomizirano upisuju u relacijsku bazu podataka te priređen središnji sustav za upravljanje podacima (CMS) regulira pristup i otklanja moguće sigurnosne probleme. Potencijalni rizik može predstavljati neovlašten pristup treće strane u institutski server, ali ne očekuje se veći rizik od takve pojave jer nije riječ o osjetljivim podacima. U CMS baze podataka ugrađena je i izrada sigurnosne kopije.
	Kako ćete upravljati zaštitom autorskih prava i drugog intelektualnog vlasništva? Tko će biti vlasnik podataka? Koje će se licencije primjenjivati na podatke? Koja će se ograničenja primjenjivati na ponovnu uporabu osobnih podataka?	Primijenit će se osnovna licencija <i>Creative Commons</i> za konzultiranje podataka na javnoj dostupnoj verziji. Podatci će biti dostupni akademskoj zajednici i javnosti na slobodno korištenje, uz navođenje izvora.
3.	Pohrana i čuvanje podataka	
	Kako će radne verzije podataka biti pohranjene tijekom projekta? Kako će se napraviti sigurnosne kopije tih podataka ( <i>backup</i> )? Koja je očekivana količina podataka koja će se prikupiti i čuvati tijekom projekta (izraženo u MB/GB/TB)?	Podatci s radnim verzijama spremeni su na serverima Instituta za hrvatski jezik i jezikoslovlje u formatima koji su zadani pozadinskim sustavom softvera <i>Laravel</i> , a za eventualne potrebe razmjene s drugim projektima izradit će se kopije u formatu koji je čitljiv na različitim platformama (npr. <i>JSON</i> ). Podatci će ostati javno dostupni na mrežnim stranicama nakon završetka projekta. Sigurnosna kopija pohranjuje se na sigurnoj mapi na serverima Instituta za hrvatski jezik.
	Kako će se završne verzije podataka dugotrajno pohraniti i čuvati (i nakon završetka projekta)? U kojim će se formatima čuvati podaci? Koja je očekivana količina podataka koja će se trajno pohraniti	Podatci iz završne inačice čuvat će se također na serverima Instituta za hrvatski jezik u formatima koji su zadani pozadinskim sustavom softvera <i>Laravel</i> (format <i>JSON</i> ). Kako je riječ o atomiziranim tekstualnim podacima, ne previđa se opseg koji bi zauzimao previše smještajnog prostora (do 100 MB).

	(izraženo u MB/GB/TB)?	
4.	Dijeljenje i ponovna uporaba podataka	
	Kako i gdje će se podaci dijeliti? Koji repozitorij će se koristiti za dijeljenje podataka? Kako će potencijalni korisnici doznati za podatke?	Podatci su pretraživi na mrežnoj stranici projekta, a korisnici za to doznaju tijekom diseminacijskih aktivnosti projekta.
	Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja.	Nema ograničenja jer je riječ o javno dostupnim rezultatima istraživanja projektom koji financira Hrvatska zaklada za znanost, a analiziraju se rečenice koje su ionako javno dostupne u slobodno pretraživim korpusima hrvatskoga jezika i znanstvenim publikacijama i ne podliježu autorskim pravima ili etičkim ograničenjima.
	Potvrdite da ćete se koristiti digitalnim repozitorijem koji je u skladu s načelima <i>FAIR-a</i> .	Da.
	Potvrdite da ćete se koristiti digitalnim repozitorijem koji održava neprofitna organizacija (ako ne, objasnite zašto ne možete dijeliti podatke na digitalnom repozitoriju koji nije komercijalan).	Da, Institut za hrvatski jezik kao javna ustanova nije profitna organizacija.