Hrvatski prijedlozi u upotrebi – semantička i sintaktička analiza (PUP)

Matas Ivanković, Ivana

Data management plan / Plan upravljanja istraživačkim podacima

PLAN UPRAVLJANJA ISTRAŽIVAČKIM PODACIMA (PUP)

Opće informacije		
	Ime i prezime predlagatelja	Ivana Matas Ivanković
	Matična organizacija	Institut za hrvatski jezik
	Naziv projekta	Hrvatski prijedlozi u upotrebi – semantička i sintaktička analiza
	Upravitelj podacima	Ivana Matas Ivanković
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)	Za istraživanje će se koristiti podaci iz hrWaC-a, javno dostupnog korpusa, koji sadržava preko 1.2 milijardu riječi, odnosno preko 67 milijuna rečenica. Za pristup korpusu koristit će Sketch Engine (https://www.sketchengine.eu/) i mrežno sučelje dostupno na adresi https://www.clarin.si/noske/all.cgi/first_form?corpname=hrwac Podatci će se nakon obrade spremati u relacijsku bazu podataka, posebno dizajniranu za potrebe projekta, a moći će se izvesti i u CSV datoteku. Cilj je projekta prikupiti minimalno 17000 zapisa o barem 125 prijedloga.
	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)	Za pretraživanje postojećeg korpusa koristit će se alati navedeni u prethodnom dijelu. Iz dostupnih podataka uzimat će se dijelovi važni za istraživanje, obradit će se u nekom od pomoćnih alata (Notepad++, Word) i nakon toga će se spremati u relacijsku bazu pomoću aplikacije izrađene u sklopu projekta.
	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, *ReadMe* datoteke i sl.)	Na javno dostupnoj mrežnoj aplikaciji bit će detaljno objašnjeno značenje pojedinih vrsta podataka i zapisa. Također, korisnicima će biti dostupne upute za korištenje aplikacije, odnosno za pretraživanje podataka i pregled detalja o zapisima.
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 sporazum o povjerljivosti. Korpus na kojem se temelji istraživanje javno je dostupan i bit će referenciran u rezultatima istraživanja, javnom sučelju i objavljenim radovima. Ne prikupljaju se podatci o osobama i u projektu se ne radi s osjetljivim podatcima.
	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 će se nalaziti u relacijskoj bazi podataka pa sam sustav za rad s bazom regulira pristup i otklanja potencijalne sigurnosne probleme. Potencijalni rizik jest napad na server ili neku od aplikacija, koji može uzrokovati neželjenu izmjenu ili brisanje podataka. Zbog navedenog će se periodički raditi sigurnosna kopija, koja će se spremati na drugo, udaljeno računalo.
	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?	Vlasnik podataka bit će Institut za hrvatski jezik. 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 (*backup*)? Koja je očekivana količina podataka koja će se prikupiti i čuvati tijekom projekta (izraženo u MB/GB/TB)?	Sigurnosne kopije podataka u bazi podataka i podataka nastalih u međukoracima periodički će se pohranjivati na udaljena mjesta, odnosno u oblak (jednom mjesečno). Za izradu sigurnosne kopije baze podataka upotrebljavat će će se sustav za sam sustav za upravljanje, a sigurnosne kopije ostalih podataka izrađivat će sudionici na projektu. Očekivana je veličina baze podataka 100 MB, a ako se u obzir uzmu i podatci vezani uz sigurnost aplikacije i izradu dnevnika, veličina će biti 1 GB.
	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 (izraženo u MB/GB/TB)?	Podatci će ostati dostupni kroz grafičko sučelje i nakon završetka projekta. Napravit će se zadnja sigurnosna kopija i pohraniti na sigurnu, udaljenu lokaciju. Osim navedenog, kako bi se osigurala buduća upotreba i kompatibilnost s aplikacijama, napravit će se izvoz podataka u CSV datoteku. Očekivana je veličina navedene datoteke 50 MB.
4.	Dijeljenje i ponovna uporaba podataka	

	Kako i gdje će se podaci dijeliti? Koji repozitorij će se koristit za dijeljenje podataka? Kako će potencijalni korisnici doznati za podatke?	Podatci će biti dostupni javnosti kroz grafičko sučelje. Aplikacija će biti dostupna sa službenih stranica Instituta. Korisnici će za aplikaciju doznati i kroz objavljene radove, a na zahtjev će moći dobiti pristup i izvezenim podatcima u CSV formatu.
	Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja.	Ne postoje.
	Potvrdite da ćete se koristiti digitalnim repozitorijem koji je u skladu s načelima *FAIR-a*.	Potvrđujem.
	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).	Potvrđujem.