Višerječni izrazi u hrvatskome jeziku – leksikološki, računalnolingvistički i glotodidaktički pristup – MWE-Cro

Blagus Bartolec, Goranka

Data management plan / Plan upravljanja istraživačkim podacima

PLAN UPRAVLJANJA ISTRAŽIVAČKIM PODACIMA (PUP)

Opće informacije		
	Ime i prezime predlagatelja	Goranka Blagus Bartolec
	Matična organizacija	Institut za hrvatski jezik
	Naziv projekta	Višerječni izrazi u hrvatskome jeziku – leksikološki, računalnolingvistički i glotodidaktički pristup – MWE-Cro
	Upravitelj podacima	Goranka Blagus Bartolec
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)	U okviru projektnog prijedloga prikupljat će se podatci za ukupno 5 *online* baza podataka u skladu s ciljevima projekta. Kao podatci prikupljat će se višerječni izrazi (skupine od dvije ili više riječi) i cijele rečenice (poslovice, idiomi, primjeri uporabe). Podatci će biti prikupljeni iz javno dostupnog korpusa hrWaC korištenjem licenciranog korpusnog alata Sketch Engine te iz javno dostupne preliminarne *Kolokacijske baze.* Ti će se podatci obraditi i obogatiti potrebnim sadržajem u skladu s ciljevima stvaranja proširene i proširive baze podataka. Za svih 5 baza podataka prikupljat će se podatci u količini od 40.000 jezičnih jedinica (višejezičnih izraza).
	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)	Podatci će se skupljati u relacijsku bazu podataka Access i SQL bazu.
	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.)	Korisnici će na mrežnim stranicama imati upute u načinima i mogućnostima pretraživanja podataka, te će biti naveden kontakt za komunikaciju korisnika sa stručnjacima u vezi s pretragom podataka u bazi.
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 skupljaju se nikakvi podatci o korisnicima na javnom mrežnom mjestu. Provedba projekta nije ograničena obradom povjerljivih 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?	Mrežna stranica bit će zaštićena HTTPS protokolom za izmjenu osjetljivih podataka. Sigurnost pohrane osjetljivih podataka osigurana je SQL bazom podataka s pripadajućom tajnom lozinkom i imenom kreirane baze.
	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?	Upravljanje zaštitom podataka regulira Institut za hrvatski jezik svojim pravnim aktima (ugovorima s informatičarima i vanjskim tvrtkama) te na internetskim stranicama na kojima će podatci biti objavljeni. Vlasnik svih podataka i baza podataka u okviru projekta bit će Institut za hrvatski jezik, što će se osigurati licencama (Creative Commons) pod kojima će podatci biti objavljeni i dostupni korisnicima.
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)?	Podatci će se periodično (jednom mjesečno) automatski pohranjivati na serveru i na računalima suradnika na projektu u obliku sigurnosnih kopija te na računalu voditelja projekta u programima Access i Excel. Također, svi podatci će se periodično pohranjivati u oblaku i na USB. Očekivana količina podataka je 122 MB.
	Kako će se završne verzije podataka dugotrajno pohraniti i čuvati (i nakon završetka projekta)? U kojim će se formatima čuvati Koja je očekivana količina podataka koja će se trajno pohraniti (izraženo u MB/GB/TB)?	Podatci će biti javno dostupni na serveru Instituta za hrvatski jezik u grafičkome sučelju. Podatci će se izvesti u CSV datoteku i u Access relacijsku bazu. Očekivana količina podataka iznosi 120 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 se dijeliti u okviru javno dostupne baze na mrežnoj domeni IHJJ-a (ihjj.hr i jezik.hr). Korisnici će svim podatcima pristupati s pomoću tražilice te diseminacijom objavljenih rezultata i radova unutar projekta, ponajprije na mrežnim stranicama projekta i Instituta za hrvatski jezik.
	Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja.	Nema takvih podataka.
	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.