Plan upravljanja istraživačkim podacima

|  |  |  |
| --- | --- | --- |
| Opće informacije | | |
|  | Ime i prezime predlagatelja | Prof. dr. sc. Aneta Stojić |
|  | Matična organizacija | Sveučilište u Rijeci, Filozofski fakultet |
|  | Naziv projekta | Metaforičke kolokacije – sintagmatske sveze između semantike i pragmatike |
|  | Upravitelj podacima | Aneta Stojić, aneta.stojic@ffri.uniri.hr |
| 1. | Prikupljanje podataka i dokumentacija | |
|  | Koje ćete podatke prikupljati, obrađivati, stvarati ili se ponovno njima koristiti? (navedite format, vrstu i opseg podataka) | Podaci koji će se prikupljati metaforičke su kolokacije u četiri jezika: hrvatskom, njemačkom, engleskom i talijanskom. U prvom će se koraku izlučivati najfrekventnije imenice u hrvatskom jeziku kao i njihovi najučestaliji kolokacijski partneri, a odredit će se automatskim računalnim alatima. Računalnim alatom SketchEngine pretraživat će se korpus hrvatskoga jezika hrWaC (Ljubešić, 2014) kojeg čine tekstovi sakupljeni s interneta, a sadrži preko 1,2 milijarde riječi. Isti će se postupak provesti na paralelnim korpusima (InterCorp, version 12) pa će se dodatnim postupkom izlučiti one imenice koje se kao najfrekventnije pojavljuju u komparabilnim korpusima (hrWac, deWac, ukWac i itWac) za hrvatski, njemački, engleski i talijanski jezik. Podaci će biti pohranjeni u XLSX datotekama za svaki pojedini jezik, prosječne veličine od 1 MB. |
|  | Kako će se podaci prikupljati, obrađivati ili stvarati? (ukratko navedite metodologiju i procese osiguranja kvalitete te načine organiziranja podataka) | Podaci će se prikupljati korpusno-analitičkim utvrđivanjem kookurencija, tj. riječi koje se statistički nadprosječno često supojavljuju u svezi, uglavnom u određenom maksimalnom razmaku od pet pozicija. Spomenuti računalno-lingvistički pristup kombinirat će se s teorijsko-semantičkim koji se temelji na utvrđivanju smislene veze između obiju sastavnica. Semantički međuovisne sastavnice kolokacijske sveze utvrđivat će se manualnim ekstrahiranjem kolokacija nakon izgradnje kolokacijskih profila odabranih imenica. Iz kolokacijskih profila selektirat će se one kolokacijske sveze koje iskazuju metaforično značenje. Semantičkom analizom utvrdit će se prototipni obrasci te analizirati korišteni mehanizmi metaforizacije. Temeljem toga izradit će se prijedlog kriterija za kategoriziranje/tipologiziranje dobivenih metaforičkih kolokacijskih sveza. |
|  | Koju ćete dokumentaciju i metapodatke ustupiti osim podataka? (navedite koje su informacije potrebne korisnicima kako bi mogli čitati i interpretirati podatke u budućnosti te koji će se standardi koristiti pri tumačenju podataka) | Dokumentacija uz podatke bit će: ogledni primjerak kolokacijskog profila s upisanim kategorijama aspekata analize.  Dokumentacija će sadržavati tablični prikaz relevantnih gramatičkih relacija ostvarenih između baze i kolokatora, frekvenciju pojavljivanja specifičnih kolokacija u mrežnom korpusu, te semantičko-pragmatičko određenje sintagmatske sveze s primjerima iz upotrebe. |
| 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 pohranjuju informirani o tome i jesu li dali privolu? Kojim ćete se metodama koristiti u svrhu zaštite osjetljivih podataka (GDPR - posebne kategorije osobnih podataka)? | Pri izvedbi ovog projekta neće se kršiti etička načela. |
|  | 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? | Podaci će se obrađivati i njima upravljati u zaštićenom nemrežnom okruženju koristeći se virtualnom  desktop tehnologijom. |
|  | Kako ćete upravljati zaštitom autorskih prava i intelektualnog vlasništva? Tko će biti vlasnik podataka? Koje će se licencije primjenjivati na podatke? Koja će se ograničenja primjenjivati na ponovnu uporabu osobnih podataka? | Podaci su prikladni za dijeljenje. Podaci su dobiveni promatranjem (dakle jedinstveni su) i mogli bi se  koristiti za druge analize. Prilike za novu upotrebu su velike. Zbog toga je naš cilj da omogućimo najširu moguću upotrebu podataka i objavit ćemo ih pod Creative Commons CC0. |
| 3. | Pohrana i čuvanje podataka | |
|  | Kako će podaci biti pohranjeni i kako će biti napravljena sigurnosna kopija podataka (*backup*) tijekom istraživanja? Koji su kapaciteti čuvanja podataka kojim raspolažete? Kojim se procedurama koristite za sigurnosnu kopiju (*backup*)? | Podaci će se tijekom istraživanja s računala glavnog istraživača kopirati u nacionalni sustav za pohranu i  dijeljenje podatka Puh (https://www.srce.unizg.hr/puh) koji članovima projektnog tima omogućava pristup  aktualnoj verziji podataka i na kojem se dnevno automatizirano izrađuje sigurnosna kopija podataka. Uz to, glavni  istraživač dnevno radi sigurnosnu kopiju s računala na vanjski disk. |
|  | Koji je vaš plan čuvanja podataka? U kojim će se formatima čuvati? | Podatke ćemo čuvati trajno u institucionalnom repozitoriju Filozofskog fakulteta u Rijeci (<https://repository.ffri.uniri.hr/>) uspostavljenom na sustavu Dabar. Tablične podatke čuvat ćemo u XLSX obliku (Microsoft Office Excel), a tekstualne u DOCX (Office Open XML) te PDF-a obliku. DOC oblik obavezno će se konvertirati u DOCX oblik. |
| 4. | Dijeljenje i ponovna uporaba podataka | |
|  | Kako i gdje će se podaci dijeliti? Na kojem repozitoriju planirate dijeliti podatke? Kako će potencijalni korisnici doznati za podatke? | Konačnu verziju skupa podatka voditelj projekta podijelit će putem institucionalnog repozitorija Filozofskog fakulteta u Rijeci uspostavljenog u nacionalnom sustavu Dabar gdje će biti pohranjene i publikacije i ostala projektna dokumentacija. Podaci će biti objavljeni pod CC0 licencom. Institucijski repozitorij u sustavu Dabar odabrali smo jer podržava FAIR načela: skupovima dodjeljuje trajni identifikator URN:NBN, osigurava vidljivost podataka putem OpenAIRE portala i Google Scholara te tražilice dabar.srce.hr, a ujedno doprinosi vidljivosti i transparentnosti rada Sveučilišta u Rijeci. |
|  | Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja. | Ne postoje takvi podaci, stoga nisu potrebna ograničenja. |
|  | Potvrdite da ćete se koristiti digitalnim repozitorijem koji je u skladu s načelima *FAIR-a*. | Institucijski repozitorij u sustavu Dabar odabrali smo jer podržava FAIR načela. |
|  | 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). | Dabar je nacionalni akademski arhiv i repozitorij, te dio nacionalne javne istraživačke infrastrukture. Osnovan je pri Računalnom centru Sveučilišta u Zagrebu (SRCE), a partnere mu čine Institut Ruđer Bošković, Nacionalna i sveučilišna knjižnica u Zagrebu, Filozofski fakultet Sveučilišta u Zagrebu te Medicinski fakultet Sveučilišta u Zagrebu. |

Ref:

[1] Celjak, D., Dorotić Malič, I., Matijević, M., Poljak, Lj., Posavec K. i Turk, I.: „Istraživački podaci - što s njima?“ [Istraživački podaci - što s njima? : priručnik o upravljanju istraživačkim podacima | Digitalni repozitorij Srca (unizg.hr)](https://repozitorij.srce.unizg.hr/islandora/object/srce:327)