European Language Resource Coordination
Connecting Europe Facility

Radionica ELRC-a u Hrvatskoj

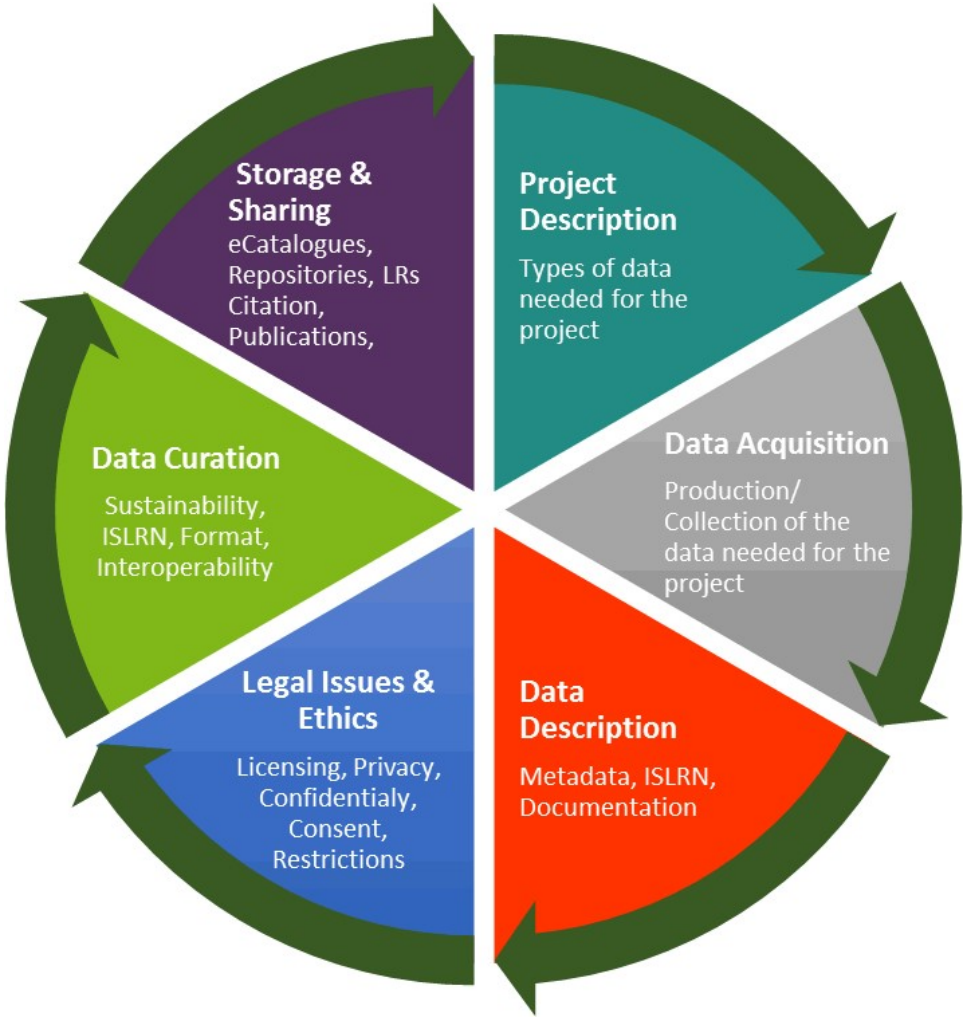
Zagreb, 2019-02-12

Pronalaženje i upravljanje podatcima: pitanja i odgovori

Prof. dr. sc. Marko Tadić
Sveučilište u Zagrebu, Filozofski fakultet

Plan upravljanja podatcima (PUP)

Plan upravljanja podatcima određuje kako će se upravljati podatcima tijekom procesa njihove proizvodnje i nakon njega. Pokriva čitav životni vijek podataka i određuje načine kako će se učinkovito upravljati podatcima ne bi li se osigurala njihova trajna dostupnost.



European Language Resource Coordination
Connecting Europe Facility

Pitanja kod sastavljanja PUP-a

- **Predvidjeti sva moguća pravna pitanja**
  - Osigurati da su u slučaju vaših podataka autorskih prava riješena
  - Osigurati da vanjski isporučitelji na vašu ustanovu prebacuju pravo vlasništva i sva druga prava nad podatcima
  - Osigurati da svu svi interno proizvedeni dokumenti vaši (npr. prijevodne memorije)
  - Unaprijed provjeriti pitanja privatnosti i napraviti plan anonimizacije ako je potrebno
- **Odrediti plan upravljanja podatcima s obzirom na zadatak**
  - Ovo se mora odnositi na glavni cilj (npr. pisanje dokumenta, prijevod dokumenta itd.)
- **Izraditi plan za višestruku uporabu podataka** (od dokumentacije do jezičnih resursa)
  - Zatražite podatke u uporabivu zapisu (ne samo PDF, nego i TMX/XML itd.)
  - Osigurajte da su vaši podatci smješteni na suvremenim medijima (npr. ne CD-ovima!)
- **Predvidite buduće objave i dijeljenje podataka** kao podataka iz javnoga sektora (PSI)

European Language Resource Coordination
Connecting Europe Facility

Pitanja?

Pitanja: autorska prava

Ako javna agencija da na vanjsko prevođenje tekst nad kojim ima autorska prava, čija su autorska prava nad prevedenom inačicom toga teksta? Može li se prijevod ustupati dalje?

To ovisi o tome kako su autorska prava regulirana ugovorom s vanjskim prevoditeljima. Javne bi agencije ugovorno morale osigurati da nad vanjskim naručenim prijevodima imaju sva prava ponovne uporabe i njihova dijeljenja u obliku prijevodnih memorija.

Pitanja: autorska prava

Sastavio sam korpus literarnih tekstova za potrebe svoga istraživanja. Mogu li ga donirati ELRC-u?

Svi tekstovi uključeni u korpus moraju biti s razriješenim autorskim pravima. Neki tekstovi, osobito stariji, već mogu biti u javnoj domeni (npr. autorska su prava istekla). Za ostale tekstove mora se od nositelja autorskih prava pribaviti licencija za redistribuciju trećim stranama.

Pitanja: autorska prava

Imam autorska prava nad prijevodom, ali ne i nad izvornim tekstom (ili obrnuto). Smijem li dijeliti paralelne podatke? Koje bi korake morao poduzeti u takvom slučaju?

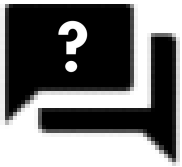
Kako bi se paralelni podatci mogli (re)distribuirati, autorska prava moraju biti razriješena i za izvorni i za prevedeni tekst. Ukoliko je izvorni tekst (ili njegov prijevod) zaštićen autorskim pravima, od vlasnika autorskih prava mora se pribaviti licencija kojom se dopušta dijeljenje teksta trećim stranama. Stoga je prvi korak kontaktirati vlasnika autorskih prava i provjeriti je li tekst dostupan pod nekom otvorenom licencijom, ili se mora dogovoriti kakav drukčiji licencijski mehanizam.

Pitanja: autorska prava

Sastavili smo dvojezični terminološki resurs iz naših vlastitih podataka i taj nam je resurs interno dostupan uz neke druge korpuse i rječnike koje inače koristimo. Nismo sigurni smijemo li distribuirati novonapravljene resurse s nekom od CC licencija.

Ako se novi resurs izgrađuje iz nekoliko već postojećih resursa, autorska prava za sve resurse moraju se razriješiti. Licencije bi morale dopustiti redistribuciju postojećih resursa i iz njih izvedene nove resurse (npr. bez CC-ND). Ako ste vi vlasnik resursa i dali ste pravo redistribucije trećoj strani, osigurajte u distribucijskome sporazumu da i vi zadržite dodatno pravo distribucije resursa pod nekom CC-licencijom.

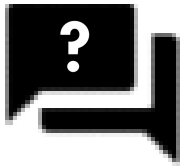
Pitanja: autorska prava



Sastavio sam korpus iz tekstova s Wikipedije. Wikipedija se objavljuje pod licencijom CC-BY-SA 3.0. Pod kojom se licencijom mora objaviti resurs izveden iz nje? CC v3.0 ili CC v4.0?

Tko god prilagođava podatke s licencijom BY-SA, mora na izvedne podatke primijeniti licenciju usklađenu s BY-SA. U slučaju BY-SA 3.0, sve buduće inačice su usklađene s BY-SA tako da se izvedeni podatci modu objaviti pod CC-BY-SA 4.0

Pitanja: autorska prava



Imamo skup podataka koji nije pod CC-licencijom. Smijemo li distribuirati terminološke podatke ili jezični model koji je izveden iz polaznoga skupa podataka?

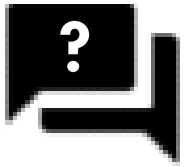
Ako izvedeni resurs sadrži znatne dijelove izvornih podataka (npr. duge citate, čitave odlomke itd.), onda se od nositeljaautorskih prava mora pribaviti licencija kako bi se izvedeni resursi moli distribuirati. Međutim, ako izvedeni resurs ne sadrži znatne dijelove izvornoga skupa podataka (npr. sadrži samo statističke podatke o broju riječi, pojedinačnih potvrda, kolokacija, itd.) takav se izvedeni resurs najvjerojatnije može distribuirati bez pribavljanja licencije od nositelja autorskih prava izvornoga skupa podataka. Ovo se obično rješava od slučaja do slučaja.

Pitanja: zaštita podataka

Što nam je činiti s resursima koji sadrže osobne podatke?

Ne moraju se svi osobni podatci anonimizirati.
Ako imate dvojbi što učiniti s resursima u kojima se nalaze osobni podatci, kontaktirajte tim ELRC-a. ELRC nudi pravnu pomoć kao i uslugu anonimizacije doniranih podataka.

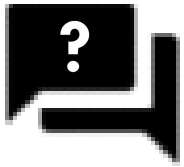
Q&A: Data protection



What is the scope of personal data restrictions? I have created a corpus of newspaper articles. They contain personal data such as names of people arrested, convicted for crimes and the like. Can I share this corpus? Should it be anonymized?

To be checked on a case-by-case basis.

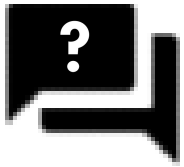
Pitanja: zaštita podataka



Imamo skup javno dostupnih dvojezičnih dokumenata iz tijela javnoga sektora, npr. iskazi interesa, pozivi za javnu nabavu, itd. Oni uključuju i osobna imena, npr. imena ravnatelja, direktora, članova odbora itd. Potpadaju li ti dokumenti pod ograničenje s obzirom na osobne podatke? Treba li anonimizirati takve dokumente?

Ovakvi su dokumenti nastali tijekom javnih aktivnosti tijela iz javnoga sektora, pa stoga ne pripadaju u područje zaštite privatnosti.

Pitanja: obrada podataka



Imamo podatke, ali nemamo načina i sredstava za pronalaženje relevantnih podataka i njihovu obradu.

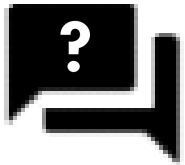
ELRC vam može pomoći u identificiranju relevantnih skupova podataka. ELRC također nudi uslugu obrade jezičnih podataka ustanovama iz javnoga sektora (pretvorbu podataka, uklanjanje oznaka, preoblikovanje, čišćenje, sravnjivanje, provjera metapodataka itd.) Moguća je i pomoć na licu mjesta tj. u vašoj ustanovi. Ove su usluge načelno besplatne.

Pitanja: obrada podataka

Imamo golemu količinu skeniranih PDF-ova. Možemo li zatražiti pomoć na licu mjesta? Hoćemo li dobiti natrag strojno čitljiv tekst (tj. rezultat OCR-a)?

Rezultati OCR-a nad skeniranim PDF-ovima znaju se značajno razlikovati u kakvoći (ovise o jeziku/jezicima, stanju papira, rezoluciji skena itd.). Neki skenirani PDF-ovi mogu biti korisni za daljnju obradu do razine strojno čitljivoga teksta i ELRC ih može obraditi dalje do razine paralelnih korpusa. I u takvim slučajevima ELRC nudi uslugu pomoći i procjene izvedivosti na licu mjesta.

Pitanja: uporabivost podataka



Većina naših podataka je brojčane naravi (npr. HNB, DZS) popraćenih s nešto teksta? Jesu li i oni uporabivi?

ELRC se ponajprije usredotočuje na tekstne podatke. Međutim, ako brojčani podatci sadrže i tekst koji bi ipak mogao biti koristan (npr. u slučaju dvo- ili višejezičoga teksta), onda bi i takvi podatci mogli ELRC-u postati uporabivi.

Pitanja: uporabivost podataka

Uzorak iz Hrvatskoga nacionalnoga korpusa je dostupan putem drugoga digitalnoga repozitorija (npr. CLARIN). Možemo li i njega proslijediti ELRC-u?

Samo ako je riječ o još neobjavljenim dijelovima toga jezičnoga resursa. Naime, ELRC ima pristup resursima iz drugih digitalnih repozitorija.