Info trend, 1.12.2009, str. 3, A3

Evidencijski broj / Article ID: 9878010
Vrsta novine / Frequency: Mjesečna
Zemlja porijekla / Country of origin: Hrvatska
Rubrika / Section:

Izradba tražilice za pretraživanja tekstova pravnih propisa Republike Hrvatske

Projekt CADIAL

Većina javnih tijela državne vlasti, kao što su parlamenti i javna administracija, indeksiraju (tj. dodatno označavaju ključnim pojmovima) sve pravne, administrativne i druge dokumente u svrhu lakšega pretraživanja. Za razliku od Interneta gdje su web-stranice u potpunosti indeksirane i gdje je omogućeno pretraživanje po punome tekstu propisa, službeni se dokumenti dodatno označavaju korištenjem koncepata (ključnih riječi) iz uređenoga pojmovnika

eGovernmenteHrvatska

Prednost takvoga dodatnoga označavanja jest u tome što korisnik može zadati upit, primjerice, „Ravnopravnost spolova", i naći sve bitne dokumente, iako se ta fraza doslovce ne pojavljuje u samome tekstu dokumenta. Kada je pojmovnik višejezični, korisnik može zadavati upite u jednome jeziku, a dohvaćati dokumente koji su napisani na drugome jeziku. Jedan primjer takvoga pojmovnika jest Eurovoc, službeni pojmovnik Europske unije koji postoji na tridesetak jezika pa tako i na hrvatskome jeziku.

Označavanje dokumenata ključnim riječima (tj. indeksiranje) zahtijeva znatan intelektualni napor jer podrazumijeva znanje o sadržaju dokumenta, kao i poznavanje pojmovnika kojim se dokumenti označavaju. Profesionalni dokumetarist označi približno oko 30-ak dokumenata dnevno. Taj je rad vrlo skup i vremenski zahtjevan, a zbog zahtjevnosti indeksiranja dokumenti su često neujednačeno označeni. Automatsko strojno indeksiranje dokumenata nameće se kao rješenje toga problema.

CADIAL (*Computer Aided Document Indexing for Accesing Legislation*) www.cadial.org međunarodni je projekt čiji je cilj bio razviti tehnologiju za automatsko označavanje dokumenata ključnim riječima iz pojmovnika Eurovoc usredotočujući se na tri glavna problema: (1) rješavanje problema djelotvornoga pretraživanja tekstnih podataka na morfološki bogatome hrvatskome jeziku, (2) razvoj interaktivnoga sustava za strojno potpomognuto indeksiranje dokumenata i (3) istraživanje različitih algoritama i njihovih parametara za automatsko indeksiranje dokumenata. Iako su rješenja problema automatskoga indeksiranja poznata, automatsko označavanje dokumenata ključnim riječima Eurovoca (tzv. deskriptorima) posebno je teško zbog velikoga broja deskriptora (oko šest tisuća hijerarhijskih organiziranih

Info trend, 1.12.2009, str. 4, A4

Evidencijski broj / Article ID: 9878010
Vrsta novine / Frequency: Mjesečna
Zemlja porijekla / Country of origin: Hrvatska
Rubrika / Section:

deskriptora) i zbog nejednolike raspodijeljenosti deskriptora dodijeljenih dokumentima. Projekt CADIAL osigurao je trajnu infrastrukturu za automatsko označavanje službene dokumentacije RH deskriptorima Eurovoca, kao i infrastrukturu za javnu dostupnost svih pravnih propisa Republike Hrvatske.

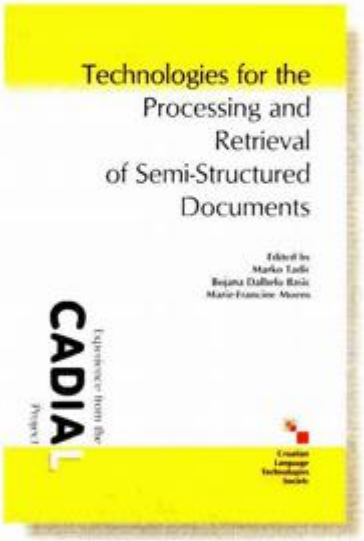
Projekt CADIAL financirale su Vlada Flandrije i Ministarstvo znanosti obrazovanja i športa Republike Hrvatske, a na projektu su radili znanstvenici i stručnjaci iz Katoličkoga Sveučilišta u Leuvenu, Fakulteta elektrotehnike i računarstva, Filozofskoga fakulteta i Hrvatske informacijsko-dokumentacijske referalne agencije. Rad interdisciplinarnoga međunarodnoga tima vodile su prof.dr.sc. Marie-Francine Moens i prof.dr.sc. Bojana Dalbelo Bašić, a sam je projekt izvrstan primjer uspjele primjene znanstvenih rezultata u praksi.

Sadržajno označeni pravni propisi kroz javno dostupnu tražilicu CADIAL stavljeni su na javnu uporabu tijelima javne vlasti, stručnjacima te svim građanima Republike Hrvatske.

Time je kroz projekt CADIAL dan doprinos ostvarenju prava građana na informacije, kao i doprinos približavanju Hrvatske standardima Europske unije.

eGovernmenteHrvatska

Knjiga „Technologies for the Processing and Retrieval of Semi-Structured Documents"



Knjiga „Technologies for the Processing and Retrieval of Semi-Structured Documents" sažima sve aspekte i sve rezultate projekta CADIAL. Knjigu su uredili prof.dr.sc. Marko Tadić (Filozofski fakultet), prof. dr.sc. Bojana Dalbelo Bašić (FER) i prof.dr.sc. Marie Francine Moens (Katoličko Sveučilište Leuven, Belgija). Knjiga je podijeljena u dva dijela. Prvi dio (Languages Technologies for Information Retrieval) pokriva jezične tehnologije za predobradu i pretraživanje polustrukturiranih dokumenata. Primjer su takvih tehnika postupci za svođenje različitih oblika riječi hrvatskoga jezika na jedinstven oblik u svrhu dohvaćanja dokumenata koji sadržavaju određeni pojam, bez obzira u kojem se morfološkome obliku (rod, broj, padež) riječ u tekstu nalazi . U tome su dijelu opisani i postupci ekstrakcije višerječnih izraza iz teksta. Drugi dio (Knowledge Technologies for Information Retrieval) pokriva područje tehnologija znanja za pretraživanje informacija koje obuhvaćaju postupke strojnoga učenja, automatske klasifikacije dokumenata i pretraživanja dokumenata. Knjiga je značajan doprinos razvoju jezičnih tehnologija za hrvatski jezik, no iako su opisane i razvijene tehnologije primijenjene na zbirku pravnih dokumenata na hrvatskome jeziku i tražilicu **cadial.hidra.hr**, one su univerzalne i primjenjive na bilo koju drugu zbirku dokumenata koju je potrebno automatski klasificirati, pretraživati, izdvajati višejezične izraze ili svoditi različite oblike riječi hrvatskoga jezika u toj zbirci na njihov natuknički oblik (lemu).

Bojana Dalbelo Bašić