# RAZŠIRITEV ISKALNIKA Z ORODJI ZA SEMANTIČNO ISKANJE V SLOVENSKI POLICIJI

*Mladen Tomaško*[1], *Dunja Mladenić*[2]
[1]Služba generalnega direktorja policije, Ministrstvo za notranje zadeve RS
Litostrojska cesta 54, 1000 Ljubljana, Slovenia
Tel: +386 1 4773419; fax: +386 1 4251038
e-mail: mladen.tomasko@policija.si
[2]Institut Jožef Stefan in Mednarodna Podiplomska šola Jožefa Stefana,
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: dunja.mladenic@ijs.si

**POVZETEK**

Spreminjanje katere koli komponente v obsežnih informacijskih sistemih z velikim številom uporabnikov, je predvsem zaradi inercije takšnih sistemov in pogosto tudi tehnične zahtevnosti, težavna naloga. Uspešnost rešitve je odvisna od natančne preučitve navad uporabnikov, preproste uporabe, čim bolj neopazne integracije in kvalitetne podpore (usposabljanja, pomoči v realnem času, …) Pri tem je še posebej pomembno, da je rešitev vgrajena v obstoječi sistem tako, da čim manj spreminja dosedanji način dela.Pri tehnični izvedbi lahko naletimo na težave s podatki, ki so neustrezno razvrščeni, vsebujejo veliko napak, ali pa so nepopolni. To predstavlja dodatne izzive pri uporabi orodij semantičnih tehnologij. Po drugi strani pa je takšna rešitev dodana vrednost, katere prednosti se pokažejo na daljši čas. Zahteva pa začetno privajanje uporabnikov na spremembe (v tem času storilnost lahko celo upade), dolgoročno pa računamo na opazen prihranek časa in kakovostnejše poslovanje. Prispevek predlaga pristop za razširitev standardnega iskalnika s pomočjo metod semantičnih tehnologij. Predstavljena rešitev je trenutno pred fazo testiranja v vsakdanjem delovnem procesu.

## 1 UVOD

Ministrstvo za notranje zadeve republike Slovenije uporablja za svoje pisarniško poslovanje IBM Lotus Notes[5]. Sistem je razbit na več kot 40 samostojnih strežnikov, ki so do pred kratkim omogočali iskanje le po lokalnih vsebinah. Zaradi tega smo razvili iskalnik, ki išče po vseh Lotus Notes zbirkah in omogoča tudi ustrezno implementacijo varnostnega pravilnika, ki omogoča za vsakega posameznika definiranje ustreznih dostopov do dokumentov[3]. Dostop do večine dokumentov je namreč omejen, ker so v njih podatki, ki jih lahko obdelujejo le osebe, ki za to imajo zakonsko podlago. Že pri integraciji novega iskalnika se je pokazala potreba po dopolnitvi obstoječega iskalnika z orodji, ki bi omogočala učinkovitejše razvrščanje dokumentov, oz. bi omogočila prikaz tudi tistih dokumentov, ki so neposredni povezavi z uporabnikovim iskanjem, čeprav jih uporabnik ni direktno iskal. Pomembno pri vsem tem je, da nam nova rešitev omogoča nadaljnje izboljšave sistema zato smo v predlagani rešitvi za dopolnitev funkcij iskalnika uporabili odprtokodni program Lucene/Sorl [6][7], za naprednejše delo s podatki pa prav tako odprtokodno rešitev iz področja semantičnih tehnologij OntoGen[1]. V prispevku so na kratko opisani koraki pri razvoju te rešitve.

## 2 PREDSTAVITEV PODATKOV

Iskalnike uporabljamo za preiskovanje različnih zvrsti podatkov, ki so lahko zapisani v podatkovnih bazah, pa tudi v manj urejenih zbirkah podatkov ali na svetovnem spletu. V prispevku smo za primarni vir podatkov vzeli zbirke IBM Lotus Notes in to tisti del, ki sestavlja sistem SPIS (SRC.SI pisarniški informacijski sistem) v katerem so shranjeni vsi dokumenti, katerih pomembnost je takšna, da morajo biti zabeleženi in shranjeni v informacijskem sistemu MNZ in Policije.

### 2.1 Lastnosti podatkov

Sistem IBM Lotus Notes je razpršen na 42 strežnikih. V njem so združene podatkovne zbirke za pisarniško poslovanje, podatkovne zbirke elektronske pošte in veliko namenskih podatkovnih zbirk. V naši rešitvi smo se omejili le na podatkovne zbirke, ki vsebujejo dokumente, ki nastajajo pri pisarniškem poslovanju, se pravi izhodne, vhodne in lastne dokumente. Količina podatkov v sistemu pa se že meri v stotinah terabajtov. Podatkovne zbirke vsebujejo več kot pet milijonov dokumentov. Podatki v njih so shranjeni v strukturirani in nestrukturirani obliki. Strukturirani podatki vsebujejo vse osnovne podatke o avtorju dokumenta, naslovniku, področju, ki ga dokument zajema, datumu in času nastanka, podatke, ki jih rabi varnostni pravilnik. Pri takšni količini podatkov se pojavlja tudi veliko napačno razvrščenih dokumentov, ki otežujejo iskanje s klasičnim iskalnikom. Te napake so shranjene v strukturiranih poljih,. Po navadi je napačno vneseno področje, včasih tudi naslovnik, posebej ko je dokument

naslovljen na službo ki je sestavljena iz več sektorjev in oddelkov. V nestrukturiranih podatkih – nas zanima predvsem polje v katerem se nahaja celotno besedilo dokumenta – je teh napak praviloma manj, čeprav je besedilo zapisano skupaj z glavo, naslovnikom, in ostalimi podatki, ki sestavljajo dokument pripravljen za tiskanje in pošiljanje po navadni pošti. Prav ta podvojenost podatkov, ki so drugače že zapisani v strukturiranih poljih, vsebuje pa jih tudi besedilo, otežijo pravilno grajenje ontologije in jih je bilo treba na določen način izločiti. Del smo jih izločili že pri izvozu podatkov iz podatkovnih zbirk. Tu smo izločili strukturirana polja, pustili smo le tista, ki smo jih pozneje rabili zato, da smo lahko preverjali kako uspešno je OntoGen razvrščal dokumente.

Izločanje podatkov iz dokumentov, ki so bili v nestrukturiranih poljih je bilo bolj zahtevno. V pomoč nam je bilo, da je v večini dokumentov besedilo urejeno na enak način. Tako smo lahko določili točki med katerima je območje besedila dokumenta in smo to besedilo shranili v datoteke. Del podatkov smo pa očistili tako, da smo pripravili seznam besed (predvsem kratic), ki niso relevantne za naš problem (n.pr. MNZ, UKP, UUP, …) in jih OntoGen ni upošteval pri postopku razvrščanja.

### 2.2 Varnostni pravilnik

V Policiji je varnost podatkov izrednega pomena. Zato mora vsaka rešitev upoštevati stroga varnostna merila, ki omogočajo nadzor dostopov do dokumentov in preprečujejo dostop nepooblaščenim uporabnikom. Sistem Lotus Domino vsebuje zmogljiv varnostni model, ki dostope do podatkov preverja na več ravneh [4]. Varnostni model sestoji iz šestih nivojev:

1. **varnost omrežja,**
2. **overovitev uporabnikov,**
3. **dostop do strežnika,**
4. **zaščita zbirke podatkov,**
5. **varnost oblikovnih elementov,**
6. **omejitev dostopa do dokumentov.**

Močan varnostni model je nujen zaradi velike količine osebnih podatkov, ki jih lahko vidijo le pooblaščeni uporabniki. Poleg tega pa zbirke vsebujejo tudi tajne podatke, ki so še posebej varovani z močnim šifriranjem.

## 3 PREDLAGANA REŠITEV

Na Sliki 1 je grafično podan predlog rešitve. Že prej smo omenili, da je osnovna zahteva bila dodati obstoječemu iskalniku rešitev, ki bo uporabljala semantična orodja in ponudila uporabnikom podobne dokumente oz. jim pomagala pri njihovem razvrščanju. Kot je iz Slike 1 razvidno je rešitev zgrajena tako, da prvi del iskanja opravi osnovni iskalnik v Lotus Notes, ki je zgrajen na podlagi odprtokodne rešitve Lucene/Sorl [6][7]. Komercialne rešitve so se izkazale za prezapletene in prezahtevne za prilagajanje specifičnemu okolju Policije. Zato se je prej omenjena odprtokodna rešitev izkazala kot ekonomsko najbolj upravičena. Enostavna implementacija in učinkovito delovanje Lucene sta glavna razloga, da je Lucene uporabljen kot rešitev za iskanje po zbirkah Lotus Notes. Iskalnik išče po indeksiranih podatkih, indeksiranje pa se ves čas odvija v ozadju. Tukaj je treba omeniti, da je bilo treba v iskalnik vgraditi varnostni pravilnik, ki omogoča nadzorovan dostop do dokumentov.

Drugi del "iskanja" oz. pravilneje rečeno razvrščanja podatkov opravi prav tako odprtokodna rešitev OntoGen [2]. OntoGen je polavtomatski in podatkovno usmerjen urejevalnik ontologij. Osredotoča se na urejanje tematskih ontologij (sklop tem, povezanih z različnimi vrstami odnosov). Sistem združuje metode podatkovnega rudarjenja, z učinkovitim uporabniškim vmesnikom kar pomeni manj porabljenega časa in preprostejše delo. Na ta način zapolnjuje vrzel med zapletenimi orodji za urejanje ontologij in ga lahko uporabljajo strokovnjaki na posameznih področjih, ki nimajo nujno znanja potrebna za gradnjo ontologij. S pomočjo OntoGena smo pripravili ontologijo, ki nam pomaga pri razvrščanju dokumentov, predlaga podobne dokumente ko pripravljamo nov dokument, ideja pa je, da bi nam ponujal besede, ki jih lahko uporabimo v iskalniku, da bi dobili boljše zadetke. Na nek način, bi nas učil kako bolje izrabiti iskalnik.

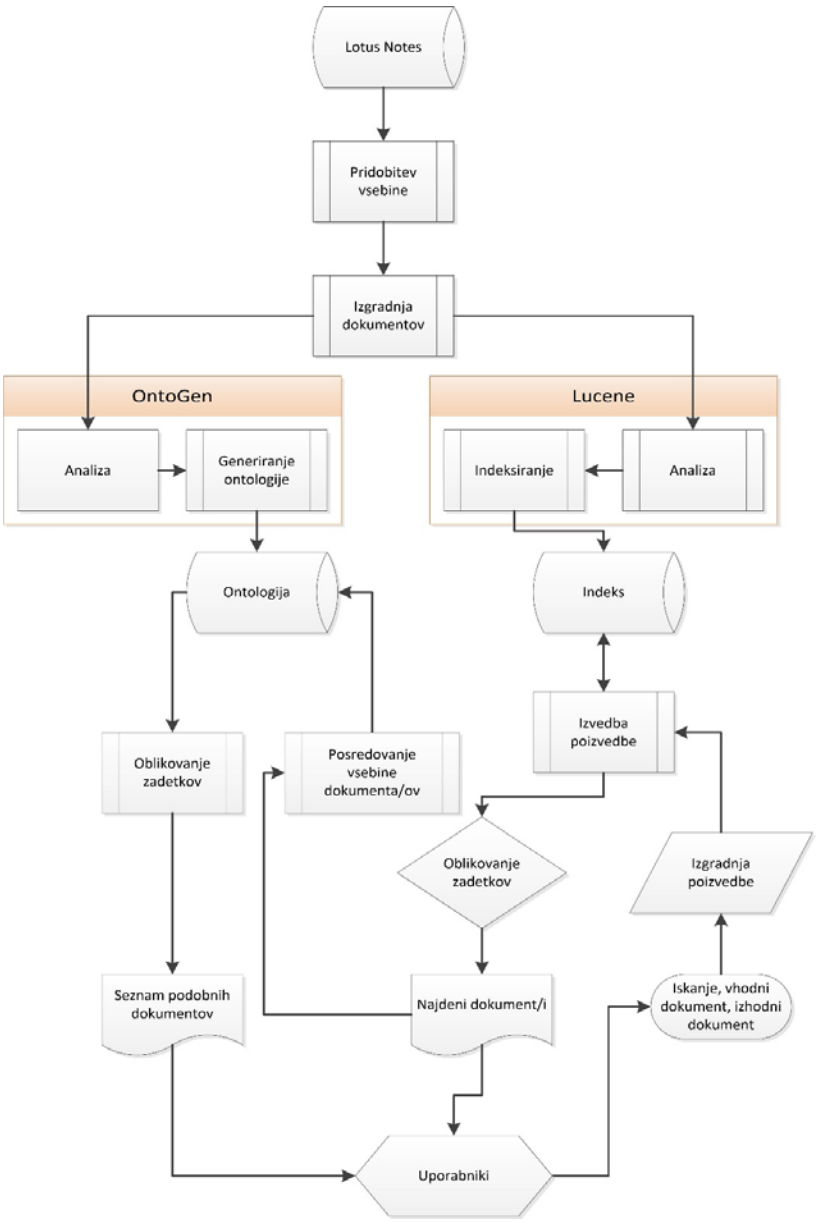
### 3.1 Priprava podatkov

Zbirka Lotus Notes vsebuje več kot pet milijonov dokumentov. Za pripravo ontologije smo izmed njih naključno izbrali 3000 dokumentov, izmed katerih smo 300 uporabili za gradnjo ontologije, na preostanku pa smo preverjali kako uspešen je naš model.

V dokumentih se pojavlja veliko število kratic, ki večinoma predstavljajo organizacijske enote ministrstva ali so okrajšave nekaterih mednarodnih organizacij. Te kratice so velikokrat povzročile napačno razvrščanje dokumentov in je nekatere izmed njih bilo treba izločiti. Začetek gradnje ontologije je bil zaradi tega poln poskusov priprave ustrezne oblike dokumentov in iskanja besed, ki jih je bilo treba uvrstiti na stop seznam (izločiti iz obdelave). Pokazalo se je, da ključ k uspešni gradnji ontologije predstavljajo dobro izbrani in dobro pripravljeni podatki.

### 3.2 Gradnja ontologije

Gradnja ontologije je potekala v več korakih. Na začetku se je pokazalo, da podatki, ki smo jih ponudili OntoGenu niso bili dovolj dobro izbrani (kljub temu, da smo jih izbirali naključno niso vsebovali dokumentov iz vseh področij, ki jih zajema delo Policije), zato smo morali postopek izbire dokumentov nekajkrat ponoviti. Tudi kakovost priprave podatkov na začetku ni bila ustrezna in smo morali izločati neustrezne besede in izraze. Težava je bila tudi s pretvorbo različnih kodnih zapisov. Vsi dokumenti niso bili v enakih kodnih zapisih in jih je bilo treba poenotiti. Po večkratnih




Slika 1. Prikaz arhitekture predlagane rešitve, pri čem je iskanje razdeljeno na dva dela: Lucene opravi osnovno iskanje po bazi, OntoGen omogoča semantično iskanje za namen pridobivanja podobnih dokumentov.

poskusih, so se pričeli kazati zametki ontologije, ki jo je bilo treba še ročno optimizirati. Težavo predstavlja velika razvejanost organizacije MNZ in prepletenost delovnih področij. Npr. kaznivo dejanje lahko obravnava navaden policist na policijski postaji, policist kriminalist, kriminalist iz policijske uprave ali kriminalisti generalne policijske uprave. Razlika je predvsem v tem ali kaznivo dejanje zajema lokalno, regionalno ali državno raven. Ontologija precej natančno sledi tem pravilom in omogoča uspešno razvrščanje dokumentov glede na vsebino.

Naša rešitev v osnovi sledi organizacijski strukturi MNZ in delovnim procesom na vseh nivojih s tem, da so pri gradnji ontologije bile upoštevane določene specifičnosti posameznih področij (predvsem v primerih, ko gre za prepletanje delovnih procesov na različnih nivojih). Pravila, ki so vgrajena v ontologijo omogočajo ustrezno identifikacijo dokumentov in v prihodnje bodo upoštevali tudi umeščenost uporabnika, ki je sprožil iskanje, v organizacijsko strukturo.

### 3.3 Oblikovanje zadetkov poizvedbe

Naša rešitev in iskalnik delujeta neodvisno drug od drugega. Iskalnik Lucene razvršča zadetke po tem kako ustrezajo iskanim besedam in kako pogosto se te besede pojavijo v najdenih dokumentih. Pri iskanju ne upošteva druge ključne besede in izraze, ki se pojavljajo v najdenih dokumentih. Naša rešitev ponudi podobne dokumente, ki niso neposredno vezani na iskane izraze, temveč se na posebnem seznamu

pokažejo še dokumenti, ki se tematsko ujemajo s tistimi, ki jih je našel iskalnik. Poleg tega ima uporabnik možnost prikaza podobnih dokumentov v času, ko sam sestavlja nek dokument, oziroma mu program sam predlaga umestitev novega dokumenta v klasifikacijsko strukturo.
Pravila zapisana v ontologiji nam pomagajo poiskati podobne dokumente, oziroma predlagati razvrstitev obstoječih. Uspešnost razvrščanja je odvisna od kompleksnosti delovnega področja iz katerega je dokument. Na bolj kompleksnih je stopnja učinkovitosti nekoliko nižja kar pomeni, da bo treba ontologijo še nekoliko dodelati. Tega se bomo lotili po zaključenem testiranju, ko bomo imeli na razpolago več podatkov.

## 4 TESTIRANJE UČINKOVITOSTI MODELA

Merjenje učinkovitosti smo zasnovali na dveh nivojih: vprašalnik (bolj subjektivno) in samodejno spremljanje uporabnikov (bolj objektivno). Struktura testiranih uporabnikov odraža strukturo celotne organizacije. Vzorec obsega 30 do50 uporabnikov, ker bi pri manjšem številu težko dosegli pravilno zastopanost vseh razredov.

### 4.1 Anketiranje uporabnikov

Vprašalnik je sestavljen iz dveh sklopov. V vsakem sklopu so še kontrolna vprašanja, ki omogočajo, da nadziramo kakovost odgovorov. Testno obdobje bo trajalo od 14 – 21 dni, tako, da bodo v tem obdobju zajeta najpomembnejša opravila, ki se pojavljajo na mesečni ravni

Vprašalnik je, vsaj deloma, subjektivna ocena uporabnika kako je doživljal delo z novimi funkcionalnostmi pri svojem vsakdanjem delu, zato je treba rezultate, ki preveč odstopajo od povprečja na ustrezen način dodatno analizirati in ugotoviti zakaj je do teh odstopanj prišlo.
Analiza rezultatov je izredno pomembna ker nam omogoča ugotoviti ustreznost rešitve, oceniti prihranke zaradi izboljšanih delovnih procesov in v končni fazi oceniti ali je rešitev zrela za implementacijo v celoten sistem. Če se izkaže, da so na določenem področju pomanjkljivosti bo treba pristopiti k dodatnim izboljšavam rešitve in ponoviti postopek testiranja.

### 4.2 Samodejno spremljanje obnašanja uporabnikov

Za potrebe spremljanja obnašanja uporabnikov bomo pripravili dodatno rešitev, ki bo zapisovala klike na povezave, število iskanj, število obdelanih dokumentov, čas potreben za posamezen dokument, morebitne neustrezne klike in podobno.
Ti podatki nam bodo v pomoč pri končnem ocenjevanju rešitve in smiselnosti njene implementacije v sedanji obliki. Dobljene rezultate bo treba pred končno analizo ustrezno ovrednotiti in jih pripraviti za medsebojno primerjavo. Vsekakor bo treba upoštevati tudi zahtevnost delovnega mesta, zahtevnost pripravljenih dokumentov in ne samo njihovo število ali čas potreben za obdelavo.

## 5 ZAKLJUČEK

Predlagali smo razširitev obstoječega iskalnika po internih bazah Slovenske policije. Pri tem smo uporabili semantične tehnologije, s pomočjo katerih smo vnaprej zgradili ontologijo vsebin dokumentov, ki so shranjeni v obstoječih bazah. Priprava ontologije za organizacijo s tako zapleteno organizacijsko strukturo zahteva veliko ročnega dela in dobro poznavanje tako organizacijske strukture kot delovnih procesov. Posebno težavo predstavlja prepletanje delovnih procesov med različnimi nivoji, tako da je včasih težko izluščiti kateri način dela je najbolj pravilen. Zato ko je ontologija enkrat v grobem zgrajena, sledi zahtevno dopolnjevanje in optimiziranje. Kljub temu, da je ta proces časovno precej potraten se na koncu obrestuje z izboljšanimi delovnimi procesi, povečano učinkovitostjo in prihranki na nivoju celotne organizacije.
Kljub temu, da je predlagana rešitev še v fazi testiranja, brez natančno definiranega varnostnega pravilnika, se že vidijo razlike v uspešnosti razvrščanja dokumentov. Koliko se bo to pokazalo kot izboljšanje delovnih procesov in s tem povezanih finančnih učinkov bo pokazalo spremljanje rezultatov dela v daljšem časovnem obdobju. Takrat bo narejena tudi analiza, ki bo ovrednotila razmerje med vloženimi sredstvi in morebitnimi prihranki. Po drugi strani pa je vsak poskus uporabe naprednih tehnologij v tako kompleksnih sistemih dobrodošel, ker na praktičnih primerih pokaže na katerih vse področjih te tehnologije lahko uporabimo. S tem lahko pri menedžmentu dobimo možnost nadaljnjega razvoja informacijskega sistema, ki bo primernejši tako z vidika zaposlenih, kot z vidika uporabnikov uslug policije.

## Viri


[1] B. Fortuna, M. Grobelnik, D. Mladenic: Semi-automatic Data-driven Ontology Construction System. In: *Proc. of the 9th International multi-conference Information Society IS-2006*, Ljubljana, Slovenia (2006).

[2] B. Fortuna, M. Grobelnik, D. Mladenic: Semi-automatic Ontology Editor, In*: Proc. of the 12th International Conference on Human-Computer Interaction,* Beijing , China (2007).

[3] P. Skale. Načrtovanje in razvoj iskalnika za potrebe policije. *Msc Thesis*. Fakulteta za računalništvo in informatiko. Univerza v Ljubljani. 2012.

[4] IBM, *Inside Lotus: The Architecture of Notes and the Domino Server*, IBM Press, 2000.

[5] Zgodovina razvoja Lotus Notes in Domino. http://www.ibm.com/developerworks/lotus/library/ls-NDHistory/. (dostop avgust 2014)

[6] Lucene. http://lucene.apache.org. (dostop julij 2014)

[7] Apache Solr. http://lucene.apache.org/solr. (dostop avgust 2014)

[8] Zgodovina razvoja Lotus Notes in Domino. http://www.ibm.com/developerworks/lotus/library/ls-NDHistory/ (dostop julij 2014)