Pretraživački sistem specijalizovane namene

Julijana Mirčevski, Nikola Popović

***Sadržaj* — Rad opisuje realizaciju prototipa sistema za pretraživanje namenjenog upravljanju tehničkom dokumentacijom. Implementacija je izvedena u database software for Windovs (FileMaker), jeftinom komercijalnom programskom paketu. Prototip je razvijen na primerima tehničke dokumentacije prospekata za automatske prekidače sa ciljem da se obezbedi njihova višejezička interpretacija.**

***Ključne reči* — Document Management System, Technical documentation, Information retrieval, Multilanguage systems**

I. Uvod

Sistemi za upravljanje dokumentima (Document Management Systems – DMS) [1] predstavljaju složene programske komplekse veoma potrebne ali visoko komercijalno vrednovane. Komercijalne verzije ovih sistema uglavnom su koncipirane za *enterprise* nivo (većih preduzeća ili državnih organa) što podrazumeva veliki broj korisnika, visoki *data storage* zahtevi i strogi kriterijumi zaštite.

Svrha ovog rada je prikaz prototipa programa za DMS zasnovanog na koncepciji Web Information System-a (WIS) i korišćenja nekih elemenata „programiranja agenata“. Realizovan je primenom jeftinog komercijalnog softvera, FileMaker Pro v10 Pro koji spada u klasu RAD (Rapid Application Development) alata. [1], [2]. Ovakav izbor učinjen je sa namerom da se minimizira potreba za programiranjem. Niska cena baznog softvera je od značaja, zato što trenutna situacija u ovoj oblasti obuhvata dve krajnosti – ili je softver koji se koristi predvidjen za rad na nivou preduzeća (državnih organa i sl.) ili se zahteva ogromna dodatna količina programiranja od strane programera sa specijalizovanim znanjima. Obe margine su vezane visokom ukupnom cenom vlasništva (TCO – Total Cost of Ownership) proizvoda što je i glavni uzrok slabe zastupljenosti DMS sistema u nas. Praktična saznanja iz upotrebe ovakvih sistema pokazuju da se o realnom korisniku ne vodi mnogo računa – u svim slučajevima se zahteva da korisnik svoje zahteve prilagodjava karakteristikama implementiranog sistema što ne retko dovodi do odustajanja od upotrebe. Ovaj rad je nastao kao rezultat ispitivanja mogućnosti da se najznačajnije funkcije WIS zasnovanih sistema realizuju bez velikih inicijalnih troškova [1], a da se na bazi korišćenja prototipa obavlja evaluacija procesa eventualne nabavke i implementacije složenijih sistema kada se za tim ukaže stvarna potreba.

Julijana Mirčevski, Asocijacija " Jelena Anžujska", ,Beograd, Srbija (telefon: 381-64-6100100, e-mail: julijana@afrodita.rcub.bg.ac.rs).

Nikola Popović (autor za kontakte), Ministarstvo spoljnih poslova Republike Srbije, Kneza Miloša 24-26, 11000 Beograd, Srbija (telefon 381-11-306-8154, e-mail: nikolapopovic1949@gmail.com).

II. Analiza problema

Primarni cilj rada je razvoj aplikacije za evidenciju i po potrebi skladištenje skupa medjusobno povezanih dokumenata različitog tipa i povezivanje dokumenata u skladu sa tekućim potrebama korisnika. Jasno je da ovakvi sistemi već postoje na tržištu i/ili u OpenSource verziji, medjutim u našim uslovima njihova primena je ispod minimum (EMC Documentum, Alfresco itd.).

Neslužbeni podaci o nekoliko lokalnih implementacija Enterprise Wide DMS ukazuju da većina njih ili nije u funkciji ili barem ne na planiranom nivou iz sledećih razloga:

- Komplikovano korišćenje za standardnog korisnika,
- Unificiran interfejs koji traži prilagodjavanje korisnika sistemu u suviše velikom obimu,
- Najčešće je primenjen model jedinstvenog repozitorijuma dokumenata
- Efikasan (?) ali složen sistem zaštite dokumenata najčešće uslovljen jedinstvenim repozitorijumom dokumenata
- Višejezičnost podržana ali na nivou pretraživanja punog teksta dokumenta
- Sistemi za upravljanje dokumentima nemaju kapacitet MS Office paketa (interakcija različitih tipova aplikacija) koji se očekuje u državnoj upravi ili malim preduzećima
- Pod dokumentima se podrazumevaju samo klasične forme dokumenata kao što je MS Word dokument odnosno DMS ne vidi baze podataka kao tip dokumenta.

Tehnološki problem migracije podataka i funkcija, a posebno cena migracije, retko se pominje.

Izabrani program FileMaker Pro v10, pored niske cene po radnoj stanici, raspolaže važnim pogodnostima:

- efikasno spaja koncept relacione baze podataka sa dobro razvijenim funkcijama Information Retrieval modela
- omogućava istovremeno korišćenje modela klasične i Web aplikacije bez potrebe za dodatnim programiranjem u jednostavnijim slučajevima primene.

Uobičajeno korišćenje Data Storage sistema u sklopu DMS sistem pored svih svojih prednosti značajno ograničava mobilnost programa i upotrebu podataka čak i u uslovima primene virtuelnih mašina. Slična primedba može da se postavi i sa aspekta veoma čestih reorganizacija državne uprave [3], i redistribucije organizacionih jedinica na različitim lokacijama. U poslednje vreme na tržištu je prisutan sistem FileMaker Go, predvidjen za korišćenje putem iPhone-a i iPod-a.

Treba naglasiti da je projektovanje upotrebe DMS sistema većih dimenzija uslovljeno definisanjem jasnih

zahteva korisnika. Za domen primene posmatran u ovom radu, to predstavlja značajnu smetnju. Često se javljaju zahtevi za hitnim analitičkim radom sa dokumentima nepoznatog sadržaja, unapred nepoznatih veza sa drugim dokumentima. Lokacija rada ne mora da odgovara korišćenju velikih sistema (terenski poslovi, SCADA sistemi itd.). Praktično iskustvo pokazuje da se korisnici velikih sistema sudaraju sa restrikcijama softverskih sistema kako u pogledu manipulacije dokumentima ili delovima dokumenata, mogućnošću njihove agregacije i/ili konsolidacije, tako i ograničenjima organizacionog vrste. Posebno su značajna ograničenja uslovljavanjem smeštanja dokumenata istog tipa i nemogućnost linkovanja prema nezaštićenim domenima podataka - internet ili drugi segmenti sistema van kontrole korisnika. Za velike agregacije dokumenata redovno postoje definisani kriterijumi i pravila za obeležavanje dokumenata i opis sadržaja (meta podaci). U analitičkom radu često je prisutna upotreba „folksonomije“ tj. ad-hoc formirane klasifikacione šeme koja je primenljiva samo za zadati skup dokumenata u zadatoj situaciji. Za odredjeni problem može da se formira i više različitih „folksonomija“ prema konkretnim potrebama korisnika (tenderi, havarije i sl.).

Veliki broj primera potrebe za ovakvim redukovanim ili *small scale* sistemom može da se nadje u organima državne uprave kao naprimer baza podataka štampe, u cilju regularnog praćenja ko i šta piše i u vezi kojih tema. Drugi primer su izjave zvaničnika vlade u medijima i na konferencijama za štampu kada je za potrebe analitike neophodno redovno pratiti i čuvati multimedijalnu dokumentaciju. Treći slučaj, obradjen ovde, je vezan za tehničku dokumentaciju koja mora uvek da bude ažurna. Poznati su slučajevi urgentnih situacija sa težim posledicama zato što izvedeno stanje instalacija nije odgovaralo projektnoj dokumentaciji (tj. nije izvršeno ažuriranje dokumentacije prema izvedenom stanju). Slična situacija se javlja i kada naše kompanije treba brzo da razumeju, obrade i odgovore na zahteve medjunarodnih tendera U posmatranom kontekstu potrebno je naglasiti aspekt mobilnosti baze podataka odnosno koncept mikro informacionog sistema kako bi podaci bili dostupni na terenu, u toku pregovora i sl.

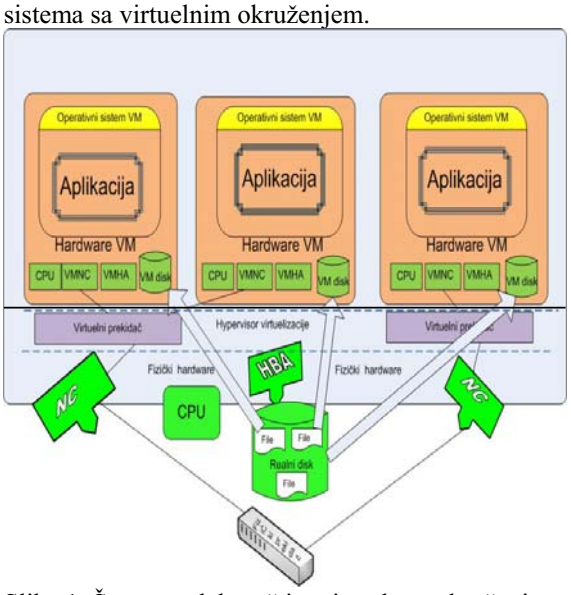
Aplikacija "Rečnik" u fazi razvoja implementirana je i u virtuelnom okruženju u cilju izrade što mobilnije varijante. Neophodno je ukazati na značaj zaštite aplikacije budući je namenjena za korišćenje i u "terenskim uslovima" gde na raspolaganju nema mnogo drugih informatičkih resursa a obično ni stručnog nadzora u domenu sigurnosti sistema.

Analiza raspoloživih informacija, kako u pisanoj literaturi tako i na internetu, ukazuje na postojanje slabosti u domenu sigurnosti informacija ali su uglavnom u pitanju komentari nakon korekcije nedostataka od strane proizvodjača virtuelnih mašina.

Mogućnosti napada virusima su realne ukoliko se pristupa internetu na uobičajeni način. Prenos virusa na host mašinu je takodje sasvim realna opasnost u slučaju deljenih foldera (posebno se odnosi na mašine sa Windows operativnim sistemom). Zasada nisu nadjeni podaci o tome da je guest mašina „zaražena“ preko host mašine a da nije bila uspostavljena bilo kakva druga konekcija.

Virtuelne mašine olakšavaju uvođenje rigidnijih mera zaštita podataka [3], [4], posebno sa aspekta zaštite privatnosti u državnim organima. Naime, samo po sebi uvodjenje virtuelne mašine znači uvodjenje još jednog sloja koji treba da se probije pri upadu u sistem. Nezavisno od slabosti, situacija se značajno komplikuje za napadača. Isto važi i za uvodjenje različitih operativnih sistema u kombinaciji Linux→VM→Windows.

Virtuelna mašina dozvoljava fizičku zaštitu tako što se virtuelna mašina instalira na eksternom disku koji se kada nije u upotrebi isključuje i napr. zatvara u kasu (slično kao i laptop računari). Vid logičke zaštite je da se disk sa podacima koje treba štititi „šeta“ po skupu računara i tako otežava njegovo lociranje od strane neovlašćenog korisnika. Na slici 1. prikazan je jedan model zaštite sistema sa virtuelnim okruženjem.



Slika 1. Šema modela zaštite virtuelnog okruženja

Isporučioci efikasnog zaštitnog software-a zahtevaju od korisnika da kupi i njihov hardware, strogo posvećen izvršavanju sigurnosnih programa i pritom vrlo skup. To onda višestruko povećava cenu virtuelizacije odnosno snižava benefit na koji se računa prilikom uvodjenja virtuelizacije [3].

Virtuelizacija uvodi nova pitanja na temu zaštita sigurnosti informacionog prostora posebno u domenu državnih organa. IT profesionalci upozoravaju da virtuelizacija ne može da se širi više nego što može da bude kontrolisana sigurnost njenog funkcionisanja. Virtuelno okruženje kojim može efikasno da se upravlja zahteva da bude gradjeno na jasno definisanim ciljevima, arhitekturi i postavom upravljanja prema nivou performansi [3], [4].

III. MODEL

Model se sastoji od baze podataka koju čine tabele sa tekstovima uputstava i prevoda iz zadatog stručnog domena (prekidači), kao tabele relacija odnosno URL (Unified Resource Locator) prema internim i eksternim linkovima. Korisnika interesuje naprimer prevod ili tumačenje nekog dela teksta. Prvo se selektuje posmatrani deo teksta i zatim aktivira proces pretraživanja baze podataka rečnika. Rezultat može da bude sledeći:

1) u rečniku ne postoji selektovani deo teksta,
2) postoji jedan prevod ili
3) postoji više prevoda.

U slučaju broj 1 potrebno je da se izmeni selekcija teksta i ponovo pokuša sa pretraživanje, naprimer usled moguće greške u ispisivanju teksta. U slučaju da je traženi tekst važan za korisnika potrebno je da se izradi odgovarajući ulaz u rečnik. Ili da se napravi referenca na odgovarajući alternativni ulaz rečnika odnosno sinonim.

U slučaju broj 2 korisnik može da posle uvida u prevod i definiciju pridruži ulaznom tekstu selektovani pojam kao tag - oznaka ili klasa kojoj pripada pojama. Tagiranjem se uspostavlja direktna veza segmenta ulaznog teksta i rečnika preko liste tagova. Tagiranjem se formira osnova za izradu „folksonomije".

U slučaju broj 3 korisnik treba da izabere jedan pojam iz rečnika i da ga pridruži selektovanom tekstu ulaznog dokumenta. U listi tagova može da se pojavi i više pojmova iz rečnika ukoliko su srodni kao što su sinonimi i drugi tipovi veza.

U skladu sa navednim primerima tagiranje i „folksonomiju" formiramo sa osnovnim rečnikom pojmova koji se javljaju u tehničkoj dokumentaciji. Ali isto tako formiramo prema potrebi i paralelne šeme tagiranja koje mogu da se odnose naprimer na: uočene kvarove kod prekidača, vek trajanja, izdržavanje preopterećenja, definisanje prekidača koji mogu da budu zamena zadatog tipa, odstupanja u dimenzijama i načinu montaže itd. Korisno je imati na raspolaganju distribuciju prekidača po fizičkim lokacijama što obezbedjuje i odredjene elemente za projektovanje nivoa rizika. Ovakvu koncepciju praćenja tehničke dokumentacije po zadatim elementima može uglavnom da obezbedi WEB tehnologija [1], [4], pošto osnovna baza podataka može da bude i centralizovana i/ili decentralizovana u zavisnosti od zadate situacije.

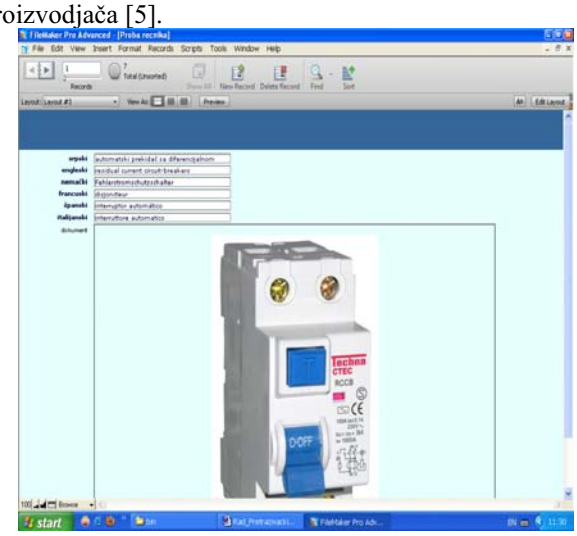
IV. STRUKTURA REČNIKA

Rečnik u ovom slučaju nije samo jednostavna tabela pojmova sa prevodima. Ulaz u tabeli rečnika može da bude reč, sintagma, rečenica, apstrakt, hiperlink kao i slika ili zvuk (naprimer video zapis sa uputstvima kako da se nešto uradi – uputstvo za montažu, ili snimak nečijeg govora sa TV, presek nekog uredjaja itd.). Hiperlink može da bude veza prema drugoj bazi podataka ili prema internetu, dislociranoj video prezentaciji itd.

Tabela 1. Prikaz osnovne strukture rečnika

ID1	Termin 1 Prevod 1 Prevod 2.	Definicija 1 Prevod Def. 1	...	Rečnik sinonima
ID2	Termin 2 Prevod 1.	Definicija 2 Prevod Def..1 Prevod Def.2	Hyperlink 2 Hypellink2a Hypelink2b	Internet parametri
ID3	Termin 3 Prevod 1 Prevod 2.	Definicija 3 Prevod Def. 1 Prevod Def. 2	Hyperlink 3	Parametri DBMS

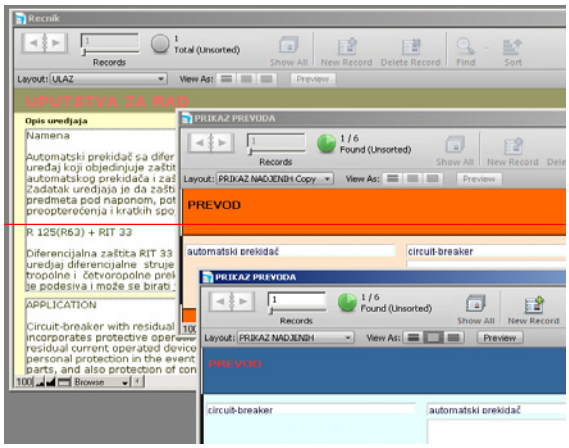
Svaki termin mora da ima barem jedan ekvivalent na stranom jeziku (uobičajeno engleski ili bilo koji drugi posmatrani strani jezik u zavisnosti od potrebe posla odnosno od isporučioca opreme). Prevodi su smešteni u posebne tabele sa identičnom strukturom, Slika 2. Posmatrano sa aspekta integracija u Evropsku uniju višejezične baze podataka postaju nužnost i prioritet. Primera radi u slučaju havarije na velikim postrojenjima kada se istražuju uzroci štete raspolaganje kompletnom, svima dobro razumljivom dokumentacijom je minimalni uslov radi naplate osiguranja. Stručni termini i slike uredjaja oslanjaju na prospektnu dokumentaciju proizvodjača [5].



Slika 2. Podaci iz tabele ''Rečnika'' sa slikom uredjaja

U osnovi svaki od elemenata tabele čini deo složene strukture poznate kao tezaurus. Detaljnija analiza tokom realizacije prototipa pokazala je da potrebna osnovna struktura nije klasični višejezički rečnik već „višejezički distribuirani tezaurus". Elementi tezaurusa po klasičnoj definiciji mogu da budu reči – termini ili sintagme i njihove relacije. Posmatrani slučaj je karakterističan jer se pojavila potreba da kao elementi tezaurusa figuriraju podaci uopšte, ali moraju da se uvedu i slike, dokumenti, pojedinačne baze podataka ili drugi izvori podataka. Naime, hiperlinkovi u kontekstu ove aplikacije objektivno imaju funkciju relacija tezaurusa što proizlazi iz koncepta semantičkog Web-a. Naprimer, lista mogućih zamena za prekidače predstavlja skup RT (Related Terms) relacija u formi skupa hiperlinkova ka Web sajtovima respektivnih proizvođača. Specifikacija tehničkih karakteristika prekidača sa aspekta bezbednosti može da se predstavi kao BT (Broader Term) relacija ili da se predstavi kao SN (Scope Note). Prevod termina na strani jezik ima karakteristike sinonima UF (Used For) ili USE tj. koristi strani naziv prekidača. Pokazuje se da u domenu tehničke dokumentacije osim hijerarhijskih i semantičkih relacija mogu da se uvedu i drugi tipovi uslovnih ili „slabih" relacija. Upravo zato je za realizaciju izabran FileMaker koji veoma dobro spaja sve ključne osobine relacionih baza podataka i Information Retrieval softvera. U urgentnim slučajevima nužno je što pre pronaći potrebne dokumente, pri čemu se često ne zna ni koje, a hijerarhijska organizacija u tim slučajevima ne mora da bude najpogodnija. U literaturi se tezaurus obično posmatra kao hijerarhijska struktura međutim u pitanju je pogrešna interpretacija pošto uvedene relacije stvaraju mrežnu strukturu između elemenata liste termina.

Elementi strukture (osim ID oznake) imaju svoju podstrukturu koja se odnosi na prevode i sinonime (na različitim jezicima se koriste isti ali i različiti nazivi za isti pojam), na definicije pojmova, na načine pristupa vezanim pojmovima (na internetu, drugoj bazi podataka itd.). Na slici 3. je prikazan jedan slog podataka za aplikaciju " Rečnik".



Slika 3. Korisnički interfejs aplikacije "Rečnik"

U slučajevima pristupa drugoj bazi podataka često je neophodno da se definiše „agent“ koji će na odgovarajući način da izvrši pretraživanje i vrati podatke u potrebnom formatu. U ovom slučaju „agent“ se pojavljuje u rudimentarnoj formi složenog izračunatog upita i interakcije sa drugim izvorima podataka.

V. PROCES PRETRAŽIVANJA

Specifičnosti jezika (posebno deklinacije u srpskom jeziku) uslovljavaju složenije kontrolisanje procesa formiranja i pretraživanja rečnika. Korišćeni bazni program FileMaker Pro v10 ispunjava najveći deo zahteva za kvalitetnim pretraživanjem teksta u savremenom smislu.

U WEB verziji aplikacije pojavljuje potreba za nešto drugačijom organizacijom programa. Naime, u radu sa klasičnom aplikacijom moguće je da se za prenos selektovanih podataka (podataka za prevod) koristi *clipboard*. WEB verzija zahteva da se selektovani podaci prebace u posebno polje forme a zatim se dalji proces obrade aktivira preko posebnog dugmeta. Ovo je u svakom slučaju specifičnost vezana za WEB aplikacije pošto *clipboard* na klijentskoj strani nema nikakve veze sa serverskom stranom aplikacije. Pošto je na raspolaganju bila volumenska licenca FileMakera testirana je i simultana interakcija tri kopije aplikacije na jednoj realnoj i dve virtuelne mašine (u virtuelnoj mreži). Ista postavka je korišćena i za testiranje WEB verzije aplikacije.

Kao veoma koristan element u FileMaker programu pokazala se ugrađena tzv. WebViewer kontrola koja dozvoljava jednostavan pristup proizvoljnoj Web stranici. URL stranice se određuje na osnovu tekućih podatka u polju baze podataka ili kalkulisanoj na osnovu zadatih internih i eksternih podataka. Time se 1) omogućava pristup mobilnim skupovima dokumenata (koji u principu mogu da se „šetaju“ kroz mrežu), 2) omogućava dinamičko praćenje promena u sadržaju dokumenata (identifikacija ažurnosti dokumentacije).

Model kao i radne procedure zasnovani su na primeni osnovnih principa tzv. „programiranja agenata“. Svaka instanca FileMaker-a instaliranog na radnoj stanici ili u formi virtuelizovane aplikacije ima elementarne karakteristike „agenta“. Komunikacija se odvija ili u remote modu (razmenom SQL upita) i/ili korišćenjem skript jezika (AutoIt3 i WSH).

VI. ZAKLJUČAK

Aplikacija "Rečnik" daje mogućnost višejezičke interpretacije pojmova iz uže zadate stručne oblasti. Ispitane su mogućnosti realizovanja multimedijalne baze podataka. U okviru toga ispitani su metodi formiranja Web zasnovanog informacionog sistema korišćenjem standardnog komercijalnog softvera niske cene.

Aplikacija ima veoma skromne memorijske zahteve (reda pola gigabajta ukupno sa FileMaker-om), jednostavna je za korišćenje i mobilna. Rad u virtuelnom okruženju pokazao je potpunu stabilnost funkcionisanja.

Dalji rad će obezbediti konfornije prevodjenje srpskih termina upotrebljenih u različitim gramatičkim formama (padeži, množina) i savršeniji korisnički interface. Trenutno se radi na dopuni tabele termina i to stručnim terminima na ruskom jeziku kao i često korišćenim skraćenicama

LITERATURA

[1] Document Management System, Document Locator, http://www.documentlocator.com/Solutions/Document-Management-System/

[2] FileMaker Pro 10 User’s Guide, August 2009.

[3] Julijana Mirčevski, Nikola Popović, O problemu steganografske analize javnih WEB sajtova, Međunarodni naučno-stručni skup INFORMACIONA BEZBEDNOST 2009, Beograd, februar 2009. godine

[4] Nikola Popović, Julijana Mirčevski, Mogućnosti primene virtuelizacije u sistemu državnih organa i problemi zaštite, Konferencija ISDOS, Beograd, septembar 2009.

[5] Prospektni materijali proizvoda "Minel – ELPRO", Beograd

ABSTRACT

Abstract — The paper describes the realisation of the retrieval system protype dedicated to a professional technical documentation management. The implementation was made in data-base software for Windows (FileMaker), the low price commercial software package. Prototype has been developed on the catalogue and leaflets of technical documentation for residual current circuit-breakers in order to provide their multilingual interpretation.

SPECIALIZED SEARCHING SYSTEM

Julijana Mirčevski, Nikola Popović