Sinteza
Međunarodna naučna konferencija Univerziteta Singidunum

Impact of Internet on Business Activities in Serbia and Worldwide
Uticaj Interneta na poslovanje u Srbiji i svetu

DOI: 10.15308/SINTEZA-2014-957-964

UTVRĐIVANJE IDENTITETA OSOBE NA OSNOVU LIČNOG IMENA SA PRIMENAMA U AKREDITACIJI I ANALIZI AFILIJACIJA NAUČNIH RADOVA

Irena D. Mitrović[1], Jelica Ž. Protić[2], Ivana P. Kostić - Kovačević[3]
[1]Univerzitet u Beogradu, ICUB
[2]Univerzitet u Beogradu, Elektrotehnički fakultet
[3]Univerzitet Sigidunum, Beograd

Abstract:
Prilikom utvrđivanja identiteta osobe na osnovu ličnog imena nailazi se na više tipova problema. U prvu grupu spadaju problemi kada ime nije napisano po zahtevanoj sintaksi: različit redosled imena/prezimena od zadatog, različito pismo (ćirilica/latinica), nepoznato srednje slovo, izostavljanje dijakritičkih znakova, navođenje ili izostavljanje crtice između dva prezimena itd. Ime može biti i pogrešno napisano usled greške u kucanju. U drugu grupu spadaju problemi prepoznavanja identiteta osobe kada je ime bitno izmenjeno, na osnovu konteksta ili dodatnih podataka (promena prezimena, dodavanje prezimena, navođenje nadimka i sl.). Problemi identifikacije osobe u slučaju postojanja više osoba sa istim imenom i prezimenom spadaju u treću grupu problema. Postoje softverska rešenja koja se bave uparivanjem aproksimativno navedenih imena sa tačnim imenima po zadatom formatu. Specifičan problem uparivanja predstavlja problem povezivanja naučnih radova koji se identifikuju preko DOI-ja sa autorima koji se identifikuju ličnim imenima, na osnovu afilijacije, korišćenjem tehnika prepoznavanja imena, kao i pridruživanjem instituciji, grupi autora, naučnoj oblasti i ključnim rečima. Na međunarodnom nivou postoje pokušaji da se jedinstveno identifikuju istraživači i da se tako ovaj problem reši na egzaktan način (ResearcherID i ORCID). U ovom radu prikazani su problemi i prednosti obaveznog uključivanja naših istraživača u ove sisteme jedinstvene identifikacije, kao i moguća uloga tih identifikatora u informacionim sistemima naših obrazovno-naučnih institucija. Takođe su prikazani primeri primene prepoznavanja imena u formularima i bazama podataka, u postupku parsiranja, kao i procene grešaka u utvrđivanju identiteta u praksi.

Key words:
identitet osobe,
parsiranje,
akreditacija,
afilijacija.

UVOD

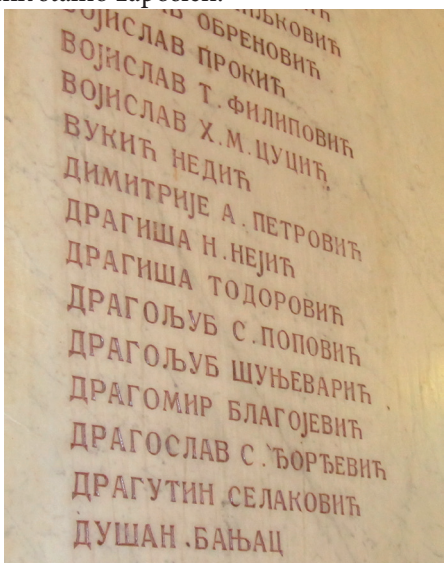
Uobičajeno je da lično ime predstavlja kombinaciju imena osobe dobijenog po rođenju, zatim srednjeg imena koje se u nekim kulturama koristi kao standardni deo imena osobe, i porodičnog imena, odnosno prezime. U zavisnosti od kulture i jezika zavise pravila davanja imena, kao i način pisanja istog [1]. Po pravilu kultura sa zapada prvo se piše ime pa prezime, dok se u nekim istočnim kulturama prvo piše prezime pa ime, što je kod nas uobičajeno i u nekim administrativnim dokumentima.

Sveopštom digitalizacijom podaci koji su neophodni u radnim procesima koji se odvijaju na visokoškolskoj ustanovi prebačeni su u elektronski oblik. Međutim, ne postoji standard kojim bi se na jedinstven način odredio format koji se tom prilikom koristi, a kao pismo kod nas se koristi i ćirilica i latinica, dok se neki tekstovi starijeg datuma, pa i lična imena, čak mogu naći u formi bez dijakritičkih znakova, navedena isključivo velikim slovima (npr. BOZIC umesto Božić) itd. Najveći problem je što se podaci nalaze u najraznovrsnijim oblicima i formatima: deo podataka se nalazi u *MS Word* ili *MS Excel* datotekama, deo podataka se nalazi u bazama informacionih sistema, a deo dokumenata koja sadrže imena (na primer neki stari naučni radovi, stranice radnih knjižica zaposlenih ili njihovi ugovori o radu) su samo skenirani i ne nalaze se u tekstualnom obliku. Same visokoškolske ustanove vrlo često koriste različita softverska rešenja za potrebe nastavnog procesa, za potrebe kadrovske evidencije, za potrebe finansija i računovodstva ili za potrebe praćenja naučnih rezultata zaposlenih, koristeći pri tome različite softverske platforme, pa kompatibilnost i interoperabilnost nije obezbeđena ni na nivou pojedinačne ustanove.

Dok je u svakodnevnom životu ime i prezime, a ponekad i nadimak, sasvim dovoljna odrednica za jedinstvenu identifikaciju osobe, na nivou digitalizovanih podataka i u bazama informacionih sistemima, situacija može biti drugačija. Kako imena i prezimena ne određuju identitet na jedinstven način, informacioni sistemi najčešće koriste numerički negovoreći (tj. bez posebnog značenja) identifikator. Iz tog razloga se vrlo često uz lično ime čuvaju numerički podaci, kao što su broj lične karte, pasoša, zdravstvene knjižice ili jednostavno broj koji je generisan u cilju jedinstvene identifikacije u tom konkretnom sistemu. Jedinstveni matični broj (JMBG) predstavlja numerički identifikator namenjen za prepoznavanje identiteta. JMBG je identifikacioni broj koji je dodeljivan svim građanima u bivšoj SFRJ počev od 1976. godine prema mestu rođenja, a za građane rođene pre 1976. godine prema mestu prebivališta. U okviru ovog broja nalaze se informacije o datumu i mestu rođenja, kao i o polu individue, pa se on smatra govorećim. Ovaj lični broj je još uvek u upotrebi u više država nastalih iz bivših republika SFRJ. Nažalost, u Srbiji ovaj broj nije dovoljno pouzdan jer sadrži dosta grešaka, detektovane su osobe sa više različitih JMBG, pa se ne može koristiti sa stopostotnom sigurnošću kod utvrđivanja identiteta.

S obzirom na specifičnost delatnosti visokoškolskih ustanova, praksu publikovanja radova i konkurisanja za naučno-stručne projekte, podaci o zaposlenima jedne visokoškolske ustanove mogu se nalaziti i u bazama podataka biblioteka i ministarstava koja kao resore imaju prosvetu i nauku. Takođe, u indeksnim bazama postoje podaci o zaposlenima koji su autori naučnih radova.Veliki napori se ulažu, kako kod nas tako i u svetu, da se autori naučnih radova jedinstveno identifikuju.

Univerziteti imaju svoje ustanove članice, fakultete i institute, a nastavnici često rade na više ustanova, učestvuju u komisijama na drugim ustanovama i u drugim oblicima saradnje, pa se imena nastavnika nalaze i u bazama drugih visokoškolskih ustanova, pored one u kojoj je nastavnik stalno zaposlen.



Slika 1 Prikaz liste uklesanih imena studenata poginulih za otadžbinu 1912-1919. godine, koja se nalazi u auli Rektorata Univerziteta u Beogradu. Uočavaju se imena bez srednjih slova, samo sa tačkom na mestu za srednje slovo, sa dva srednja slova, kao i ime Vukić Nedić, u kome nije sigurno šta je ime a šta prezime.

Ukoliko je potrebno da se podaci iz više izvora međusobno ukrštaju radi dobijanja izveštaja koji se traže za procese akreditacije, evaluacije ili rangiranja visokoškolskih ustanova tada to zahteva uparivanje podataka iz jednog izvora, na primer baze podataka, sa podacima iz drugog izvora. Prilikom uparivanja može se javiti čitav niz problema od prepoznavanja samog imena koje nije napisano na isti način, do prepoznavanja identiteta.

Prilikom akreditacije visokoškolskih ustanova jedan od najvažnijih priloga jeste lista imena i JMBG nastavnog osoblja, gde se zahteva da ova imena budu napisana ćirilicom, po redosledu: prezime, srednje slovo sa tačkom, ime. Formiranje ispravne liste imena pokazalo se kao veoma podložno greškama, a prikazani spisak sa Slike 1 star čitav vek, pokazuje da su problemi slični kao i pre 100 godina.

PROBLEMI PREPOZNAVANJA IMENA

Lično ime može imati nekoliko različitih, ali ispravnih varijacija. Broj varijacija zavisi od kulture i pisma koje se koristi. Pored ispravnih varijacija, postoje i varijacije imena koje su nastale usled grešaka [2][4]. U stranim imenima varijacije mogu poticati od grešaka prilikom spelovanja (npr. Smyth и Smith), od grešaka usled fonetske sličnosti (npr. Sinclair и St. Clair) ili zbog postojanja dvostrukog imena (npr. Jaun-Claude i Jaun Claude).

Prilikom prepoznavanja ličnog imena koje treba da bude napisano na srpskom jeziku po zadatom formatu često se ne mogu uparitio imena prostim poređenjem stringova, jer su u navođenju imena uočeni sledeći problemi:

- ime i prezime je napisano obrnutim redosledom od zahtevanog,
- lično ime je napisano latinicom umesto ćirilicom ili obrnuto, a moguče je i da je kombinovana latinica i ćirilica,
- prilikom pisanja izostavljeni su dijakritički znaci,
- prilikom pisanja dvostrukog prezimena izostavljene su crtice,
- prilikom pisanja ličnog imena napisani su i nazivi titula (najčešće *dr* ispred imena lekara),
- prilikom navođenja imena napravljena je greška u kucanju.

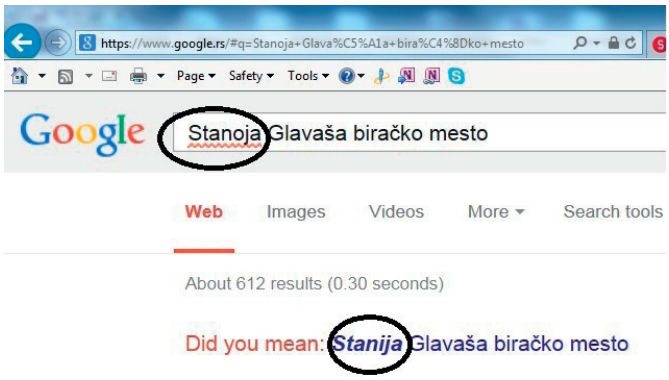
Ranija istraživanja na engleskom jeziku utvrdila su da oko 80% grešaka u kucanju čine pojedinačne greške [11]. Pojedinačne greške se mogu svrstati u sledeće kategorije:

- zamenjivanje pojedinačnog slova drugim slovom,
- izostavljanje pojedinačnog slova,
- umetanje dodatnog pojedinačnog slova,
- zamena redosleda dva susedne slova.

Pored navedenih pojedinačnih grešaka, pojavljuju se i višestruke pravopisne greške, kao što su: umetanje dva ili više dodatnih slova, izostavljanje dva ili više slova, međusobna zamena dva slova oko trećeg itd.

Sve varijacije ličnog imena, bez obzira da li potiču od grešaka ili ne, značajno otežavaju automatske obrade podataka u kojima je potrebno odrediti precizno šta je ime, srednje ime, a šta prezime. Ukoliko se radi o upotrebi imena i prezimena na srpskom jeziku, problem može pred-

stavljati i primena padeža. Naime, kod upotrebe punog srednjeg imena za identifikaciju, problem može proisteći iz formulacije "ime jednog od roditelja" gde neki korisnici pri unosu podataka u jednu formu navode ime oca, dok u drugoj formi navode ime majke.



Slika 2. Problemi sa prepoznavanjem imena usled upotrebe padeža u srpskom jeziku, koji uzrokuju konfuziju između muškog i ženskog imena

Čak i u slučaju da se za srednje ime navodi isti roditelj, u nekim slučajevima unosi se njegovo ime u nominativu, a u nekim u genitivu, gde u drugom slučaju ime može odgovarati ženskom imenu (na primer Dragan, Dragana). Problem padeža u kombinaciji sa muškim i ženskim imenima reflektuje se u primeru prikazanom na Slici 2, gde je jedan korisnik preko *Google* pretraživača tražio biračko mesto u ulici Stanoja Glavaša, a dobio predlog za ispravku imena Stanoja u ime Stanija, jer u tom kontekstu ime ulice nije nađeno, a Stanija u datom trenutku ima mnogo više pogodaka (817.000) nego Stanoje (253.000) ili Stanoja (210.000).

PROBLEMI UTVRĐIVANJA IDENTITETA OSOBE

Prepoznavanje ličnog imena u dokumentima, formularima ili bazama najčešće se vrši radi utvrđivanja identiteta osobe za koju je potrebno uraditi neku dodatnu akciju.

Najčešći problem prepoznavanja identiteta osobe jeste promena prezimena, koja se može izvršiti dodavanjem jednog prezimena na postojeće porodično prezime ili potpunim preuzimanjem drugog prezimena. Dodavanje prezime predstavlja lakši problem jer novo kombinovano prezime sadrži prethodno prezime. S druge strane, praksa pokazuje da osobe sa dva prezimena srazmerno često koriste samo jedno od prezimena ili obrću redosled ta dva prezimena ne poštujući zvanično lično ime iz identifikacionih dokumenata.

Dešava se i da autori radova nisu dosledni prilikom pisanja srednjeg imena u okviru afilijacije, a ovakav scenario posebno dolazi do izražaja kada se navodi samo prvo slovo imena i kada se ne može ustanoviti da li je došlo do greške prilikom pisanja. Kako se sve češće za podnošenje radova koristi specijalizovani softver, način unošenja može zavisiti od postojanja i predviđene duzine pojedinih polja za unos.

Najređe greške događaju se zbog navođenja nadimaka ili pseudonima, što praktično onemogućava uparivanje sa pravim ličnim imenom i dalju automatsku obradu podataka [3].

Ukoliko se ustanovi da se sa prepoznatim imenom ne može odrediti identitet osobe, onda se prelazi na alternativne metode. Automatski se može primeniti metoda kojom se porede dodatni podaci koji se tom prilikom obrađuju, a najčešće je to titula, datum rođenja, institucija kojoj pripada osoba i sl. Ukoliko se automatski ne može utvrditi identitet osobe preostaje manuelno utvrđivanje identiteta na osnovu konteksta podataka. U slučaju naučnih radova, taj kontekst može podrazumevati naučnu oblast, grupu koautora, a može se koristiti i aktuelna email adresa, pa neki časopisi i organizacije insistiraju na zvaničnoj, institucionalnoj email adresi.

PREGLED SPECIFIČNIH ALGORITAMA ZA ODREĐIVANJE LIČNIH IMENA

Proces uparivanja imena u najjednostavnijem slučaju se svodi na proces određivanja da li su dva stringa identična, dakle uparivanje imena predstavlja upoređivanje parsiranog stringa sa odgovarajućim referentnim stringom[5]. Ukoliko se samo proverava da li su dva stringa identična, onda se može govoriti o *exact name matching* algoritmima. Pošto su greške u pisanju imena česte, ovakav algoritam nije dovoljno robustan. Algoritmi koji pokušavaju da isprave greške u imenu nazivaju se *approximate name matching* algoritmi [2][4].

Approximate name matching najčešće koriste koeficijent sličnosti prilikom odlučivanja o slaganju imena. Postoje dve klase algoritama koje spadaju u ovu grupu: *pattern matching* i *phonetic coding.*

Pattern matching

Tehnike iz kategorije *pattern matching* predstavljaju najčešće korišćene tehnike kod *approximate name matching* algoritama. Najznačajnije dve klase ovih algoritama su *edit distance* i *n-gram distance.*

Edit distance algoritmi računaju distancu između pogrešno napisane reči i reči u rečniku, odnosno kolika je najmanja promena potrebna da bi se od pogrešne reči dobila reč u rečniku. Dozvoljene operacije su zamenjivanje slova, brisanje slova, umetanje slova i transponovanje dva susedna slova. Što je edit distanca između dva stringa manja, to se smatra da su stringovi sličniji [10].

Postoji nekoliko implementacija minimalnog edit distance algoritma, a za sve njih je zajedničko da se izračunavanjem koeficijenta sličnosti dobija vrednost između 1 (stringovi su identični) i 0 (stringovi su potpuno različiti). Početna ideja je bila da se generišu sve moguće varijacije pogrešno napisane reči, koje se od originalne reči razlikuju najviše za jednu grešku. Potencijalno ispravne varijacije su zatim bile testirane u odnosu na rečnik upotrebom exact string matching-a. Reč dužine n i pismo od p slova su generisali ukupno p(2n + 1) + n -1 mogućih varijacija koje je trebalo testirati. Npr. engleska reč od deset slova bi generisala 555 reči za uparivanje sa rečnikom. Neke od ovih varijacija mogu biti zanemarene pre testiranja s

obzirom da sadrže nemoguće kombinacije slova, tj. redosled slova, i kao takve ne postoje u rečniku. Primeri ovih algoritama su: Damerau–Levenshtei*n metrika, Weighted edit distance, Edit distance sa gornjom granicom i cut-off kriterijumom i Edit distance trie.*

N-gram metrika podrazumeva jedino poređenje slova i kao takva je nezavisna u odnosu na jezik. *N-grami* predstavljaju podstringove dužine *n* u dužim stringovima. Stringovi se smatraju utoliko sličnijim ukoliko imaju više zajedničkih *n-grama*. Najviše se upotrebljavaju unigrami ($n = 1$), bigrami ($n = 2$) i trigrami ($n = 3$). Npr. bigrami prezimena *Petrović* su {*Pe*, *et*, *tr*, *ro*, *ov*, *vi*, *ić*}, a trigrami su {*Pet*, *etr*, *tro*, *rov*, *ovi*, *vić*}. *N-gram* veličina sličnosti između dva stringa se izračunava prebrojavanjem zajedničkih *n-grama* u oba stringa i deljenjem ili brojem *n-grama* kraćeg stringa, ili brojem *n-grama* dužeg stringa, ili srednjim brojem *n-grama* oba stringa.

Phonetic coding

Sve tehnike fonetskog kodiranja pokušavaju da konvertuju string sa imenom u kod na osnovu toga kako se ime izgovara. Dva stringa, tj. imena, se smatraju identičnim ukoliko zvuče slično prilikom izgovaranja. Osnovna prednost ovih tehnika je brzina, a osnovni nedostatak je jezička zavisnost. Većina do sada razvijenih tehnika uzima u obzir samo engleski jezik, ali postoji i nekoliko tehnika prilagođenih drugim jezicima..

Soundex je najstariji *phonetic coding* algoritam patentiran za englesko govorno područje od strane Odell i Russell-a 1918. godine. *Soundex* prevodi ime u kod od četiri karaktera na osnovu zvuka svakog slova. Algoritam pretvaranja imena u kod se sastoji iz četiri koraka. U prvom koraku se sva slova, osim početnog, transformišu u brojeve na osnovu tabele 1. Nakon kodiranja kompletnog imena u drugom koraku se u kodu eliminišu sva uzastopna ponavljanja istih brojeva na samo jedan broj (npr. '222' se zamenjuje sa '2'), a u trećem se eliminišu sve nule. U poslednjem četvrtom koraku se dobija finalni kod koji je potrebno svesti na četiri karaktera, a to se postiže uklanjanjem suvišnih brojeva sa kraja koda ili dodavanjem nula na kraj koda sve dok se on ne svede na početno slovo i tri broja. Npr. *Soundex* kod za prezime Dickson se dobija na sledeći način: Dickson → D022205 → D0205 → D25 → D250.

Tabela 1 Soundex kodovi

Slova	Kod
a e h i o u w y	0
b f p v	1
c g j k q s x z	2
d t	3
l	4
m n	5
r	6

Kasnije je *Soundex* algorotam dalje usavršavan i doživeo je veliki broj varijacija kao što su: *Phonix, NYSIIS, Metaphone i Double–Metaphone*.

PROBLEMI PREPOZNAVANJA IDENTITETA U AFILIJACIJAMA RADOVA U NAUČNIM ČASOPISIMA

U Srbiji ne postoji definisan softver na nacionalnom nivou, a u velikom broju slučajeva ni na nivou visokoškolske institucije, koji egzaktno prati naučne rezultate istraživača. Jedini relevantan i ažuran izvor jeste baza KOBSON-a u koju se periodično iz indeksne baze Web of Science prenosi obrađeni deo podataka o autorima iz Srbije (*kobson.nb.rs*).

Pri pokušaju formiranja jedinstvene baze na nivou univerziteta, sa ciljem da se podaci prikupe radi apliciranja za poznata međunarodna rangiranja, analizirane su baze koje sadrže podatke o naučnim radovima autora: podaci koje su same institucije prikupile, podaci iz Scopusa, podaci iz Web of Science [12][13].

Iako je jedinstven format navođenja ličnog imena i prezimena u sve tri baze, i dalje je prisutan problem utvrđivanja identiteta autora. Pored problema utvrđivanja identiteta same osobe, u međunarodnim ideksnim bazama postoji problem i utvrđivanja institucije u kojoj je autor zaposlen. Ovaj problem nastao je kao posledica loših afilijacija. Usled nedostatka propisa kako se tačno navodi ime institucije, katedre, departmana, centra i laboratorije, autori su po svom nahođenju pisali afilijacije na engleskom jeziku, tako da se događa da za istu instituciju ima više desetina različitih načina navođenja imena i adrese.

Svi problemi prepoznavanja identiteta na osnovu ličnog imena mogu se kao primer naći u problemima identiteta osoba u afilijacijama naučnih publikacija. Specifičan problem koji se ovde javlja je navođenje samo prvog slova imena u formatu koji diktiraju pojedini časopisi, kojim se otvara mogućnost da više osoba ima isto prezime i prvo slovo imena. U tim slučajevima identitet osobe može se utvrditi na osnovu institucije u afilijaciji, na osnovu oblasti u kojoj je rad objavljen ili na osnovu drugih autora koji su potpisani na tom radu. Primer filtriranja na osnovu oblasti mogao se videti u softveru *Publish or perish,* ali ova opcija je isključena redizajnom Google Scholar-a 2012. godine.

PREGLED TRENUTNIH MEĐUNARODNIH REŠENJA ZA JEDINSTVENU IDENTIFIKACIJU AUTORA

Kako se vrednovanje naučnog rada sve više vrši kroz utvrđivanje broja objavljenih radova u časopisima sa SCI/SCIe/SSCI liste, kao i broja citata tih radova, tako sve više raste problem utvrđivanja jedinstvenog identiteta autora. Imena i prezimena ne predstavljaju jedinstvene entitete i kada se autori pretražuju na svetskom nivou često se dolazi do problema višeznačnosti, odnosno situacije u kojoj jedno isto ime i prezime odgovara većem broju autora. Registar autora koji bi dodelio jedinstven broj autoru, slično kao što registar DOI (digital object identifier) dodeljuje jedinstven broj samom naučnom radu, mogao bi da bude rešenje problema višeznačnosti[9].

Bilo je nekoliko pokušaja da se uspostavi registar. Najraniji pokušaji vezani su za specifične aplikacije ili

baze: arXiv Author ID (2005), Scopus Author ID (2006) i Thomson Reuters ResearcherID (2008). Postoje i pokušaji na nivou države kao što su DAI (Digital Author Identifier) za istraživače Holandije i Names Project za registrovanje istraživača Velike Britanije.

Na osnovu softvera ResearcherID, koji je razvio Thomson Reuters, napravljen je prototip Open Researcher Contributor Identification Initiative (ORCID) sa ciljem da napravi otvoreni nezavisni sistem koji bi bio dostupan svim istraživačima i kompatibilan sa indetifikatorima drugih sistema. Prvi ORCID servisi pušteni su u rad 2012. godine.

ORCID predstavlja podskup od International Standard Name Identifier (ISNI) uvedenog od strane International Organization for Standardization. ISNI jedinstveno utvrđuje identitet osoba koji su pisci, novinari, umetnici i sl.

ResearcherID

ResearcherID predstavlja komercijalno rešenje kojim se autorima dodeljuje jedinstveni broj[6]. Na sajtu ResearcherID autorima je omogućeno da linkuju svoje radove sa svojim jedinstvenim brojem čak i ako ti naučni radovi ne postoje u bazi Web of Science. Na ovaj način omogućeno je autorima da imaju na jednom mestu ažuran pregled svih svojih radova.

Pregled liste objavljenih radova značajna je mogućnost za istraživače koji najviše koriste recenzirane radova sa konferencija ili istražuju u oblasti gde je fokus na štampanju knjiga.

Kombinacijom DOI-a i ResearcherID-a omogućeno je jedinstveno povezivanje autora i naučnih radova. Thomson Reuters omogućio je razmenu između ResearcherID sistema i ORCID-a.

ORCID

ORCID predstavlja identifikator od 16 alfanumeričkih znakova i nije komercijalan[7]. ORCID ima dve osnovne funkcije, prva je održavanje registra jedinstvenih identifikatora i baze aktivnosti istraživača, a druga je razvoj aplikacija kojima se omogućava komunikacija sa drugim sistemima kao što su Scopus, ResearcherID ili LinkedIn.

ORCID, takođe, predviđa mogućnost registracije organizacija koje mogu linkovati svoje rekorde sa drugim ORCID identifikatorima, ažurirati ORCID podatke, primati novosti od ORCID-a i registrovati svoje zaposlene i studente.

Podaci iz ORCID registra su dostupni u skladu sa željama autora, a preporuka je da se autori gde god mogu pozovu na svoj ORCID identifikator.

Od oktobra 2012. godine Elsevier je sponzor i partner ORCID-a i planirana je integracija sa ORCID-ovim proizvodima i servisima[8].

ANALIZA MOGUĆNOSTI UKLJUČIVANJA NAŠIH AUTORA U MEĐUNARODNE REGISTRE ISTRAŽIVAČA

U Srbiji ne postoji identifikator koji bi poslužio za jedninstveno prepoznavanje srpskih istraživača, stoga ostaje mogućnost formiranja nacionalnog registra ili pridruživanje nekom od međunarodnih. Od međunarodnih registara dolaze u obzir ResearcherID i ORCID. Ukoliko se pristupi formiranju nacionalnog registra on bi trebalo da bude kompatibilan sa ORCID-ovim brojem radi moguće integracije dostupnih servisa.

Uključivanje u međunarodne registre može se obaviti na nivou institucija ili na nivou istraživača. Ukoliko se sprovede institucionalno, veća je garancija da svi istraživači budu registrovani i omogućava se lakše praćenje produktivnosti na nivou institucije. Međutim, registracija institucija se naplaćuje kod oba registra.

Formiranje centralnog registra istraživača koji dodeljuje jedinstvene identifikatore doprinelo bi sledećem:

- smanjila bi se potreba da autori višestruko popunjavaju podatke o sebi aplicirajući za razne potrebe,
- lakše bi se povezivali podaci ukoliko autor menja prezime, instituciju ili institucija menja naziv,
- rezultati naučnog rada postaju transparentniji, lakše pretraživi i ne zavise od oblasti istraživanja i međunarodnih okvira,
- omogućava se lakše povezivanje informacionih sistema kako pri ministarstvima tako i pri visokoškolskim ustanovama.

PREPOZNAVANJE IDENTITETA OSOBA RADI FORMIRANJA BAZE NASTAVNOG OSOBLJA NA UNIVERZITETU

Skoro svakodnevno se od visokoškolskih ustanova traži generisanje izveštaja koji zahtevaju uparivanje podataka iz dva izvora ili sama institucija ima potrebu da proširi svoju bazu podataka [14].

Za potrebe formiranja baze podataka o nastavnicima, saradnicima i istraživačima Univerziteta u Beogradu prikupljani su formulari u MS Excel-u. Excel formular je bio dobro definisan, sa postavljenim ograničenjima. Jasno su bila podeljena polja sa imenom, srednjim imenom, prezimenom, JMBG-om, zvanjem i institucijom zaposlenja. Pored definicija, formular je bio i zaključan tako da nije bilo moguće promeniti strukturu i pravila. Lični podaci služili su za identifikaciju, a sami formulari su prikupljani više puta u zavisnosti od problematike za koju su zahtevani podaci.

Pristigli formulari obrađivani su parserom koji je napisan na programskom jeziku Java. Parser je tražio osobe u bazi informacionog sistema Univerziteta i podacima iz formulara dopunjivao bazu [16].

Prilikom parsiranja nije bilo grešaka u samom pisanju imena već je bilo sledećih problema:

- zamenjeno popunjavanje polja ime i prezime,
- korišćeni blanko karakteri,
- ubacivanje titule u ime,
- korišćene duple tačke.

Parser je tražio osobu sa identičnim podacima JMBG, ime, srednje ime, prezime, zvanje i institucija. Ukoliko takva osoba nije pronađena softver je tražio osobu koja ima bar 5 identičnih podataka od ponuđenih 6. Takvom aproksimacijom su skoro svi podaci ažurirani, a problematični podatak prepisan je najnovijom verzijom. Najče-

šći problem ovakvog uparivanja je bio loš JMBG, a zatim pogrešno uneseno srednje ime, odnosno inicijal srednjeg imena.

PREPOZNAVANJE IDENTITETA OSOBA U PROCESU AKREDITACIJE VISOKOŠKOLSKE USTANOVE

Ozbiljan problem prilikom unosa podataka o novom studijskom programu jesu imena nastavnika koji drže pojedine predmete. Podaci o studijskom programu obično stižu u vidu više tabela, u kojima su na raznim katedrama i departmanima podaci unošeni na razne načine, najčešće u tekst procesoru i bez poštovanja precizne sintakse koja je ključna za proračun opterećenja nastavnika. Tako prilikom pridruživanja imena nastavnika predmetima neki unose imena latinicom, drugi ćirilicom, treći stavljaju srednje slovo, neki umesto imena unose samo inicijal, na nekim mestima je uneto prezime pa ime, na drugim ime pa prezime, negde i nedostaju dijakritički znaci kod slova čćđšž (Slika 4), a ima i slučajeva da se u okviru imena piše i titula, akademsko zvanje... U početku su ti podaci sređivani ručno, što nije težak, ali jeste obiman posao, koji na fakultetima sa mnogo programa može trajati danima, a ipak se u rezultujućem spisku nađe dosta grešaka, koje kasnije treba opet ručno ispravljati.

U ovom specijalnom slučaju postoji okolnost koja omogućava da se značajno ubrza ispravljanje imena i njihovo dovođenje u propisani format. Naime, postoji spisak svih nastavnika i saradnika u ustanovi i taj spisak je uvek unapred napravljen, tako da su imena napisana u predviđenom formatu i to je obično višestruko provereno u kadrovskim službama ustanove, tokom pripreme Knjige nastavnika neophodne pri akreditaciji i prikupljanja radnih knjižica koje se moraju skenirati. Dakle, koliko god da je neko ime u formularu za studijski program uneseno "slobodnim stilom" i ma koliko mu nedostajali neki podaci (srednje slovo, drugo prezime itd.), sa sigurnošću se zna da je to zapravo oblik nekog od imena sa spiska nastavnika i saradnika same ustanove.

Softver za proračun akreditacionih parametara na osnovu elektronskih formulara u *MS Excel*-u razvijen je na Elektrotehničkom fakultetu, a opis softvera i njegove funkcionalnosti su prikazane na sajtu Komisije za akreditaciju i proveru kvaliteta www.kapk.org. Kao dopunska funkcionalnost vezana za prepoznavanje imena razvijen je relativno jednostavan VBA (*Visual Basic for Applications*) modul za *MS Excel*, koji najpre učita i parsira knjigu nastavnika, upisujući u odgovarajuću matricu ime nastavnika, prezime (ili dva prezimena) i srednje slovo [15]. Zatim program kreće kroz listu "slobodno unetih" podataka, uzima prvi od njih, pretvara ga u latinicu i velika slova, razdvaja ga na reči (smatrajući i tačku delimiterom reči) i pokušava da svaku od reči pronađe među ispravno unetim podacima iz ranije učitane matrice. Svako uspešno uparivanje dodaje neki skor - najviše za upareno prezime, zatim

Slika 4. Na slici je prikazana detekcija i ispravljanje imena prilikom uparivanja fajlova sa podacima o nastavnicima sa podacima iz popunjenog elektronskog formulara u kojem se, između ostalog, radio i obračun opterećenja nastavnika i saradnika.

za upareno ime i najmanje za srednje slovo, ako ono postoji. Sortiranjem tih skorova dobija se lista potencijalnih imena koja najviše odgovaraju analiziranom unosu. Zatim se u susednu kolonu, pored analiziranog unosa, upisuje ime koje mu najviše odgovara (najviši skor poklapanja) a u naredne kolone eventualna druga imena koja imaju visok skor.

Operator zatim ima zadatak da prođe kroz tabelu i vidi da li ime sa najvišim skorom zaista odgovara unetom imenu osobe koja drži predmet. U praksi pomaže ako se pored imena iz Knjige nastavnika unese i sa koje je katedre i u kom je zvanju, ali to obično nije neophodno, osim na ustanovama sa jako mnogo nastavnika. Pokazuje se da je ime sa najvećim skorom gotovo uvek (preko 95% slučajeva) pravo ime, a tamo gde nije, jednostavno se neko od imena iz sledećih kolona, koje je imalo manji skor, prekopira ručno u kolonu rezultata i najzad pokrene procedura koja tako ispravljena imena vraća na list podataka o obaveznim i izbornim predmetima. U praksi se događaju i neke neregularne situacije koje moraju da se rešavaju ručno, recimo pokaže se da neki od nastavnika zaista drže predmet, a nisu upisani u Knjigu nastavnika, pa se prema tome mora modifikovati Knjiga nastavnika, ali se ukupno gledano posao koji je trajao više dana svede na sat ili dva provere, posle koje su podaci garantovano ispravno formatirani.

ZAKLJUČAK

U svakodnevnoj govornoj komunikaciji za prepoznavanje identiteta osoba je dovoljno navesti ime i prezime pa čak iako je pri tome načinjena greška. Automatsko prepoznavanje imena i prezimena, uz utvrđivanje identiteta osobe, još uvek predstavlja izazov. U cilju rešavanja problema rađene su analize tipova grešaka koje se javljaju prilikom pisanja ličnog imena, kao i algoritmi za uparivanje imena.

Problem identiteta koji se nadovezuje na prepoznavanje ličnog imena utiče na mogućnosti kombinovanja i uparivanja podataka iz više izvora. U Srbiji ovaj problem je naročito izražen jer su podaci o nastavnicima, saradnicima i istraživačima nalaze u različitim formatima u različitim elektronskim verzijama, od dokumenata do baza u različitim institucijama koje se tiču visokog obrazovanja, a koriste se. i dva pisma, latinično i ćirilično.

Jedan deo rešenja utvrđivanja identiteta leži u formiranju nacionalnog registra istraživača, koji može biti deo nekog međunarodnog sistema identifikacije. Drugi deo problema, koga čine formiranje izveštaja na osnovu uparivanja podataka iz različitih izvora, rešava se pojedinačnim softverskim rešenjima, kojim se podaci parsiraju, uparuju i na kraju povezuju i obrađuju.

Pošto korišćenje JMBG-a nije prihvatljivo zbog poverljivosti podataka, kao i zbog prisustva stranaca koji JMBG nemaju (npr. lektori za strane jezike iz različitih zemalja), uparivanje podataka bi bilo značajno olakšano postojanjem jedinstvenog identifikatora, koji bi mogao da se unese u sve informacione sisteme visokog obrazovanja i time gotovo minimizuje problem uparivanja podataka iz više baza ili dokumenata. Taj identifikator mogao bi da bude isti kao u registru istraživača ili nekom drugom registru formiranom pri Ministarstvu prosvete i nauke, ali bi univerzalno rešenje svakako bilo pridruživanje nekom od već postojećih sistema identifikacije istraživača na međunarodnom nivou.

LITERATURA

[1] R. Alford. Naming and identity: A Cross-cultural Study of Personal Naming , Practices. New Haven, CT: HRAF, 1988

[2] A. J. Lait and B. Randell. An assessment of name matching algorithms. Technical Report, Department of Computing Science, University of Newcastle upon Tyne, 1993

[3] P. Driscoll, D. Yarowsky, Disambiguation of Standardized Personal Name Variants, In Proceedings of Multisource, Multilingual Information Extraction and Summarization, RANLP, 2007

[4] P. Christen. A comparison of personal name matching: Techniques and practical issues. Tecnical Report TR-CS-06-02, Department of Computer Science, The Australian National University, Canberra, Australia, 2006

[5] P. Thompson and C. Dozier. Name searching and information retrieval. In Proceedings of Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island, 1997

[6] http://www.researcherid.com/resources/html/dsy5769-TRS.html

[7] http://orcid.org/

[8] http://www.elsevier.com/about/press-releases/science-and-technology/elsevier-joins-orcid-in-announcing-launch-of-orcid-registry

[9] D. Butler. Scientists: your number is up, Nature News, 2012

[10] F. J. Damerau. A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3):171–176, 1964.

[11] J. L. Peterson. A note on undetected typing errors. Communications of the ACM, 29(7):633–637, 1986

[12] Mitrović I., Protić J., Popović M., Romić U., Integracija podataka o naučnim radovima u informacioni sistem Univerziteta u Beogradu, TREND 2013, Maribor, 2013.

[13] Mitrović I., Protić J., Analiza podataka o naučnim radovima autora sa Univerziteta u Beogradu na osnovu izvora Web of science, ETRAN 2013, Zlatibor, 2013

[14] I. Odžić, J. Protić, Iskustva u formiranju baze podataka Univerziteta u Beogradu na temelju elektroniskih obrazaca prikupljenih iz više izvora, Telfor 2009

[15] M. Petrović. Master rad: Softver za aproksimativnu identifikaciju ličnih imena i njegova primena u pripremi akreditacije, Elektrotehnički fakultet Univerziteta u Beogradu, 2010

[16] I. Odžić. Magistarski rad: Razvoj softverske podrške za prikupljanje , ažuriranje i prezentaciju podataka o akreditovanim programima i nastavnom osoblju Univerziteta u Beogradu, aproksimativnu identifikaciju ličnih imena i njegova primena u pripremi akreditacije, Elektrotehnički fakultet Univerziteta u Beogradu, 2010

PERSON'S IDENTITY DETERMINATION BASED ON PERSONAL NAME WITH IMPLEMENTATION IN ACCREDITATION AND IN ANALYSES OF AFFILIATION OF SCIENTIFIC PAPERS

Abstract:
Several types of problems can arise in determining a person's identity based on personal name. The first group includes problems that occur when the name is not written according to the requested syntax: with different order of name / surname, in different alphabets (Cyrillic/Latin), when the middle name is unknown, when diacritical marks are missing, when a hyphen between the two names is written or omitted etc. The name can be misspelled due to typographical errors. The second group includes the problems of recognizing a person's identity when the name changed significantly, based on the context or additional information (change of surname, added names, nicknames, etc.). The challenge of distinguishing between persons with the same name belong to the third group of problems. There are software solutions that deal with the approximate matching of written names with the correct names for a given format. A specific problem is the problem of pairing the scientific paper, identified by DOI, with authors, identified by signatures, on the basis of affiliation, using the techniques of recognizing names, as well as joining them with the institution, group of authors, the scientific field and keywords. At the international level, there are attempts to uniquely identify researchers, in order to solve this problem in an exact manner such as ResearcherID and ORCID. This paper describes the problems and benefits of mandatory inclusion of our researchers in these systems of unified identification, as well as the possible role of these identifiers in the information systems of our educational and research institution. We will present some examples of the name recognition implementations applied to specific forms and databases during the process of parsing, as well as estimation of error rates in practical person's identification.

Key words:
personal identity, parsing; accreditation, affiliation.